MATHEMATICAL MODELLING IN ENGINEERING & HUMAN BEHAVIOUR

PROCEEDINGS

im²
Instituto Universitario de Matemática Multidisciplinar

GENERALITAT VALENCIANA
Conselleria d'Innovació, Universitats, Ciència i Societat Digital
This event is partially financed by Grant CIAORG/2022/23.

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Modelling for Engineering & Human Behaviour 2023

València, July 11th-14th, 2023

This book includes the extended abstracts of papers presented at XXV Edition of the Mathematical Modelling Conference Series at the Institute for Multidisciplinary Mathematics *Mathematical Modelling in Engineering & Human Behaviour 2023*.

ISBN: 978-84-09-57681-4
Edited by:
I.U. de Matemàtica Multidisciplinar, Universitat Politècnica de València.
J.R. Torregrosa, J-C. Cortés and A. Vidal-Ferràndiz.

Cover Page Image by Jaume Hervás Raga, http://www.hagamosalgointeligente.com

Report any problems with this document to imm@imm.upv.es.

Instituto Universitario de Matemática Multidisciplinar

UPV

GENERALITAT VALENCIANA
Conselleria d'Innovació, Universitats, Ciència i Societat Digital

Grant CIAORG/2022/23 of Generalitat Valenciana: Conselleria d'innovació, Universitats, Ciència i Societat Digital.

Preface

This book includes the extended abstracts presented at XXV Edition of the Mathematical Modelling Conference Series at the Institute for Multidisciplinary Mathematics Mathematical Modelling in Engineering & Human Behaviour.

December 2023

Juan Ramón Torregrosa
MME&HB 2023
Universitat Politècnica de València

Organization

Mathematical Modelling in Engineering & Human Behaviour (MME&HB 2023) is organized by the Institute of Multidisciplinary Mathematics (IMM) of Universitat Politècnica de València.

Organizing Committee

The organizing Committee of the MME&HB 2023 is composed by:

Juan Ramon Torregrosa	Universitat Politècnica de València
Antonio Hervás Jorge	Universitat Politècnica de València
Juan Carlos Cortés	Universitat Politècnica de València
Damian Ginestar Peiró	Universitat Politècnica de València
Begoña Cantó Colomina	Universitat Politècnica de València
Dolors Roselló Ferragud	Universitat Politècnica de València
Francisco Chicharro López	Universitat Politècnica de València
Alicia Herrero Debón	Universitat Politècnica de València
Antoni Vidal Ferràndiz	Universitat Politècnica de València
Elena López Navarro	Universitat Politècnica de València

Sponsoring Institutions

- Universitat Politècnica de València

- Grant CIAORG/2022/23 of Generalitat Valenciana: Conselleria d'innovació, Universitats, Ciència i Societat Digital.

Table of Contents

III Maths and Physics

IV Mathematical Modelling in Public Health

VI Mathematical Models in Social Science and Financial Mathematics

VII Complex Networks, Graphs, and Applications

VIII Mathematical Models in Population Dynamics

IX Recent Advances in the Approximation of Matrix Functions

X Student Project's

Part I

Mathematical Methods for Engineering Problems

Balancing sustainability and occupational health in airport operations

Bruno Brentan[1], Silvia Carpitella[2], Antonella Certa[3], and Joaquín Izquierdo[4]

[1] Federal University of Minas Gerais, Av. Presidente Antônio Carlos, Belo Horizonte, Brazil
[2] California State University, Northridge, 18111 Nordhoff St, Northridge, CA, USA, silvia.carpitella@csun.edu
[3] University of Palermo, Viale delle Scienze, Palermo, Italy
[4] Universitat Politècnica de València, Cno. de Vera s/n, Valencia, Spain

Abstract. The interconnections among sustainability factors in airports and their effects on occupational health are of paramount importance. Not only are airports key drivers of economic growth, employment, and tourism, but they also have significant and far-reaching environmental and social impacts. Therefore, achieving sustainable airport operations requires a delicate balance between economic, environmental, and social factors. To this end, collaboration among stakeholders is crucial for developing innovative solutions that protect workers' health and promote sustainability. Mathematical models play a key role in decision-making and policy development to ensure sustainable airports while safeguarding workers' well-being. In this context, this contribution proposes the use of Fuzzy Cognitive Maps (FCMs) to evaluate the most significant Occupational Stress Risks (OSRs) for an Italian airport. The analysis aims to assess the subset of OSRs whose potential occurrence may likely impact the occurrence of other related OSRs. The study concludes by proposing potential prevention/reduction strategies for each of these OSRs.

Keywords: Occupational Health, Sustainability, Fuzzy Cognitive Maps

1 Introduction

Achieving sustainable airport operations requires a delicate balance between the economic, environmental, and social dimensions of sustainability [1], while also considering the health and safety of workers who are exposed to various physical and psychological hazards such as noise, vibration, air pollution, and stress. Pursuing this goal requires the effort and involvement of all stakeholders, including airport operators, airlines, passengers, local communities, and government agencies. Collaboration and partnerships among these stakeholders can help identify and implement innovative and effective solutions that promote sustainability [2] and protect the environment and public health. This necessitates an integrated and comprehensive approach that takes into account the interdependencies among various sustainability factors and their potential effects on

occupational health. Sustainable airport operations can also generate numerous benefits, such as improved air quality, reduced noise pollution, enhanced energy efficiency, and cost savings, which can contribute to the airport's long-term success and competitiveness. The complexity of this problem makes it a fascinating and challenging area for human behavior engineering modeling. By understanding how different sustainability factors interconnect and impact human behavior, we can design interventions that promote sustainable practices and protect workers' health. To this end, suitable decision-making mathematical models can help analyze the relationships among the critical factors that impact sustainability and occupational health. These models can facilitate trade-off evaluations among sustainability dimensions and provide decision-makers with a powerful tool to optimize outcomes and inform policy decisions. By leveraging the information gained from these models, we can achieve sustainable airport operations that promote economic growth and social and environmental sustainability while safeguarding the health and safety of airport workers.

High-risk environments can lead to increased stress levels, negatively impacting workers' well-being. Stress can compromise focus and decision-making abilities, consequently elevating the risks of accidents. Therefore, implementing supportive measures is essential to manage stress and ensure overall safety. Occupational stress encompasses the physical, emotional, and psychological strain that individuals experience due to their work. Factors like heavy workloads, tight deadlines, lack of control, conflicts with colleagues, and job insecurity contribute to this stress. Prolonged exposure to such stressors can have harmful effects on mental and physical health, job performance, and overall well-being. In high-risk environments such as the aviation industry, emergency services, and construction sites, among others, occupational stress intensifies safety hazards exponentially. It becomes crucial for companies to establish a methodological evaluation of occupational stress risks and develop effective management plans. Mathematical modeling plays a pivotal role in this process, allowing for the enhancement of safety measures and consideration of uncertain conditions. By managing risks proactively, organizations can create a safer and healthier work environment for their employees.

2 Methodological approach

Fuzzy Cognitive Maps (FCMs) are advanced mathematical models designed to represent complex relationships and interactions between different concepts or variables. These models utilize fuzzy logic, which allows them to handle the inherent uncertainty and imprecision in human thinking and decision-making. In the context of occupational stress risk evaluation, FCMs offer valuable advantages. They help capture the intricate nature of stress risks, providing a deeper understanding of the factors involved. By building FCMs, informed decision-making becomes possible, leading to the development of effective stress reduction strategies.

Specifically, FCMs can be useful in several ways in the context of occupational stress risk evaluation, as described in the following.

- Identifying key factors: FCMs help identify the most influential factors contributing to occupational stress, highlighting the critical areas that require intervention or management.
- Assessing risk interdependencies: FCMs enable the evaluation of how different stress risks interact with and influence each other. This allows for a comprehensive understanding of the interconnectedness and potential cascading effects of stress risks within an occupational setting.
- Predicting stress outcomes: FCMs can be used to simulate different scenarios and predict the potential impact of interventions or changes in stress risk factors on occupational stress outcomes. This helps in developing targeted strategies for stress prevention and management.

A detailed application description can be consulted in [3].

3 The case of an Italian airport

As previously observed, airports can be highly stressful environments for operators, subjecting them to significant stress risks. An analysis of an Italian airport focused on identifying and understanding four key risk categories and, for each of these areas, a set of three Occupational Stress Risks (OSRs) that affect workers (twelve OSRs in total). To gain deeper insights into the complex relationships linking these OSRs with each other, an FCM was built on the set of twelve identified OSRs. The elements of analysis are summarized in the following.

- Workload and Time Pressure
  - OSR1: High workload and demanding schedules leading to increased stress levels.
  - OSR2: Tight deadlines and time pressure to ensure efficient operations.
  - OSR3: Balancing multiple tasks simultaneously, causing work overload and time constraints.
- Customer Service and Interactions
  - OSR4: Dealing with difficult or upset passengers, leading to emotional stress.
  - OSR5: Managing conflicts and resolving disputes between passengers.
  - OSR6: Constantly maintaining friendly and professional conduct while handling complaints or requests.
- Shift Work and Irregular Hours
  - OSR7: Irregular and rotating shifts disrupting sleep patterns and causing fatigue.
  - OSR8: Difficulty in maintaining work-life balance due to unpredictable schedules.
  - OSR9: Social and personal life limitations resulting from working during weekends, holidays, or night shifts.

– Sustainability and Job Insecurity
  - OSR10: High turnover rates due to insecurity hinder sustainable practices.
  - OSR11: Insufficient career development challenges airport sustainability.
  - OSR12: Insecure employment undermines employee well-being and hinders airport sustainability.

Figure 1 presents linguistic evaluations of preference which have been collected with the support of an expert in the field of safety and security working at the airport. These evaluations refer to the intensity of causality that an element imparts to another one, expressed as: very low (VL), low (L), medium (M), high (H), very high (VH).

	OSR1	OSR2	OSR3	OSR4	OSR5	OSR6	OSR7	OSR8	OSR9	OSR10	OSR11	OSR12
OSR1	0	H	VH	H	H	H	VH	VH	VH	M	L	L
OSR2	VH	0	VH	H	H	M	VH	H	VH	M	L	L
OSR3	VH	VH	0	H	H	H	H	VH	VH	H	L	L
OSR4	H	H	VH	0	VH	VH	L	H	H	H	L	H
OSR5	H	H	H	VH	0	VH	M	H	M	M	M	H
OSR6	H	H	H	VH	VH	0	H	L	L	L	L	L
OSR7	M	M	M	VH	VH	VH	0	VH	VH	M	M	M
OSR8	VH	H	H	VH	VH	VH	M	0	VH	M	M	M
OSR9	H	H	H	H	H	VH	H	H	0	H	H	H
OSR10	H	H	H	H	H	H	H	H	H	0	VH	VH
OSR11	M	M	M	M	M	M	M	M	H	H	0	VH
OSR12	VH	VH	VH	H	H	VH	H	VH	VH	H	VH	0

Fig. 1: Linguistic evaluations of preference.

	OSR1	OSR2	OSR3	OSR4	OSR5	OSR6	OSR7	OSR8	OSR9	OSR10	OSR11	OSR12	TE
OSR1	-	0.31	0.4	0.31	0.31	0.31	0.4	0.4	0.4	0.23	0.14	0.14	**M**
OSR2	0.4	-	0.4	0.31	0.31	0.23	0.4	0.31	0.4	0.23	0.14	0.14	**M**
OSR3	0.4	0.4	-	0.31	0.31	0.31	0.31	0.4	0.4	0.31	0.14	0.14	**M**
OSR4	0.31	0.31	0.4	-	0.4	0.4	0.14	0.31	0.31	0.31	0.14	0.31	**M**
OSR5	0.31	0.31	0.31	0.4	-	0.4	0.23	0.31	0.23	0.23	0.23	0.31	**M**
OSR6	0.31	0.31	0.31	0.4	0.4	-	0.31	0.14	0.14	0.14	0.14	0.14	**M**
OSR7	0.23	0.23	0.23	0.4	0.4	0.4	-	0.4	0.4	0.23	0.23	0.23	**M**
OSR8	0.4	0.31	0.31	0.4	0.4	0.4	0.23	-	0.4	0.23	0.23	0.23	**M**
OSR9	0.31	0.31	0.31	0.31	0.31	0.4	0.31	0.31	-	0.31	0.31	0.31	**H**
OSR10	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	-	0.4	0.4	**H**
OSR11	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.31	0.31	-	0.4	**M**
OSR12	0.4	0.4	0.4	0.31	0.31	0.4	0.31	0.4	0.4	0.31	0.4	-	**H**

Fig. 2: Defuzzified fuzzy numbers and total effect (TE) associated to each OSR, OSRs 9, 10 and 12 having associated higher influence suggesting higher priority.

These evaluations have been translated to trapezoidal fuzzy numbers, as shown in [4], which have been defuzzified by applying the centroid method, as explained in [5]. Crisp evaluations are reported in Figure 2, while the FCM displayed in Figure 3 was produced with the MentalModeler software. It shows the 132 connections identified for the 12 elements, which correspond to 11 connections per element.

The study aims to shed light on the interconnections and relationships between different OSRs within each identified category. By employing FCM, which incorporates fuzzy logic, we are able to account for the inherent uncertainty and intricacies of human thinking and decision-making processes as well as the interdependencies of OSRs. By capturing these interdependencies, decision-makers, and stakeholders can gain valuable insights into the root causes and potential cascading effects of stress risks in the airport environment.

The findings of this FCM-driven approach provided a foundation for developing targeted stress reduction strategies and interventions. With a deeper understanding of the connections between OSRs, airport management, and relevant authorities can be better equipped to implement measures that address the underlying issues effectively. To conclude our study, we propose a set of potential strategies that could be implemented by the management to deal with the most interdependent OSRs, as a result of the FCM-based analysis.

- OSR9: Social and personal life limitations resulting from working during weekends, holidays, or night shifts
  - Flexible Scheduling and Shift Rotation: Implement a system that allows employees to have a fair distribution of working hours, including weekdays and weekends off, reducing social and personal life limitations.
  - Employee Support Programs: Establish counseling services, stress management workshops, and resources to assist employees in coping with the challenges of working irregular hours and provide a supportive work environment.
  - Job Rotation and Cross-Training: Rotate employees through different roles and responsibilities to break the monotony, prevent burnout, and enhance their skills, while also allowing them to fill in for one another during undesirable shifts.
- OSR10: High turnover rates due to insecurity hinder sustainable practices
  - Enhance Job Security Measures: Implement measures to enhance job security for airport employees, such as offering long-term contracts, providing clear career progression pathways, and ensuring fair and competitive compensation.
  - Foster a Positive Organizational Culture: Cultivate a positive organizational culture that prioritizes employee well-being, open communication, and employee involvement in decision-making processes.
  - Implement Sustainable Work Practices: Establish sustainable work practices within the airport environment. This can include initiatives such as promoting work-life balance, reducing excessive workloads, and implementing stress management programs.

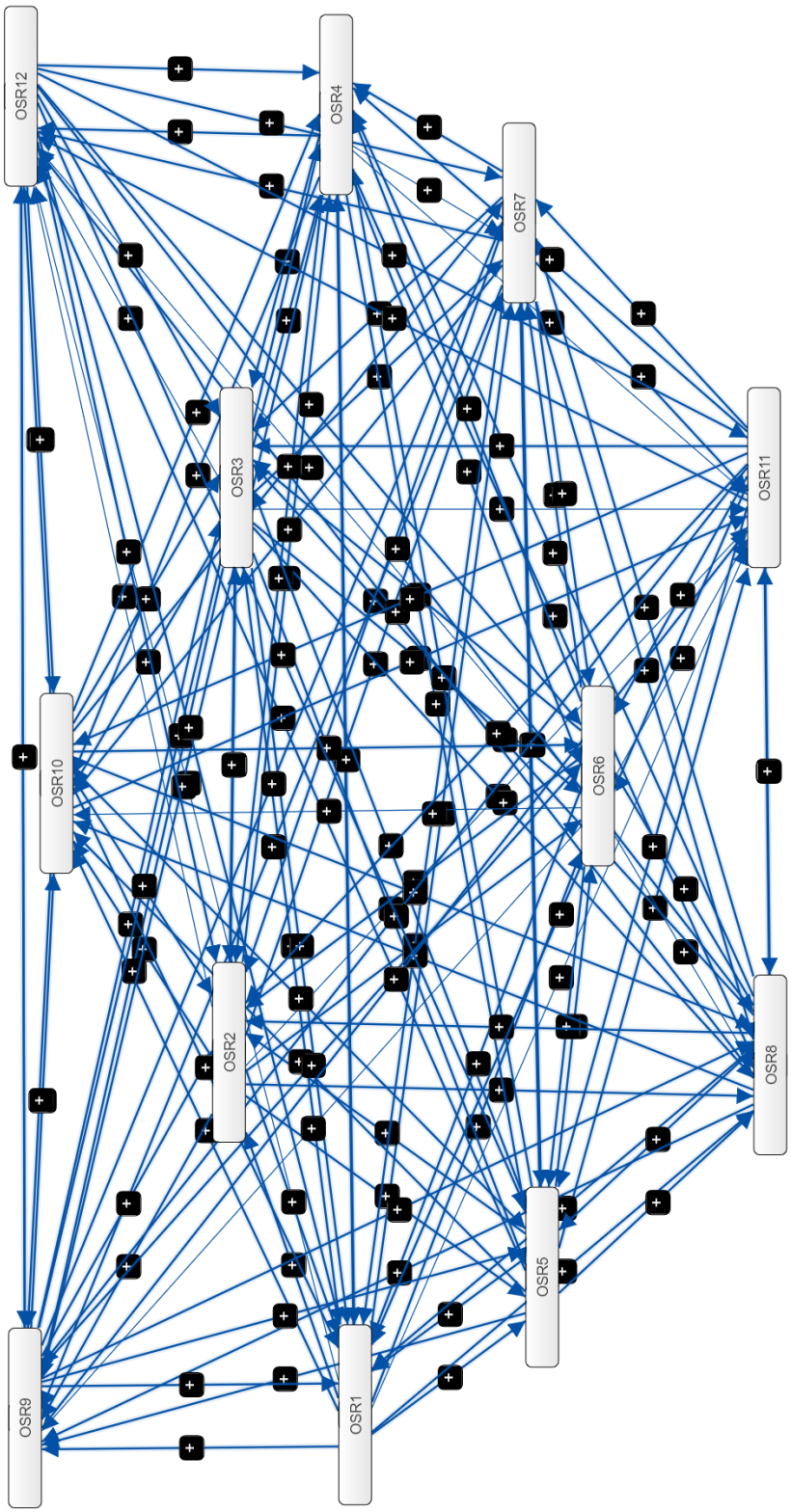


Fig. 3: FCM produced with the MentalModeler software.

– OSR12: Insecure employment undermines employee well-being and hinders airport sustainability
  - Secure Employment Contracts: Offer stable and secure employment contracts with fair compensation and clear career paths to enhance employee well-being and support airport sustainability.
  - Employee Engagement and Empowerment: Foster a culture of employee engagement, involvement, and empowerment through decision-making opportunities, skill development, and recognition, promoting well-being and alignment with sustainability goals.
  - Sustainable Workforce Practices: Implement practices that prioritize work-life balance, manage workloads effectively, and provide resources for stress management, ensuring a healthy and sustainable workforce while addressing occupational stress risks.

4 Conclusions and future directions

Occupational stress poses a substantial risk, particularly in complex work environments, impacting both employee well-being and performance. To gain comprehensive insights into stress-related factors, a meticulous analysis was herein conducted using the FCM approach at an Italian airport. This method allowed for a thorough understanding of the interconnections and causal relationships among various stress-related factors. Specifically, the FCM analysis revealed two primary areas of concern (intensity of work and sustainability) and three main OSRs (OSR9, OSR10, OSR12). In response to the identified risks, a set of mitigation strategies has been proposed, aiming at fostering a healthier and more supportive work environment for airport employees.

Future studies aim to enhance mathematical models for a complete understanding of occupational stress. Advanced modeling would enable the quantification of stressors and the development of targeted interventions while real-time data and computational techniques would enhance stress prediction and the development of personalized interventions.

References

1. Sreenath, S., Sudhakar, K., Yusop, A.F, Sustainability at airports: Technologies and best practices from ASEAN countries Journal of Environmental Management, 299:113639, 2021.
2. Ramakrishnan, J., Liu, T., Yu, R., Seshadri, K., Gou, Z., Towards greener airports: Development of an assessment framework by leveraging sustainability reports and rating tools Environmental Impact Assessment Review, 93:106740, 2022.
3. Kosko, B. (1986). Fuzzy cognitive maps. International journal of man-machine studies, 24(1), 65–75.
4. Poomagal, S., Sujatha, R., Kumar, P. S., & Vo, D.-V. N. (2020). A fuzzy cognitive map approach to predict the hazardous effects of malathion to environment (air, water and soil). Chemosphere, 263, 127926.
5. Wang, Y. M., Yang, J. B., Xu, D. L., & Chin, K. S. (2006). On the centroids of fuzzy numbers. Fuzzy sets and systems, 157(7), 919-926.

Effect of electron space-charge on the gain of a two-dimensional photomultiplier tube model

D. Esperante[1,2], B. Gimeno[1], D. Ginestar[3], D. González-Iglesias[1], J.L. Hueso[3], F. Hueso-González[1], G. Llosá[1], P. Martín-Luna[1], and J. Riera[3]

[1] Instituto de Física Corpuscular (IFIC), CSIC–UV,
c/ Catedrático José Beltrán 2, 46980 Paterna, Spain.
[2] Electronics Engineering Department, Universitat de València,
46100 Burjassot, Spain.
[3] Instituto de Matemática Multidisciplinar, Universitat Politècnica de València,
Camí de Vera, s/n. 46022 València, Spain.

Abstract. Photomultiplier tubes (PMTs) [20] are widely used as photodetectors in military, industrial and medical applications since almost one century. When faced with high-intensity light pulses, PMTs exhibit a non-linear response due to extrinsic factors like voltage supply electronics [9] as well as intrinsic ones, such as electron space-charge: a cloud of densely packed electron trajectories perturbing the electrostatic field in the vacuum of the PMT. To quantify the effect of space charge on the amplification factor (gain) of the PMT, we perform a Monte Carlo simulation of a two-dimensional model of a PMT. Based on an X-ray image of it, we manually contour the dynodes, anode, cathode and glass envelope. The electrostatic field is solved using a two-dimensional mesh after setting the boundary conditions in the faces of the contoured geometries, namely the voltages recommended by the manufacturer on each of the electrodes. The electron trajectories are then calculated by using the Boris leap-frog method [10] and the amplification is based on Vaughan's model [16]. The effect of space-charge is estimated iteratively and without time dependence, by incorporating a charge density into the solution of the electrostatic field and retracing the trajectories. In each step, the charge density is determined by the distribution of electron trajectories found out in the previous iteration. Finally, the effect of the space charge on the PMT gain is studied in terms of the stationary photocathode current for a fixed supply voltage. In the future, the extension of this study to 3D geometry and comparison with experimental measurements are planned.

Keywords: Photomultiplier tubes, space-charge effect, Poisson equation

1 Introduction

A Photomultiplier tube (PMT) is a light sensor that can detect even a single photon by producing an amplified electrical signal of it, and it has been around for

almost one century [12,20]. The use of photomultipliers pervades a wide range of industries, including environmental radiation control, homeland security, nuclear physics, life sciences, medical imaging, and treatment monitoring [7, 8, 19].

A photomultiplier consists of a photocathode followed by an electron multiplier device. Due to the photoelectric effect, electrons are ejected from the photocathode and accelerated into a metallic surface known as dynode by electric forces inside the PMT. Multiple additional electrons are produced as a result of the collision between the electron and the dynode and are then accelerated into the next dynode. A cascaded chain of about ten dynodes results in an electron gain of typically $\approx 10^6$. In this way, a low intensity light entering the photomultiplier results in a measurable current of electrons.

PMTs have been extensively studied in the literature from different points of view, [4,9,11,20], but there are few mathematical models [21] of these devices based on fundamental physics principles. In [13], a Monte Carlo method is proposed taking into account a static electric field computation, and the electron dynamics is modeled using the Lorentz's force associated with this electric field.

It is known that under high-intensity light pulses, PMTs exhibit a non-linear response due to extrinsic factors like voltage supply electronics [9] as well as intrinsic ones, such as electron space-charge: a cloud of densely packed electron trajectories affecting the electric field inside the PMT.

To quantify the effect of space-charge on the amplification factor (gain) of the PMT, we perform a Monte Carlo quasi-static simulation based on a two-dimensional model of the PMT. Following [13], the electrostatic field is solved using a two-dimensional mesh after setting the boundary conditions in the faces of the contoured geometries, namely the voltages recommended by the manufacturer on each of the electrodes. The electron trajectories are then calculated by using the Boris leap-frog method [10] and the amplification is based on Vaughan's model [16].

The effect of space-charge should be computed making use of the Vlasov-Poisson equations [18], but to solve this model in a multidimensional geometry is quite demanding from the computational point of view [3]. Instead, we assume that a quasi-static treatment can give accurate enough results, and the electric field is estimated iteratively, by incorporating a charge density into the solution of the electrostatic field and retracing the trajectories. In this way, the space-charge effect for different light intensities can be evaluated.

2 Physical models

We studied a PMT with 8 dynodes [13], whose geometry is shown in Figure 1.

We designed a 2D model of the PMT corresponding to an axial plane, as shown in Figure 2. To simulate the PMT operation, we track the trajectories of a given number of electrons that start from the cathode and, driven by the electric field, hit on the different dynodes, producing secondary electrons whose trajectories are also tracked until they are trapped in a dynode, they are absorbed

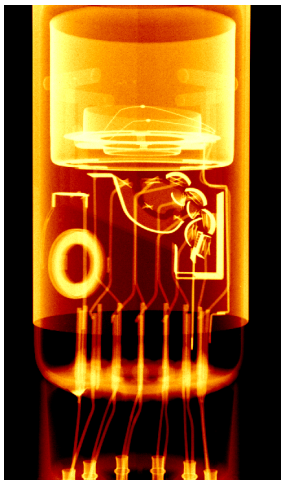
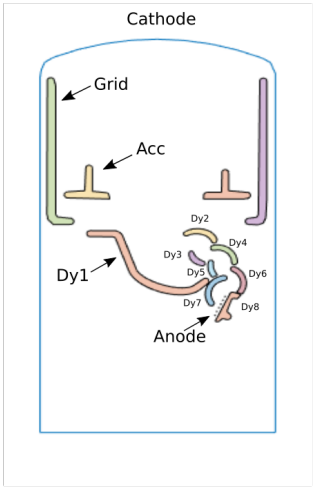
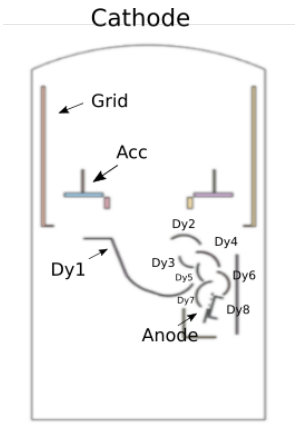


Fig. 1: X-ray image of the photomultiplier tube under study.



(a) Coarse model.



(b) Fine model.

Fig. 2: 2D model of the photomultiplier tube.

or they arrive to the anode. A plot of different electron trajectories is shown in Figure 3.

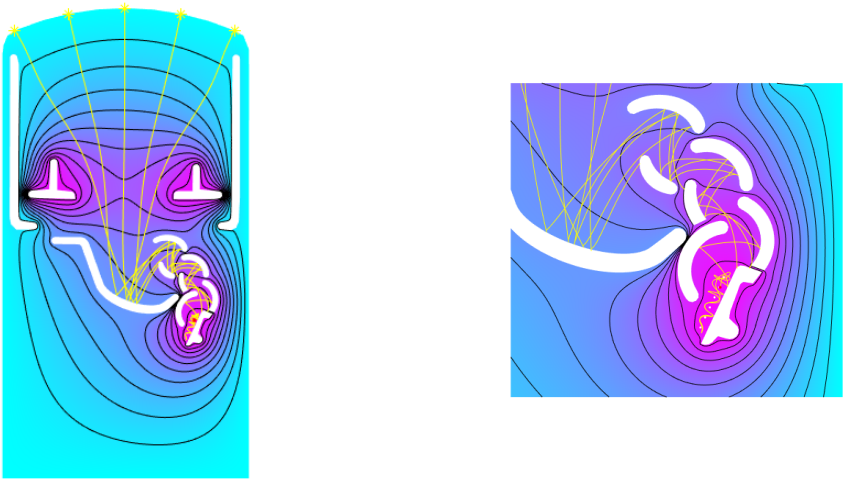


Fig. 3: Different electron trajectories from the cathode to the anode of the photomultiplier tube.

2.1 Electric field computation

To obtain the electric field, we have used the finite element method based on a triangular mesh. In particular, the electrostatic module of the Partial Differential Equations Toolbox of Matlab has been used. In general, this software solves the Poisson equation for the electric potential, V,

$$-\vec{\nabla}\left(\varepsilon\vec{\nabla}V\right) = \rho, \tag{1}$$

where ε is the dielectric permittivity, ρ is the electric charge density and the electric field yields $\vec{E} = -\vec{\nabla}V$.

The boundary conditions are the Dirichlet boundary conditions. The voltage of each dynode is constant and set by the electronics setup [2] and at the outer boundary $V = -1500\text{V}$ is considered.

2.2 Electron dynamics

The dynamics of the electrons in the electric field are governed by the relativistic Second Newton's law associated with Lorentz's force

$$\vec{F} = q\left(\vec{E} + \vec{v}\times\vec{B}\right) = m\frac{d\vec{u}}{dt}\ , \tag{2}$$

where m is the electron rest mass, q is the charge of the electron, $\vec{E}$ and $\vec{B}$ are the electric and magnetic fields, $\vec{v}$ is the velocity vector, $\vec{u} = \gamma\vec{v}$ and γ is the Lorentz factor,

$$\gamma = \frac{1}{\sqrt{1 - \left(\frac{v}{c}\right)^2}},$$

being v the magnitude of $\vec{v}$.

To discretize equation (2), we assume a discrete time step Δt, and $t_n = n\Delta t$, and we make use of Boris method [1, 14], based on the steps,

$$\begin{aligned} \vec{x}_{n+\frac{1}{2}} &= \vec{x}_n + \frac{\vec{u}_n}{2\gamma_n}\Delta t, \\ \frac{\vec{u}_{n+1} - \vec{u}_n}{\Delta t} &= \frac{q}{m}\left(\vec{E}\left(\vec{x}_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}\right) + \vec{v} \times \vec{B}\left(\vec{x}_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}\right)\right) , \\ \vec{x}_{n+1} &= \vec{x}_{n+\frac{1}{2}} + \frac{\vec{u}_{n+1}}{2\gamma_{n+1}}\Delta t \end{aligned} \tag{3}$$

assuming that the particle has an average velocity

$$\vec{v} = \frac{\vec{u}_{n+1} + \vec{u}_n}{2\gamma_{n+\frac{1}{2}}} .$$

2.3 Secondary emission yield

After each step in the electron trajectory, it is checked if it collides with any of the PMT electrodes. When the electrons strike a given electrode (dynode), which is covered with a thin layer of secondary emissive material, if they have sufficient energy, they generate a specific number of secondary electrons. This process is characterized by the total Secondary Electron Yield (SEY) coefficient, δ, that is the average number of electrons emitted per incident one. This coefficient depends on the primary kinetic energy and the incidence angle of the impacting electron, and the modified Vaughan's model has been used for its computation, [15, 17].

Typically, macroparticles (with mass $M = Nm$ and charge $Q = Nq$) are used to represent N electrons. Because the ratio of charge to mass is the same, the trajectory of these macroparticles will be the same as that of the individual electrons if we ignore interactions between electrons.

2.4 Computation of the space-charge effect

A rigorous approach to take into account the effect of the electron charge in the global electric field driving the movement of electrons through the photomultiplier is solving the Vlasov-Poisson equations where a kinetic equation (similar to (2)), for a given distribution of particles, is coupled with the Poisson equation (1), where the charge density ρ depends on the spatial distribution of electrons. But, as it has been already mentioned, this approach is very expensive for a

realistic model, even in a 2D geometry. Thus, we propose to use a quasi-static approach that initially computes the electric field without considering the electron trajectories. With this field, the trajectories of N_i initial particles emitted from the cathode are determined using equations (3), taking into account the secondary electrons generated by the impacts on the different dynodes.

These obtained trajectories are assumed to be stationary and they define a spatial electric charge distribution that is used to recalculate the electric field using again Poisson's equation (1). This process is repeated in an iterative way until the electric field does not change.

3 Numerical results

For the simulations, we have considered two different 2D models of the PMT tube depicted in Figure 2. The second model has more realistic elements requiring a finer mesh for the solution of Poisson's equation.

For each PMT model we have carried out two simulations. In one simulation we assume that the secondary electrons are emitted normally to the dynode contour in the impact point. For this simulation, 320 electrons are traced in order to compute the electric field for the next iteration, and the electric field calculation is repeated twice.

In the other simulation, the emission direction is computed projecting in the plane 3D directions from a $\cos\theta$ distribution (see [5, 6]). In this case, one needs to throw more electrons (1280) and iterate more times (four) in order to get relatively smooth results.

In all cases, the starting points of the trajectories are equispaced spanning practically the whole extent of the cathode.

We repeat these computations for different cathode illuminations, that result in different electron currents I_k in the cathode, in order to analyze their influence on the space-charge effect. The number of electrons arriving to the anode per electron emitted by the cathode is the gain of the PMT. The average gain of the trajectories is computed, first with the electric field from the Poisson's equation without the space charge effect, and then considering the one created by the electrons supposed stationary along their trajectories.

The results are shown in figures 4–7.

Considering orthogonal emission, the finer model yields a slightly superior average gain, 4.6×10^5, than that of the coarser model, 3.6×10^5. With random emission angle, the average gain drops to about 1.35×10^5 for the finer model and 1×10^5 for the coarser one.

The field effect can be observed in Figure 4 corresponding to the coarser model with orthogonal emission, from $I_k = 0.0316\,\mathrm{A}$ on. In the other cases, there is no significant difference in the average gain for different values of I_k.

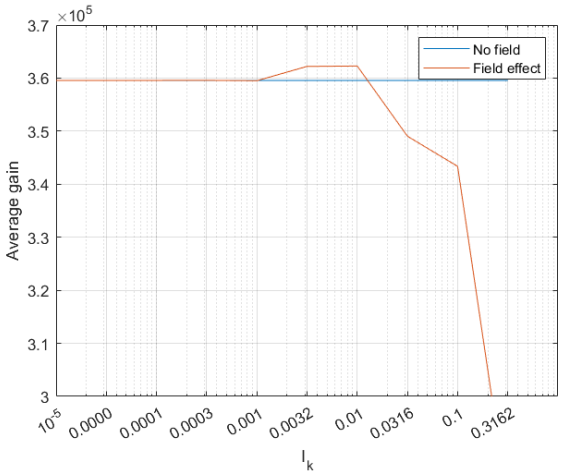


Fig. 4: Average gain for 320 throws with the coarse PMT model. Orthogonal emission.

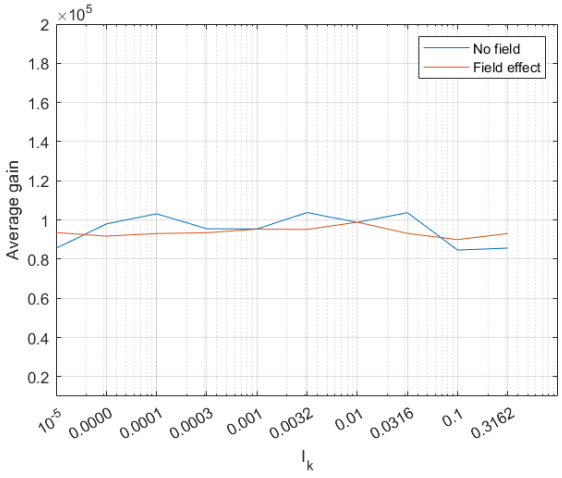


Fig. 5: Average gain for 1280 throws with the coarse PMT model. Angle emission $\cos\theta$-distributed.

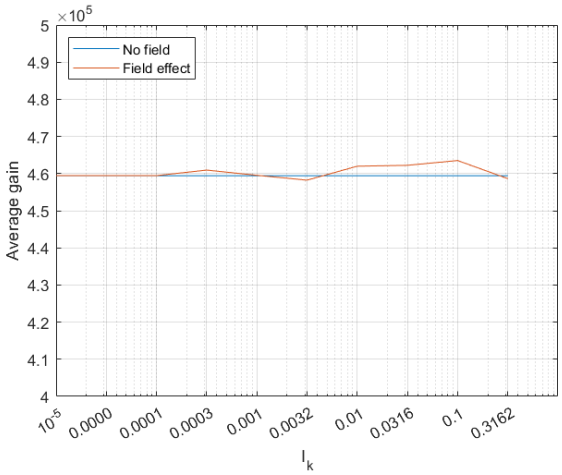


Fig. 6: Average gain for 320 throws with the fine PMT model. Orthogonal emission.

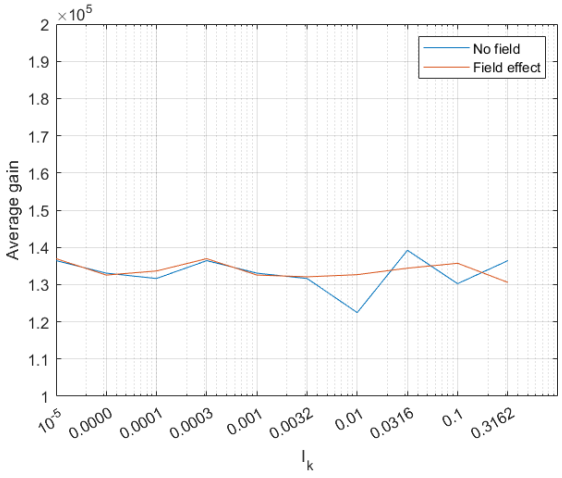


Fig. 7: Average gain for 1280 throws with the fine PMT model. Angle emission $\cos\theta$-distributed.

4 Conclusions

We have developed a quasi-static model to take into account the space charge effect in a photomultiplier device based on the correction of the electric field when the charge of stationary trajectories of electrons is considered. The gain of the PMT has been computed for different illuminations and different emission models in two 2D geometries. The obtained results show that only for unrealistic illuminations the space charge effect has to be taken in consideration for the studied PMT.

References

1. Jay P Boris et al. Relativistic plasma simulation-optimization of a hybrid code. In Proc. Fourth Conf. Num. Sim. Plasmas, pages 3–67, 1970.
2. A Fernández Prieto, A Gallas Torreira, I García Rivas, F Hueso-González, E Lemos Cid, G. Llosá, A Pazos álvarez, E. P´erez Trigo, and P. Vázquez Regueiro. Pmt system for prompt gamma-ray measurements during proton therapy treatments. In IEEE Nucl Sci Symp Conf Rec, 2022.
3. Francis Filbet and Chang Yang. Numerical simulations to the vlasov-poisson system with a strong magnetic field. arXiv preprint arXiv:1805.10888, 2018.
4. SO Flyckt and C Marmonier. Photomultiplier tubes: principles and applications, 2nd Ed. https://www2.pv.infn.it/debari/doc/FlycktMarmonier.pdf, 2002.
5. M. A. Furman and M. T. F. Pivi. Probabilistic model for the simulation of secondary electron emission. Phys. Rev. ST Accel. Beams, 5:124404, Dec 2002.
6. John Greenwood. The correct and incorrect generation of a cosine distribution of scattered particles for monte-carlo modelling of vacuum systems. Vacuum, 67(2):217–222, 2002.
7. K K Hamamatsu-Photonics. Photomultiplier tubes. Basics and applications. 4th Ed. https://www.hamamatsu.com/resources/pdf/etd/PMThandbookv4E.pdf, 2017.
8. Fernando Hueso-González, Moritz Rabe, Thomas A Ruggieri, Thomas Bortfeld, and Joost M Verburg. A full-scale clinical prototype for proton range verification using prompt gamma-ray spectroscopy. Physics in Medicine & Biology, 63(18):185019, 2018.
9. Fernando Hueso-González, Damián Ginestar, Jos´e L. Hueso, and Jaime Riera. Comments on "spice model of photomultiplier tube under different bias conditions". IEEE Sens J, 2021.
10. J. P. Boris. Relativistic plasma simulation-optimization of a hybrid code. In Proc. Fourth Conf. Numerical Simulations of Plasmas (Washington, D.C.: Naval Research Laboratory), pages 3–67, 1970.
11. Natan Krihely. Spice model of photomultiplier tube under different bias conditions. IEEE Sensors Journal, 14(10):3606–3610, 2014.
12. Bayarto K Lubsandorzhiev. On the history of photomultiplier tube invention. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 567(1):236–238, 2006.
13. P. Martín-Luna, D Esperante, A A. Fernández Prieto, N Fuster-Martínez, I García Rivas, B. Gimeno, D Ginestar, D. González-Iglesias, J L Hueso, G. Llosá, P Martinez-Reviriego, J. Riera, P Vázquez Regueiro, and F. Hueso-González. Simulation of electron transport and secondary emission in a photomultiplier tube. IEEE Sensors Journal, Submitted.

14. Bart Ripperda, Fabio Bacchini, Jannis Teunissen, Chun Xia, Oliver Porth, Lorenzo Sironi, Giovanni Lapenta, and Rony Keppens. A comprehensive comparison of relativistic particle integrators. The Astrophysical Journal Supplement Series, 235(1):21, 2018.
15. J Rodney M Vaughan. A new formula for secondary emission yield. IEEE Transactions on electron devices, 36(9):1963–1967, 1989.
16. C. Vicente, M. Mattes, D. Wolk, B. Mottet, H.L. Hartnagel, J.R. Mosig, and D. Raboso. Multipactor breakdown prediction in rectangular waveguide based components. In IEEE MTT-S International Microwave Symposium Digest, 2005, pages 1055–1058, 2005.
17. Carlos Vicente, Michael Mattes, Dieter Wolk, B Mottet, HL Hartnagel, JR Mosig, and D Raboso. Multipactor breakdown prediction in rectangular waveguide based components. In IEEE MTT-S International Microwave Symposium Digest, 2005., pages 1055–1058. IEEE, 2005.
18. Anatoli Aleksandrovich Vlasov. The vibrational properties of an electron gas. Soviet Physics Uspekhi, 10(6):721, 1968.
19. Theresa Werner, Jonathan Berthold, Fernando Hueso-González, Toni Koegler, Johannes Petzoldt, Katja Roemer, Christian Richter, Andreas Rinscheid, Arno Straessner,Wolfgang Enghardt, et al. Processing of prompt gamma-ray timing data for proton range measurements at a clinical beam delivery. Physics in Medicine & Biology, 64(10):105023, 2019.
20. A. G. Wright. The Photomultiplier Handbook. Oxford University Press, 2017.
21. Yao Zhu, Sen Qian, Qi Wu, Guoqing Zhang, Lishuang Ma, and Zhile Wang. Study on fast timing mcp-pmt in magnetic fields from simulation and measurement. Sensors and Actuators A: Physical, 318:112487, 2021.

Mathematical modeling for the analysis of thermo-optic response of the cranial implant Window to the Brain

Mildred S. Cano-Velázquez[1], Jose Bon[2] M. Llamazares[3]*, Santiago Camacho-López[4], Guillermo Aguilar[5], Juan Hernández-Cordero[1], and Macarena Trujillo[3]

[1] Instituto de Investigaciones en Materiales, Universidad Nacional Autónoma de México, Ciudad de México, México,
[2] Food Technology Department, Universitat Politècnica de València, Valencia, Spain
[3] Department of Applied Mathematics, BioMIT, Universitat Politècnica de València, Valencia, Spain,
marllalo@upvnet.upv.es
[4] Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Ensenada, B.C., México
[5] J. Mike Walker 66 Department of Mechanical Engineering, Texas A&M University, College Station, TX, USA

Abstract. Poly-crystalline zirconia based ceramics (nc-YSZ) have suitable thermo-optic and biomedical properties for the implementation of biomedical transparent cranial implant, called Window to the Brain (WttB). This implant has the advantage of allowing optical access to the brain for medical treatments with laser sources. Previous works have determined the physic parameters of the nc-YSZ, and experimental data for bacteria anti-fouling via thermal methods using laser sources have been obtained. In this work, a thin layer of copper nanoparticles has been added underneath the WttB. Mathematical models with the theorethical physical properties of nanoparticles are made for this configuration. The aim is to use the nanoparticles properties to perform bacteria anti-fouling with a lesser penetration of the temperature increments in the brain tissue. Preliminary computational results are obtained, comparing the configuration with and without the nanoparticles layer, using finite elements resolution.

Keywords: nanocrystalline yttria stabilized zirconia, cranial implant, Window to the brain, heat conduction, photothermal effect, laser sources, copper nanoparticles, computer modeling

1 Introduction

The good thermo-optic and biomedical properties of poly-crystalline zirconia based ceramics (nc-YSZ) have been widely studied [1], [2], [3], [4], [5] in order to make them suitable for the implementation of biomedical implants. In particular the transparent cranial implant called Window to the Brain (WttB) [6],

[7], [8], [9], [10]. This implant has the advantage of allowing optical access to the brain for medical treatments with laser sources, which can vary from diagnosis to surgical interventions for tumor destruction [11].

Despite its good properties, the WttB implant has some problems to solve in order to make a viable biomedical implant, one of the most important being the bacteria proliferation underneath the WttB. Bacteria proliferation is a common issue in biomedical implants, but for the WttB is critical due to its proximity to the brain. Thus, is mandatory to perform bacteria anti-fouling. Thermal methods using laser sources have been explored, and experimental data have been obtained [12], [13].

In this work, a new configuration is tested, where a thin layer of copper nanoparticles has been added underneath the WttB. The aim is to use the nanoparticles properties to perform bacteria anti-fouling with a lesser penetration of the temperature increments in the brain tissue. The objective is to obtain $\Delta T \geq 10\ ^oC$ in the bacteria for anti-fouling, while $\Delta T < 10\ ^oC$ in the brain in order to avoid neural degradation.

Mathematical model presented in [14] is extended to this configuration. The physic parameters of the nc-YSZ have been determined using experimental measurements and optimization of mathematical models [15], [14], [16]. Physical properties of copper nanoparticles layer are unknown, with no experimental measurements performed yet. A theoretical study of these properties is performed in this work, and preliminary computational results are obtained in order to test the viability of this configuration.

2 Mathematical model

Heat equation (1) is used for describing the evolution of temperature increases [17] in the mathematical model, and Beer-Lambert equation (2) is used for describing the energy absorption of the laser irradiation from the material [18]. Heat and Beer-Lambert equations are coupled by the heat source term (3);

$$\rho c \frac{\partial T}{\partial t} = \nabla(k \nabla T) + Q \tag{1}$$

where T is the temperature, ρ is the material density, c is the thermal capacity, k is the thermal conductivity and Q is the external heat source.

$$\frac{\partial I}{\partial z} = -\alpha I \tag{2}$$

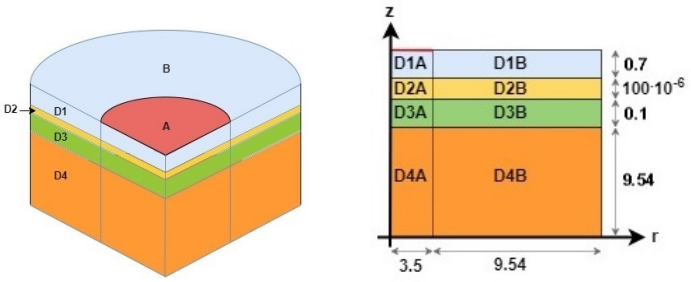
where I is the intensity function, α is the absorption coefficient of the material and z is the spatial variable.

$$Q = \frac{\partial I}{\partial z} \tag{3}$$

These equations are applied to the geometry described in Figure 1, which is a multi-layer disc with layers of WttB, nanoparticles, bacteria and agar; divided in two concentric domains, A and B. Heat equation applies to domain A and B, while Beer-Lambert equation and heat source apply to domain A.

Regarding the thermal problem, the boundary conditions are natural heat convection $\mathbf{n_1} \cdot \mathbf{q} = h \cdot (T_a - T)$ for the top surface of the D1 layer, null flux for the edges of layers D1 to D4 and constant temperature T_0=37°C of the human body for the bottom surface of the D4 layer.

Regarding the optical problem, the boundary conditions are null flux for the boundary between domains A and B, and gaussian intensity distribution of the laser source for the top surface of domain A $I(r) = I_0 \cdot (1 - R) \cdot e^{-r^2}$, where R is the reflection coefficient and r is the radial spatial coordinate.



(a) (b)

Fig. 1: Geometric representation of the multi-layer disc with axis-symmetric geometry, where D1 is the layer of WttB, D2 is the layer of copper nanoparticles, D3 is the layer of bacteria and D4 is the layer of agar. A is the irradiated domain and B the non-irradiated domain. (Spatial dimensions in mm and schemes represented out of scale).

3 Matherial properties

3.1 Known properties

The known physical properties of the Wttb and the agar, introduced in the mathematical model as parameters, are listed in Table 1. Note bacteria layer has the same physical properties as the agar, since bacteria layer composition is bacteria suspended in the agar.

Table 1: Properties of the WttB [14] [16] and the Agar [12] [13].

Parameter	Symbol	WttB Value	Agar Value
Initial temperature [ºC]	T_0	37	37
Constant temperature [ºC]	T_c	[-]	37
Room temperature [ºC]	T_a	23	23
Thermal conductivity [W/m·K]	k	2.13	0.6
Density [kg/m^3]	ρ	6042	4182
Heat capacity [J/kg·K]	c	427	998
Convection coefficient [W/m^2·K]	h	11	[-]
Absorption coefficient [m^{-1}]	α	$-3.851 \cdot 10^3 + 14.54 \cdot T$	2
Reflection coefficient [%]	R	0.19	[-]

3.2 Nanoparticles properties

Thermal capacity It has been shown experimentally that the thermal capacity of nanoparticles differs from the macroscopic thermal capacity [19]. This variation of the physical properties can be explained due to the effects of surface atoms [20].

The atoms of a material can be classified as surface atoms, these being the atoms corresponding to the exterior boundary of the material, and interior atoms, these being the atoms enveloped by the surface atoms. In a macroscopic material, the layer of surface atoms is negligible with respect to the interior atoms. However, for a thin layer of nanoparticles, the surface atoms constitute a significant percentage of the total material atoms.

This classification of atoms is relevant because the amplitude of vibration of surface atoms with temperature is greater than that of interior atoms, while the frequency of vibration of surface atoms is less than that of interior atoms. These properties can be determined theoretically using the Debye temperature [21].

Due to the vibrational properties of the surface atoms, they increase the thermal capacity of the material. Therefore, the smaller the size of the nanoparticles, the greater the percentage of surface atoms, and the greater the thermal capacity with respect to the macroscopic case. [22].

An experimental fit equation is purposed in [20] for the relation between the thermal capacity of nanoparticles c_n and macroscopic materialc_m.

$$c_n = c_m \left(1 - K_0 \frac{d}{D}\right) \tag{4}$$

where $K_0 = -0.5$ is a fiting constant, D is the diameter of the nanoparticles y d is the atomic diameter. This equation is represented in Figure 2.

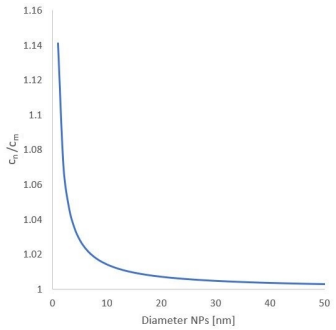


Fig. 2: Comparison between thermal capacity of nanoparticles c_n and macroscopic material c_m as a function of nanoparticles diameter.

Thermal Conductivity Thermal conductivity in metals occurs due to the transport of energy by free electrons. Its values is constant for macroscopic materials. In the case of nanoparticles, according to [20], [23], [24] thermal conductivity k_n can be calculated using the following equation

$$k_n = \frac{k_m(1-\delta\frac{4f_0d}{\eta D})^{1/2}}{1+\frac{R_k}{D}k_m(1-\delta\frac{4f_0d}{\eta D})^{1/2}} \tag{5}$$

where k_m is the thermal conductivity of the macroscopic material, D is the diameter of the nanoparticle, $d = 0.2822$ nm [25] is the diameter of the atoms of the material, $\delta \in (0,1)$ is the relaxation factor, $f_0 = 1$ is the shape factor, $\eta = 0.68$ is the atomic packing fraction and R_k is the kapitza resistance [26], [27]. Numerical values are particularized to spherical copper nanoparticles. With this numerical values, equation (5) can be approximated to

$$k_n \simeq \frac{k_m}{1+\frac{R_k}{D}k_m} \tag{6}$$

where for any value of δ, variations between equations (5) and (6) are less than 0.1%.

Equation (6) is represented in the Figure 3, where copper thermal conductivity is $k_m = 385$ W/(m/K) [25] and Kapitza resistance $R_k = 0.3 \cdot 10^{-9}$ (m^{2}K)/W. It can be observed how as the diameter of the nanoparticles increases, the thermal conductivity approaches its macroscopic value.

Density. Nanoparticles density is considered equal to the density of the material on a macroscopic scale [25].

Surface Plasmon Resonance. Plasmons, in the field of matter physics, refer to the collective vibration of free electrons in metals. These vibrations possess a

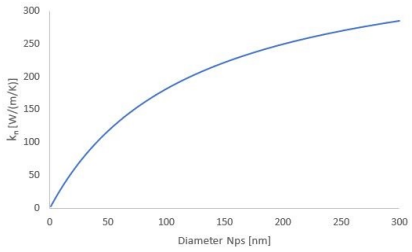


Fig. 3: Thermal conductivity of copper nanoparticles as a function of their diameter.

distinct frequency and thus, plasmons can be categorized as an excitation of a bosonic quasi-particle [29].

Plasmons may be excited by electromagnetic radiation, such as coherent irradiation from a laser [30], [31], [32], [33]. However, the effects of this excitation become negligible in metallic materials on a macroscopic scale, since the excitation, present on the surface of the metal, is dispersed inside it.

If the size of the nanoparticles is much smaller than the wavelength of the incident photons, the excitation does not propagate, thus obtaining what is known as localized surface plasmons. The resulting plasmonic oscillation is distributed over the entire volume of the nanoparticle [34].

When the frequency of the photons of the incident irradiation corresponds to the resonance frequency of the described system, the plasmons enter into resonance, generating very energetic oscillation peaks that result in the absorption of a large part of the intensity of the incident light.

This effect is known as surface plasmon resonance, and from an optical point of view, it produces peaks in the absorption coefficient of metallic nanoparticles for the characteristic resonance frequency of each metal [29], [30], [32]. In the case of copper nanoparticles, this frequency corresponds to a wavelength of 580 nm [32], [33].

In the mathematical model proposed by this work, due to the coupling between the optical problem and the thermal problem, increases in the absorption coefficient imply an increase of the effect from the external heat source term in the heat equation, which induces localized temperature increases in the nanoparticle layer, an effect that is desirable for the control of the bacterial population by thermal methods.

Absorption Coefficient. The absorption coefficient of a material can be determined experimentally [18] by measurements of the extinction coefficient ϵ and

the absorbance A. These measurements are made using solutions of nanoparticles in liquids with known optical properties, in containers whose properties are also known, such as quartz.

Absorbance determines the amount of light intensity that is absorbed by a sample solution and is defined by equation (7)

$$A = -log_{10}\left(\frac{I}{I_0}\right) \tag{7}$$

where I_0 is the incident intensity and I is the intensity measured after the light passes through the solution.

The correlation between extinction coefficient and absorbance is determined by equation (8)

$$A = \epsilon M L \tag{8}$$

where M is the molar concentration or molarity of the solution and L optical path length.

Absorption and extinction coefficients correlation is known

$$\alpha = \frac{4\pi\epsilon}{\lambda} \tag{9}$$

where λ is the incident wavelength.

And finally we can relate the absorbance and the absorption coefficient combining equations (2), (7) and (9)

$$\alpha = 2,303\frac{A}{L} \tag{10}$$

Using the experimental measurements presented in [33] to determine the relationship between absorbance and temperature, as well as measurements of the extinction coefficient and absorbance for different diameters of nanoparticles [32]; an expression of $\alpha(T)$ can be obtained, for irradiations at the characteristic frequency of copper corresponding to a wavelength of 580 nm, with a linear dependence on temperature ($R^2 = 0.975$), for copper nanoparticles particularized to a diameter of 20 nm

$$\alpha(T) = 2.61 \cdot 10^7 + 5655.6 \cdot T \tag{11}$$

where it has been taken into account the hypothesis that the functional dependence of the absorption coefficient of the nanoparticles with the temperature does not vary significantly for different diameters, and the value of the absorption coefficient of a nanoparticle can be extended to a set of nanoparticles. Dispersion phenomena are not considered.

4 Computational results

The numerical results of this work must verify the pre-set objectives of obtaining $\Delta T \geq 10\ ^oC$ in the bacteria layer for bacteria anti-fouling and $\Delta T < 10\ ^oC$ in the agar layer representing the brain, in order to avoid neural degradation.

Table 2 shows the results of the ΔT in the bacteria layer, obtained computationally for the numerical resolution with finite elements. Intensities I_2 (2W) and I_3 (3W) produce $\Delta T \geq 10\ ^oC$ in the bacteria layer.

Figure 4 shows the two-dimensional temperature distribution for an axisymmetric slice of the four-layer disk. Temperature distributions obtained from the irradiation of the laser sources I_2 and I_3 are compared, for the case with and without nanoparticles (NPs). Moreover, the isotherm for $\Delta T = 10\ ^oC$ is also represented, so that the penetration of this ΔT into the agar layer for each case study can be identified.

Table 3 displays the numerical values for the maximum penetration length L_{pm} del $\Delta T = 10\ ^oC$ in the agar layer. These values are measured on the axis of symmetry, $r = 0$, since it is on this axis where the maximum temperatures are reached.

Table 2: Computational results ΔT for each irradiation source for the case with NPs. Temperature measured at the center of the upper surface of the bacterial layer at 20 seconds.

	ΔT Superior Bacteria	ΔT Inferior Bacteria
I_1: 1 W	5.38	5.03
I_2: 2 W	11.77	10.78
I_3: 3 W	18.42	16.82

Table 3: Comparative of the penetration of the isotherm $\Delta T = 10\ ^oC$ in the agar layer for the case with and without NPs. Maximum penetration length L_{pm} measured on the symmetry axis.

	Without NPs L_{pm} $[mm]$	With NPs L_{pm} $[mm]$
I_2: 2 W	0.71	0.07
I_3: 3 W	1.31	0.56

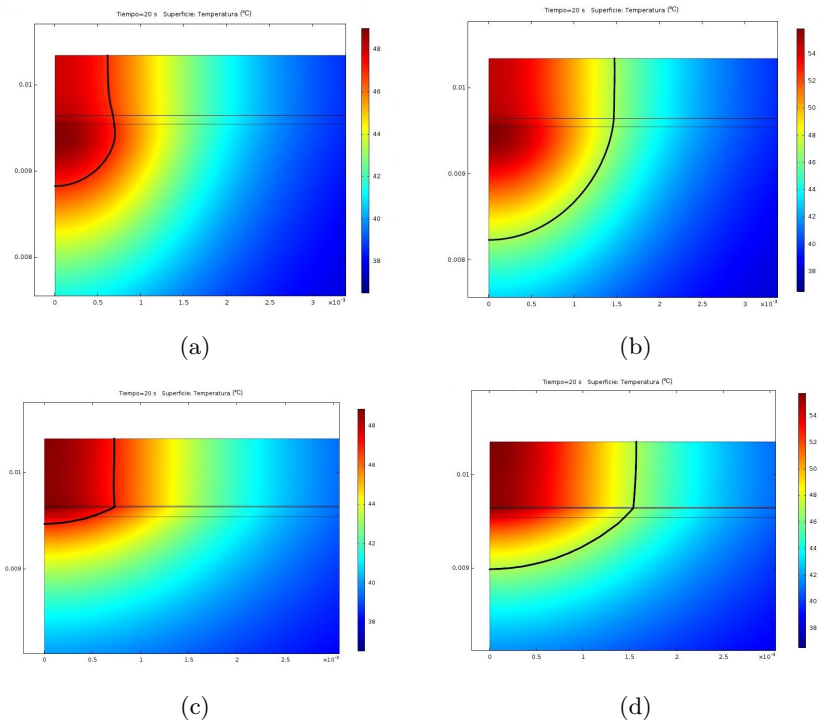


Fig. 4: Temperature distribution and isotherm of $\Delta T = 10\ ^oC$ ((black line) for the cases with and without nanoparticles at 20 seconds. Spatial dimensions in mm, temperatures in ^oC. (a) and (b) represent the results without NPs for I_2 (2W) and I_3 (3W) respectively. (c) and (d) the results with NPs for I_2 and I_3.

5 Conclusion

The multi-layer model of the WttB implant with a thin layer of nanoparticles has promising results with regard to reducing the penetration of increased temperatures into brain tissue, due to bacterial anti-fouling procedures.
A notable reduction of the penetration of the temperature increases for the different laser sources studied is appreciated.
These preliminary results motivate the production of the implant with the nanoparticle layer and experimental measurements in the laboratory, to confirm the theoretical properties of the nanoparticles and the thermal behavior of the proposed configuration.

References

1. S.F. Wang, J. Zhang, D.W. Luo, F. Gu, D.Y. Tang, Z.L. Dong, G.E. Tan, W.X. Que, T.S. Zhang, S. Li, and L.B. Kong: Transparent ceramics: Processing, materi-

als and applications. Progress in Solid State Chemistry. https://doi.org/10.1016/j.progsolidstchem.2012.12.002
2. S.R. Casolco, J. Xu, and J E. Garay: Transparent/translucent polycrystalline nanostructured yttria stabilized zirconia with varying colors. Scripta Materialia 58(6), 516-519 (2008). https://doi.org/10.1016/j.scriptamat.2007.11.014
3. Y. Xiong, Z. Fu, V. Pouchly, K. Maca, and Z. Shen: Preparation of Transparent 3 Y-TZP Nanoceramics with No Low-Temperature Degradation. Journal of the American Ceramic Society 97(5), 1402-1406 (2014). DOI:10.1111/jace.12919
4. A. Rosenflanz, M. Frey, B. Endres, T. Anderson, E. Richards, and C. Schardt: Bulk glasses and ultrahard nanoceramics based on alumina and rare-earth oxides. Nature 430(7001), 761-764 (2004). https://doi.org/10.1038/nature02729
5. U. Anselmi-Tamburini, J.N. Woolman, and Z.A. Munir: Transparent Nanometric Cubic and Tetragonal Zirconia Obtained by High-Pressure Pulsed Electric Current Sintering. Advanced Functional Materials 17(6), 3267-3273 (2007). https://doi.org/10.1002/adfm.200600959
6. J. Hernández-Cordero, R. Ramos-García, S. Camacho-López, and G. Aguilar: A New Window for Photonics in the Brain. IEEE Photonics Society Newsletter, 32(1), (2018).
7. Y. Damestani, C.L. Reynolds, J. Szu, M.S. Hsu, Y. Kodera, D.K. Binder, B.H. Park, J.E. Garay, M.P. Rao, and G. Aguilar: Transparent nanocrystalline yttria-stabilized-zirconia calvarium prosthesis. Nanomedicine: Nanotechnology, Biology and Medicine 9(8), 1135-1138 (2013). https://doi.org/10.1016/j.nano.2013.08.002
8. N. Davoodzadeh, M.S. Cano-Velázquez, D.L. Halaney, C. Jonak, D.K. Binder, and G. Aguilar: Evaluation of a transparent cranial implant as a permanent window for cerebral blood flow imaging. Biomedical Optics Express 9(10), 4879-4892 (2018). DOI: 10.1364/BOE.9.004879
9. N. Davoodzadeh, M.S. Cano-Velázquez, D.L. Halaney, C.R. Jonak, D.K. Binder, and G. Aguilar: Optical access to arteriovenous cerebral microcirculation through a transparent cranial implant. Lasers in Surgery and Medicine 51(10), 920-932 (2019). DOI: 10.1002/lsm.23127
10. D.L. Halaney, C.R. Jonak, J. Liu, N. Davoodzadeh, M.S. Cano-Velázquez, P. Ehtiyatkar, H. Park, D.K. Binder, and G. Aguilar: Chronic brain imaging across a transparent nanocrystalline yttria-stabilized-zirconia cranial implant. Frontiers in Bioengineering and Biotechnology 8, 659 (2020). https://doi.org/10.3389/fbioe.2020.00659
11. I. Mellal, A. Oukaira, E. Kengene, and A. Lakhssassi: Thermal therapy modalities for cancer treatment: A review and future perspectives. International Journal of Applied Science, 4(2), 14 (2017). DOI:10.21767/2394-9988.100064
12. W.L. Titsworth, G.J. Murad, B.L. Hoh, and M. Rahman: Fighting fire with fire: the revival of thermotherapy for gliomas. Anticancer Research, 34(2), 565-574 (2014). PMID: 24510985
13. Y. Damestani, N. De Howitt, D.L. Halaney, J.E. Garay, and G. Aguilar: Evaluation of laser bacterial anti-fouling of transparent nanocrystalline Yttria-Stabilized-Zirconia cranial implant. Lasers in Surgery and Medicine, 48(8), 782-789 (2016). doi: 10.1002/lsm.22558
14. Mildred S.Cano-Velázquez, Jose Bon, M.Llamazares, Santiago Camacho-López, Guillermo Aguilar, Juan Hernández-Cordero, MacarenaTrujillo: Experimental and computational model approach to assess the photothermal effects in transparent nanocrystalline yttria stabilized zirconia cranial implant. Computer Methods and Programs in Biomedicine 221 (2022) 106896. https://doi.org/10.1016/j.cmpb.2022.106896

15. J.E. Alaniz, F.G. Pérez-Gutierrez, G. Aguilar, and J.E. Garay: Optical properties of transparent nanocrystalline yttria stabilized zirconia. Optical Materials 32(1), 62–68 (2009). https://doi.org/10.1016/j.optmat.2009.06.004
16. S. Ghosh, D. Teweldebrhan, J.R. Morales, J.E. Garay, and A.A. Balandin: Thermal properties of the optically transparent pore-free nanostructured yttria stabilized zirconia. Journal of Applied Physics, 106,113507 (2009). https://doi.org/10.1063/1.3264613
17. McAdams, Williams H: Heat transmission. McGraw-Hill, New York, 1954 .
18. Nikolai V.Tkachenko: Optical Spectroscopy Methods and Instrumentations. Elsevier B.V. 2006
19. J. Rupp and R. Birringer: Phys. Rev. B: Condens. Matter. 987, 36, 7888.
20. Shiyun Xiong, Weihong Qi, Yajuan Cheng, Baiyun Huang, Mingpu Wang, Yejun Li: Universal relation for size dependent thermodynamic properties of metallic nanoparticles. Phys. Chem. Chem. Phys., 2011, 13, 10652–10660 (2011).
21. C. Kittel: Introduction to Solid State Physics. John Wiley and Sons, 8th edn, 2004.
22. Wenhua Luo, Wangyu Hu, and Shifang Xiao: Size Effect on the Thermodynamic Properties of Silver Nanoparticles. J. Phys. Chem. C 2008, 112, 2359-2369.
23. Ratan Lal Jaiswal, Brijesh Kumar Pandey: Modelling for the variation of thermal conductivity of metallic nanoparticles. Physica B 627 (2022) 413594.
24. H.S. Yang, G.R. Bai, L.J. Thompson, J.A. Eastman: Interfacial thermal resistance in nanocrystalline yttria-stabilized zirconia. Acta Mater. 50 (2002) 2309–2317.
25. C. Kittel: Introduction to Solid State Physics John Wiley and Sons, 8th edn, 2004.
26. R.C. Johnson, W.A. Little: Experiments on kapitza resistance. Phys. Rev. J. Archive 130 (1963) 596.
27. J.D.N. Cheeke, B. Hebral, J. Rechard: Kapitza resistance between the transition metals Fe, Co, and Ni and superfluid helium. J. Low Temp. Phys. 12 (3/4) (1973) 359–373.
28. Arijit Kumar Chatterjee, Ruchira Chakraborty and Tarakdas Basu: Mechanism of antibacterial activity of copper nanoparticles. Nanotechnology 25 (2014) 135101 doi:10.1088/0957-4484/25/13/135101
29. Vincenzo Amendola: Surface plasmon resonance in gold nanoparticles: a review. J. Phys.: Condens. Matter (2017) 29 203002
30. V. Liberman, M. Sworin, R. P. Kingsborough, G. P. Geurtsen, M. Rothschild: Nonlinear bleaching, absorption, and scattering of 532-nm-irradiated plasmonic nanoparticles. J. Appl. Phys. 113, 053107 (2013) https://doi.org/10.1063/1.4790798
31. Junqi Tang, Kunpeng Gao, Quanhong Ou, Xuewen Fu, Shi-Qing Man, Jie Guo, Yingkai Liu: Calculation extinction cross sections andmolar attenuation coefficient of small gold nanoparticles and experimental observation of their UV–vis spectral properties. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 191 (2018) 513–520
32. Oleg A. Yeshchenko, Igor M. Dmitruk a, Andriy M. Dmytrukb, Alexandr A. Alexeenko: Influence of annealing conditions on size and optical properties of copper nanoparticles embedded in silica matrix. Materials Science and Engineering B 137 (2007) 247–254
33. O.A. Yeshchenko: Temperature Effects on the Surface Plasmon Resonance in Copper Nanoparticles. ISSN 2071-0186. Ukr. J. Phys. 2013. Vol. 58, No. 3
34. Dean J. Campbell, Younan Xia: Plasmons: Why Should We Care? Journal of Chemical Education, Vol. 84 No. 1 January 2007, www.JCE.DivCHED.org

Maximum-Likelihood Expectation-Maximization method applied to unfold neutron spectra in a radiotherapy bunker

S. Oliver, B. Juste, R. Miró and G. Verdú

Instituto de Seguridad Industrial, Radiofísica y Medioambiental (ISIRYM),
Universitat Politècnica de València, València, Spain,
sanolgi@upvnet.upv.es

Abstract. The Maximum-Likelihood Expectation-Maximization (MLEM) method is an iterative computation of maximum-likelihood estimation widely used to solve ill-posed problems. An example of this, in the field of medical physics, is the neutron spectrum unfolding in a radiotherapy bunker via the raw data measured with a Bonner Sphere Spectrometer (BSS). This device consists of a central thermal neutron detector and six high-density polyethylene spheres with different diameters. Each sphere moderator is inserted in the detector, being sensitive to different energy neutrons. Each sphere-detector combination has a unique response function depending on the neutron energy. The relation with neutron spectrum, and the response function is given by the Fredholm integral equation of the first kind. Since in the described problem the number of measurements carried out, one for each sphere, is smaller than the number of energy bins of the unknown spectrum, 29 bins, the problem has an infinite number of mathematical solutions, some of them without a physical sense. By this reason, the MLEM method is suitable to obtain an accurate neutron spectrum. In this work, the neutron spectra at different locations of a real radiotherapy bunker have been obtained applying the MLEM with the aim to measure the neutron ambient dose produced in a radiotherapy treatment.

Keywords: neutron spectrum, MLEM, BSS, LinAc

1 Introduction

Radiotherapy is one of the widely used therapy for cancer. These treatments consist of several external beams delivered by a medical linear accelerator (LinAc), targeted to destroy cancer cells while sparing surrounding healthy tissues. LinAcs can generate photon and electron beams with different energies. High-energy photons (above 8 MeV), generated when the LinAc is working above 10 MV, can produce secondary neutrons during the treatment by photonuclear interactions [1, 2]. The problem with these neutrons is that they can contribute to the patient absorbed dose and can induce activation of different materials inside the treatment room. These facts expose patients and radiotherapy personnel to carcinogenic risk [3, 4], [5, 6].

The accurate determination of the neutron spectrum is crucial due to the strong dependence of the carcinogenic risk associated with secondary neutrons on their energy [7]. The biological effects of ionizing radiation on the human body are measured using the equivalent dose, which assigns a high weighting factor (ω_r) for neutrons and α particles. For neutrons, ω_r values can range from 5 to 20, resulting in equivalent doses up to 20 times larger than the absorbed dose, depending on the neutron energy. This highlights the significance of exposure to non-negligible doses of secondary neutrons, which could pose a radiological protection issue and impact people's health [8].

By these reasons, the main objective of this work is to develop and validate a methodology to unfold the neutron spectrum produced in a LinAc bunker operating at 15 MV, at different points of interest.

2 Materials and Methods

To perform the described objective and carry out the reconstruction of the neutron spectrum, the work is divided into different steps. Firstly, it is necessary to measure the secondary neutrons produced in the points of interest using a suitable detector. These three points consist of one meter from the LinAc's ISOCENTER, at the maze-room junction, and in the middle of the maze. Secondly, the detector's response to different energies needs to be known. Finally, it is essential to implement a mathematical algorithm to unfold the spectrum using the previously gathered information as input. All these steps are described throughout this section.

2.1 Bonner Sphere Spectrometer System

To be able to measure the neutron contribution, in this study, the multisphere spectrometer used was the Bonner Sphere Spectrometer (BSS) [9]. This one comprises six high-density polyethylene spheres with varying diameters (2, 3, 5, 8, 10, and 12 inches). At the center of these spheres, a thermal neutron detector is placed. This detector is composed of a scintillator crystal made of lithium iodide (^{6}LiI-Eu) with dimensions of 4 mm in diameter and 4 mm in height. To enhance detection efficiency, the scintillator crystal is coupled to a multiplier tube using a Plexiglas light pipe. Since the set of spheres acts as a neutron moderator, depending on the sphere used for measurements, neutrons from thermal energies to hundreds of MeVs can be detected by the scintillator crystal. The complete BSS system is shown in Fig. 1. Measurements in counts per second were obtained for each detector-sphere combination at each point of interest inside the bunker, with the Varian Clinac 2100C operating at 15 MV to ensure secondary neutron production.

2.2 Monte Carlo Simulations

One of the main steps to unfold the neutron spectrum involves generating the response function of the multisphere spectrometer. In this study, the matrix used

Fig. 1: Picture of the BSS detector. Left, the smaller sphere inserted in the detector, and right, the complete set of spheres and the bare detector.

was obtained through the use of Monte Carlo (MC) simulations in a previous work [10], which was validated with the same found in the literature [11].

Additionally, this work involved a set of MC simulations using MCNP6.1.1 [12] with the aim to obtain the BSS measurements also via MC simulations. On the one hand, the detailed geometry of the BSS, including the detector and six spheres, was modeled based on manufacturer information, Fig 2. On the other hand, the Varian Clinac 2100C was accurately modeled based on confidential blueprints provided by the company. The entire setup was placed within a radiotherapy bunker reproducing the specific bunker at Hospital Universitari i Politècnic La Fe de València. ANSYS SpaceClaim and Abaqus/CAE were used for 3D modeling and meshing [13], and both the LinAc and BSS models were validated in previous works [10, 14].

Since the neutron spectrum was unfolded at three different positions within the bunker, for each measurement point, one simulation by detector-sphere configuration was carried out, measuring the neutron counts in the scintillator. For each simulation, a photon beam of 15 MV obtained from [15], is used, reproducing the same conditions as in the experimental measurements.

2.3 Maximum-Likelihood Expectation-Maximization (MLEM) unfolding method

The MLEM method is used in this work to unfold the neutron spectrum at different points inside a radiotherapy bunker. This method is an iterative process that maximizes the likelihood of obtaining the measured data when convergence is achieved, providing an accurate neutron spectrum. It is employed in this work with two sets of data: the neutron measurements obtained from the BSS system experimentally or simulated, and the response function matrix for each detector-sphere configuration.

The response function for each detector-sphere combination i, is unique, denoted by $R_i(E)$ as a function of neutron energy E. The measured neutron data

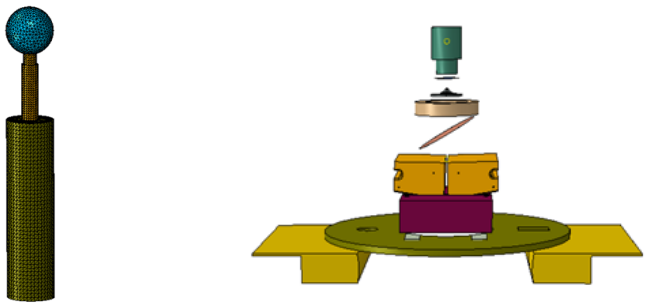


Fig. 2: One of the sphere-detector combinations meshed geometry of the BSS system (left), and complete LinAc geometry (right) used for MC simulations. Both systems are not on the same scale..

for each BSS moderator both of experimental data and MC simulation results, is denoted by m_i. The solution of the problem, i.e, the neutron spectrum obtained through the unfolding process, is denoted by $n(E)$. The relation between the neutron spectrum and the response function is described by a Fredholm integral equation of the first kind (Eq. 1), which can be represented in matrix form $m = R \cdot n$. In this case m refers to the array of dimension i which corresponds to each measured data of each sphere moderator; R is the linear operator corresponding to the rectangular response $i \times j$ matrix where j is the energy neutron beam, and finally, n corresponds to the unfolded neutron spectrum.

$$m_i = \int_E^{E+\Delta E} R_i(E) n(E) dE \tag{1}$$

The iterative form of the MLEM method is described in Eq. (2). In this equation, the summation limits, N and J, correspond to the total number of moderators and the total number of energy bins, respectively.

$$n_j^(k+1) = \frac{n_j^k}{\sum_{i=1}^{N} r_{ij}} \sum_{i=1}^{N} r_{ij} \frac{m_i}{\sum_{b=i}^{J} r_{ib} n_b^k} \tag{2}$$

Then, n_b^k corresponds to the initial spectra, uploaded until the kth iteration; n_j^k is the current spectrum uploaded in each iteration; r_{ij} represents each element of the described response matrix and finally, m_i are the experimental measurements, in counts, for each detector-sphere configuration moderator. Each step of the iteration method provides a new neutron spectrum uploading the previous one. Finally, when the stop criteria stablished by the user is accomplished, the MLEM gives the final unfolded neutron spectrum over a wide energy range.

3 Results

The neutron spectra were unfolded at three different positions within the radiotherapy bunker: 1 meter from the ISOCENTER, at the maze-room junction, and

in the middle of the maze. These results have been obtained both using experimental and simulated data. For the MLEM algorithm a uniform initial spectrum was used as input for the unfolding process. In Fig. 3, Fig. 4, and Fig. 5, the fluence rate spectra are presented in logarithmic scale, indicating the number of neutrons per cm2 per second as a function of energy for each measurement point, respectively. Both point B and C spectra fit precisely for MC and measured data with a difference less of 5% and 1% respectively. At point A, the values of the experimentally measured data cannot be directly precisely obtained. This is due to the problem of this detector, which is the separation of neutron and photon events when photon fluence is extremely high compared with neutron contribution. By this reason, at point A, only the unfolded spectrum using the initial simulated data is shown.

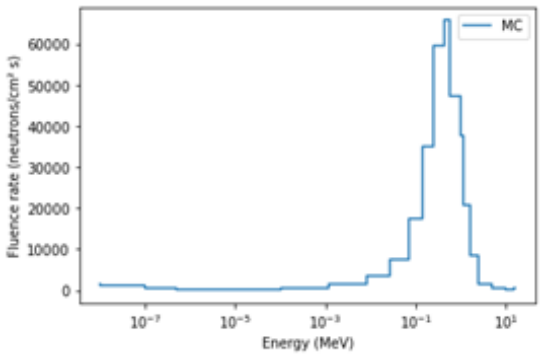


Fig. 3: Unfolded neutron spectrum obtained from Monte Carlo simulated data using LinAc geometry of Varian Clinac at Hospital Universitari i Politècnic La Fe de València, at point A.

In these figures, it can be shown that the maximum fluence is obtained at point A, showing a fast neutron peak. Additionally, at point B the thermal neutron peak surpasses the fast peak which shows 25 less intensity than the same observed at A. Finally at point C, the thermal neutron peak has decreased a factor of 7 with respect to point B.

Once the neutron spectrum has been unfolded, using the fluence-to-dose coefficients of the ICRP Publication 74 [16], and integrating the fluence rate over the energy range, the total ambient dose equivalent in each point of measurement was calculated. Moreover, the ambient dose equivalent was also measured at the same points of the bunker, using an LB6411 detector designed between Berthold and the Karlsruhe Research Center [17]. The results for the neutron ambient dose equivalent are shown in Table 1. Take into account that, for the same reason as in the BSS detector (the photon contribution is extremely high compared with the neutron one), the dose equivalent cannot be measured at point A.

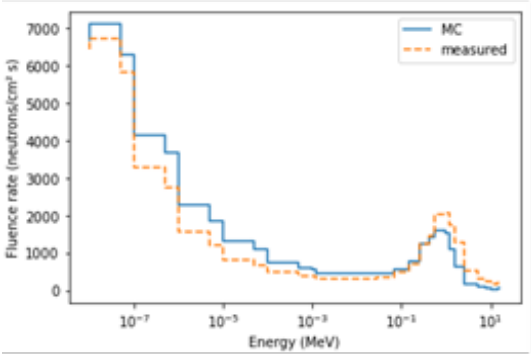


Fig. 4: Comparison between unfolded neutron spectrum obtained from experimental data (dashed lines) and from simulated data (solid lines) at point B, in facility room of Varian Clinac at Hospital Universitari i Politècnic La Fe de València.

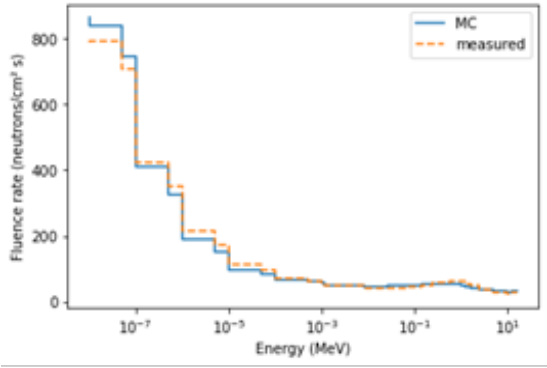


Fig. 5: Comparison between unfolded neutron spectrum obtained from experimental data (dashed lines) and from simulated data (solid lines) at point C, in facility room of Varian Clinac at Hospital Universitari i Politècnic La Fe de València.

Table 1: Total Fluence Rate, Ambient Dose Equivalent H*(10) for Unfolded Spectrum and for Measurements Using Berthold Detector at Different Locations in Facility Room of Varian Clinac at Hospital Universitari I Politècnic La Fe de València at Points A, B and C.

Point	Total fluence (n/cm^2s)	H*(10) (mSv/h) unfolded	H*(10) (mSv/h) Berthold
A	3.15e5	357.0	———
B	4.71e4	13.6	13.1
C	4.67e4	0.97	0.92

4 Conclusions

In this study, secondary neutron spectra were obtained at three points in the radiotherapy treatment room. A realistic simulation of a Varian Clinac emitting a 15 MV beam was performed, incorporating the BSS system geometry to calculate counts per second for each sphere at different LinAc bunker points. Additionally, the same counts per second were measured using the BSS system at the Hospital Universitari i Politècnic La Fe de València. The MLEM algorithm, employing previously validated neutron response curves, was used for spectrum unfolding. This unfolded spectrum at three points of interest of the bunker, shows a decrease in neutron fluence rate at greater distances from the source, with distinctive peak variations in different locations.

The dose values obtained in this work allowed the calculation of neutron ambient dose equivalent, consistent with experimental data measured with the Berthold detector. The described methodology enables dose assessment at various points in the treatment room through MC simulations, the MLEM algorithm, and neutron responses.

Finally, although in this work, the described methodology has been applied to the medical field and to measure the dose associated with the neutrons in a LinAc treatment, this methodology is versatile and applicable in other radiation facilities such as cyclotrons, nuclear power plants, or uranium enrichment factories, for assessing neutron dose contribution in any region of interest.

Acknowledgments

The authors would like to acknowledge the collaboration of the radiotherapy staff of the Hospital Universitari i Politècnic La Fe de València for their constant interest, support, availability, and collaboration in the development of this work. Thanks to Varian for all the information received related to the LinAc model.

References

1. National Council on Radiation Protection and Measurements (NCRP), Neutron Contamination from Medical Electron Accelerators, “NCRP Report 79” (1984).
2. W.P. SWANSON, “Radiological safety aspects of the operation of electron linear accelerators”, Technical Report No 188, IAEA (1979).
3. S. F. Kry et al., “AAPM TG 158: Measurement and Calculation of Doses Outside the Treated Volume from External-Beam Radiation Therapy,” Med. Phys., 44, 10 (2017).
4. W. D. Newhauser and M. Durante, “Assessing the Risk of Second Malignancies After Modern Radiotherapy,” Nat. Rev. Cancer, 11, 438 (2011); http://dx.doi.org/10.1038/nrc3069.
5. Y. Z. Wang, M. D. C. Evans, and E. B. Podgorsak, “Characteristics of Induced Activity from Medical Linear Accelerators,” Med. Phys., 32, 9, 2899 (2005); http://dx.doi.org/10.1118/1.2001767.

6. J. Alan Rawlinson, M. K. Islam, andD. M. Galbraith, "Dose to Radiation Therapists from Activation at High-Energy Accelerators Used for Conventional and Intensity-Modulated Radiation Therapy," Med. Phys., 29, 4, 598 (2002).
7. G. Baiocco et al., "The Origin of Neutron Biological Effectiveness as a Function of Energy," Sci. Rep., 6, 34033 (2016); http://dx.doi.org/10.1038/srep34033.
8. A. Ma et al., "Monte Carlo Study of Photoneutron Production in the Varian Clinac 2100C Linac," J. Radioanal. Nucl. Chem., 276, 1, 119 (2008); http://dx.doi.org/10.1007/s10967-007-0419-3.
9. LUDLUM MODEL 42-5 NEUTRON BALL CART. Product manual. (2015)
10. S. Morató, B. Juste, R. Miró, G. Verdú, V. Guardia. "Evaluation of the response of a Bonner Sphere Spectrometer with a 6LiI detector using 3D meshed MCNP6.1.1 models", Radiat. Phys. Chem, 155, 221-224. (2019).
11. M. Lacerda et al., "A New Response Matrix for A 6LiI Scintillator BSS System," Nucl. Instrum. Methods Phys.Res. Sect. A, 868, 122 (2017); http://dx.doi.org/10.1016/j.nima.2017.06.057.
12. LOS ALAMOS SCIENTIFIC LABORATORY GROUP X-6., "A general Monte Carlo code for neutron and photon transport", Los Alamos, n.m.,Dept. of Energy, Los Alamos Scientific Laboratory (1979)
13. Dassault Systemes 3D Experience Company ABAQUS 6.14, ABAQUS/CAE User's Guide. (2014)
14. S. Morató et al., "Varian Clinac 6 MeV Photon Spectra Unfolding Using a Monte Carlo Meshed Model," EPJ Web Conf., 153, 04012 (2017)
15. D. Rogers, W. Sheikh-Bagheri. "Monte Carlo calculation of nine megavoltage photon beam spectra using the beam code", Med. Phys. 29, 391-402. (2002)
16. "Converssion Coefficients for Use in Radiological Protection Against External Radiation," ICRP Publication 74, International
17. Commission on Radiological Protection (1996) Berthold Technologies website; https://www.berthold.com/en/search/ (current as of Mar. 22, 2022).

Analysis of the effectiveness of a freight transport vehicle at high speed in a vacuum tube (Hyperloop transport system)

Pellicer, D. S. and Larrodé, E.

University of Zaragoza, María de Luna Street no. 3, 50018, Zaragoza, Spain
dasapezu@unizar.es

Abstract. This paper shows the development of a numerical analysis model, which enables the calculation of cargo transport capacity of a vehicle that circulates through vacuum tubes at high speed and analyzes its effectiveness in transport. The simulated transportation system is based on the use of vehicles that move in vacuum tubes at high speed, which is commonly known as Hyperloop, but assuming the vehicle for cargo containers. For the specific vehicle proposed, which does not include a compressor and levitates on magnets, the system formed by the vehicle and the vacuum tube is conceptually developed, establishing the corresponding mathematical relationships that define its behavior. To properly model the performance of this transport system, it has been necessary to establish the relationships between the design variables and the associated constraints, such as the Kantrowitz limit, aerodynamics, transport, energy consumption, etc. Once the model was built and validated, it was used to analyze how it affects the variation of the transported load (in our case number of containers), the speed of operation and the length of the tube, with the total and specific consumption of energy. Once the most efficient configuration was found in regard to energy consumption and transport effectiveness, the complete system was calculated. The results obtained constitute a first approximation for the pre-design of this transport system and the built model allows different alternatives to be compared according to the design variables.

Keywords: High-speed Transport, Freight Transport, Sustainable Transport, Electric Vehicles.

1 Introduction

The objective of this work is to develop a calculation model that could find the best configuration of a vehicle that transports heavy goods at high speed in a vacuum tube, and thus obtain greater energy efficiency as well as greater effectiveness in the operation of transport. The process consists of defining a case study under behavioral hypotheses, parameterizing the problem through the behavioral equations corresponding to each of the physical phenomena that occur, and applying the analysis to a predesign of a vehicle that simulates the operation in real conditions. Once the behavior relations of the system are established,

the energy consumption, the performance of the system, as well as the verification of the Kantrowitz limit, are determined. This allows selecting the optimum amount of load to be transported, the most suitable operating speed, and the most appropriate tube length. Once the optimal values of these variables have been obtained, the rest of the main characteristics of the vehicle are determined.

Regarding the type of system used for vehicle-infrastructure interaction, the authors have considered two options for levitation: air bearings and electrodynamic suspension (EDS). This work focuses on the latter. The vehicle does not include a compressor to overcome the Kantrowitz limit at near-sonic speeds or airfoils; but includes batteries in the rear of the vehicle for the control, and in the EDS rotor. The vehicle also has a mechanical brake for immediate braking in the event of an emergency. It is only considered cargo contained in 20-foot aluminum Dry Van containers and each container must be placed within a single capsule. A linear geometry has been studied, with a straight tube with zero slopes, and which can be 500, 750 or 1,000 km long between origin and destination.

Due to work limitations, other issues such as technical and economic feasibility, control loops, stability, infrastructure, vehicle structure, heat transfer, EDS geometry and electrical systems are out of scope and may be eligible for additional work.

To carry out this research work, it has been necessary to review the behavioral theories of the different physical phenomena involved and the extraction of the corresponding behavioral laws: aerodynamics, electromechanics and thermodynamics.

The main contribution of this work is the determination of the most suitable masses and volumes for freight transport using containers in a vehicle that travels at high speed in a vacuum tube levitating on magnets. Thus, more efficient transportation can be achieved with lower energy consumption per ton and kilometer transported, and greater effectiveness in transportation by establishing the ideal number of containers in cargo movement.

Another contribution to highlight is the analysis procedure, taking into account all of the physical conditions of the problem, and adding the restrictions and limitations of the case to be studied. The result is the variation of the parameters sought. In this case, for example, the optimal weight and volume, which allows finding the most appropriate alternative to the proposed criterion, is aimed at the minimum energy consumption. Once the analysis procedure has been validated, the methodology is open to adding more restrictions and limitations for future research work.

2 Material and methods

The methodology applied in this work follows a deductive method, in which, through the construction of the physical problem to be solved, the behavioral equations of the laws of thermodynamics, electromechanics and aerodynamics are applied to the specific case proposed. By establishing the determined limits, the comparison variables that allow an analysis based on the variation of param-

eters are obtained, and the variation of parameters enables the acquisition of an optimal design. First, the problem to be solved is defined, consisting of establishing the behavior laws of a vehicle levitating on magnets in a vacuum tube to be transported at high speed, for which a series of hypotheses have been proposed. These hypotheses are fundamental to delimit the model of the high-speed transport system, which is defined by the physical equations of its main subsystems: aerodynamics, electromechanics and thermodynamics. Second, these equations are interrelated by auxiliary equations that are introduced later, building a system of equations that is solved by mathematical equation solving software. This software allows to solve the system of equations after configuring the input data. The parameters can be varied in the case study: the calculation is carried out by varying one parameter at a time.

2.1 Hypothesis

The following hypotheses have been regarded:

1. Subsonic speed.
2. Ideal gas theory, since the compressibility factor is around 1 under the system working conditions.
3. Isentropic compression as the vehicle moves and the air is compelled to flow into the annulus.
4. The boundary layer does not separate from the vehicle.
5. Both acceleration and deceleration are held constant.
6. The frontal area of the EDS magnets is negligible with respect to the annulus area (figure 2 (a)).
7. Any lateral forces generated by the propulsion part of the EDS are not considered. These are inherently stabilizing and low with respect to the propulsion force (Pellicer, 2019).
8. Active power losses in the EDS are modeled with a single stator resistance (figure 2 (c)).
9. The average power dissipated by the EDS drag is considered one third of the maximum during acceleration and braking. This is because the power dissipated first increases and then decreases with speed (Flankl, 2018). If it were linear with speed, then the average power would be half of the maximum, but in this case, it is less due to that decrease.
10. The diameter needed to accommodate the load is equal to the diameter of the circumference surrounding a container (figure 2 (b)).

3 Calculation

The calculation process consists of three parallel branches that conflate at a point:

a) In the left branch, the power dissipated by aerodynamic drag is computed. For that, the speed of the vehicle and its thermodynamical data are entered. At

that given speed, the tube diameter is calculated so that the Kantrowitz limit is prevented. According to the blockage ratio, the power dissipated by aerodynamic drag is computed.

b) In the middle branch, the onboard batteries which feed the rotor of the linear motor are dimensioned. Their dimensioning comes from evaluating their energy density and their discharge time, which depends on the total travel time. In turn, the total travel time relies on the operating speed, acceleration and deceleration of the vehicle through kinematics relations.

c) In the right branch, the power needed to propel and lift the vehicle is calculated. This calculus highly relies on the number of containers (that equals the number of capsules in the vehicle) and their individual masses, which depend on the filling factor of each container. These data determine how much mass is lifted and propelled and, thus, the power needed for that. These branches conflate in order to determine the energy consumption of the vehicle. This way, the energy consumption is linked to the mass transported and to the operating speed, which allows finding relations between the mass flow and the energy needed to maintain that mass flow (always considering that only one vehicle, the one that is to be optimized, is using the tube). The calculation process is shown in the next flow diagram, which shows how the different equation blocks are interrelated. Equation blocks referring to the main subsystems (aerodynamics, electromechanics and thermodynamics) are represented with a bolded contour, while auxiliary equation blocks are represented with a normal contour. The final block is represented with a doubly-bolded contour:

In this way, it is ensured that the behavior laws of the vehicle inside the tube are fulfilled under all the requirements and considering all the starting hypotheses, with which the physical phenomenon is completely characterized. Once the problem has been formulated, and the behavior equations and the input data are introduced into the software, the software finds the solution to the system of equations. Finally, the design parameters, such as the transported mass, the operating speed, and the length of the tube, are varied according to the simulation procedures. The results (the new results of the system of equations) are obtained in relation to the energy consumption of the transport operation.

3.1 List of abbreviations

The list that contains the abbreviations used in the following sections can be consulted in Annex 1 (Table 3).

3.2 System definition

System drawings

Aerodynamics The high-speed transport system runs inside a tube, and this is like a vehicle that runs inside a tunnel, whose drag coefficient increases as a result of the tunnel effect. According to Melis (2001), the relation between the

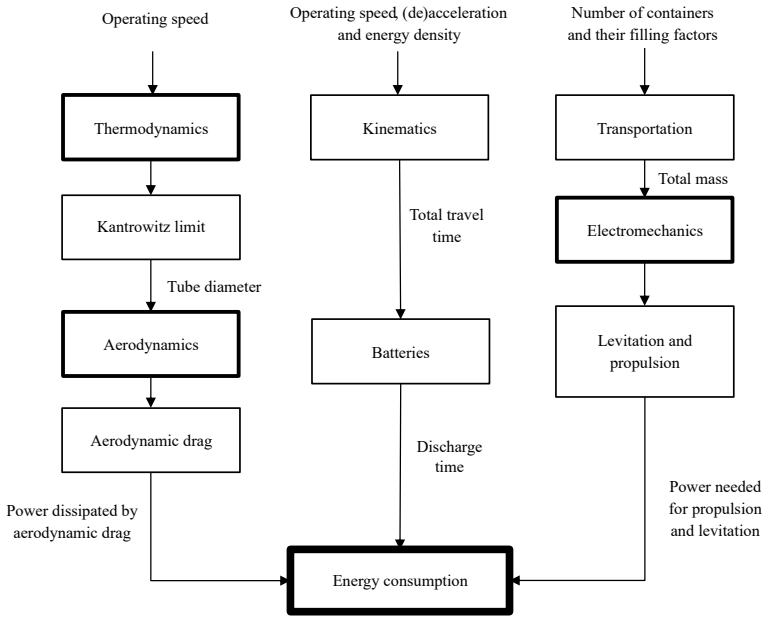


Fig. 1: Flow diagram of the calculation process. Source: Own elaboration.

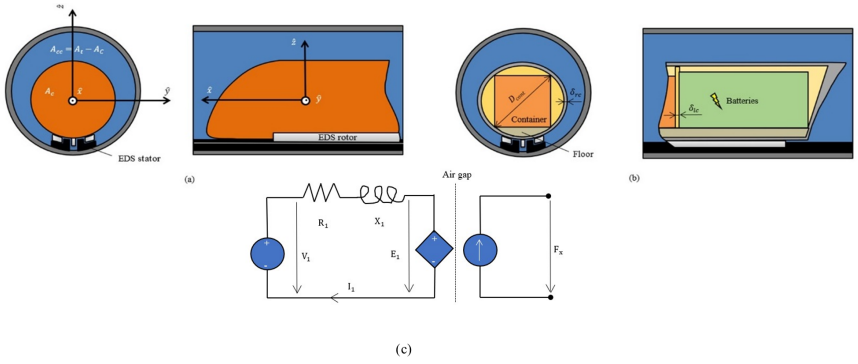


Fig. 2: (a) Cross-sectional drawing of the tube in front of the vehicle and its profile. (b) Cross-sectional drawing of the tube and the vehicle, near the rear, and its axial section. (c) Electrical model for EDS. Source: Own elaboration..

drag coefficient inside and outside is expressed as follows (Eq. 1). To calculate the coefficient of drag inside, the same reference includes this formula (Eq. 2). Furthermore, the outside drag coefficient is related to the moment section of the boundary layer (Eq. 3):

$$T_f = \frac{C_{D_t}}{C_{D_{ext}}} \tag{1}$$

$$C_{D_t} = \left(C_{D_{ext}} + \beta \left(\frac{\Delta_1}{A_f} \right)^2 \right) \left(\frac{1 - \frac{c_i}{v}}{1 - \beta \left(\frac{\Delta_1}{A_f} \right)} \right)^2 \tag{2}$$

$$C_{D_{ext}} = 2\frac{\Delta_2}{A_f} \tag{3}$$

A relationship exists between the boundary layer momentum section and the boundary layer displacement section. To find this relationship, it must be taken into account that the boundary layer will be laminar, as can be verified by calculating both the local and the global Reynolds number with some data extracted from Pellicer (2019) (first model):

$$Re_{Dc} = \frac{\rho_t v D_c}{\mu_t} \tag{4}$$

$$Re_{Lc} = \frac{\rho_t v L_c}{\mu_t} \tag{5}$$

$$Re_{D_c} = \frac{1.18 \cdot 10^{-3} \cdot \frac{1{,}220}{3.60} \cdot 1.34}{1.80 \cdot 10^{-5}} = 29,769.51 \tag{6}$$

$$Re_{L_c} = \frac{1.18 \cdot 10^{-3} \cdot \frac{1{,}220}{3.60} \cdot 25}{1.80 \cdot 10^{-5}} = 555,401.23 \tag{7}$$

$$\tag{8}$$

Where $1.80 \cdot 10^{-5}$ Pa $\cdot$ s is the dynamic viscosity for dry air at 20 ºC and 100 Pa (the variation of viscosity with pressure is neglectable for such a low pressure) (Schlichting, 2017). 25 m is approximately the length of the capsule, which can be gathered from Pellicer (2019): The passenger capsule levitates on 28 air bearings, 14 on each side and 1.5 m long each (21 m in total, to which other parts as the nose and nozzle are added). With respect to $1.18 \cdot 10^{-3}$ kg/m^3, this is the air density and comes from the ideal gas equation. It can be noted that the local Reynolds is small and not significant, whilst the global can be proper to a laminar boundary layer, since the transition from laminar to turbulent occurs somewhere between 5×10^5 and 1×10^6 for a flat plate. Assuming that it is always laminar for the high-speed transportation system, von Karman results can be used to relate the momentum thickness to the displacement thickness through the layer thickness. The process is shown below, after collecting the proper information from Schlichting (2017) (Eqs. 6 – 7). The function $u(y^{''})$ could be linear, parabolic, polynomial, etc. As a first approximation, the speed profile is assumed to be linear (Eq. 8). After integrating, the following is obtained

(Eqs. 9 – 10):

$$\delta^* = \int_0^\delta \left(1 - \frac{u(y'')}{U}\right) dy'' \tag{9}$$

$$\theta = \int_0^\delta \frac{u\,(y'')}{U} \left(1 - \frac{u\,(y'')}{U}\right) dy'' \tag{10}$$

$$u\,(y'') = \frac{U}{\delta} y'' \tag{11}$$

$$\delta^\star = \frac{\delta}{2} \tag{12}$$

$$\theta = \frac{\delta}{6} \tag{13}$$

After having gathered all this information, the aerodynamic analysis is solved. Only geometric relations and the generic formula for calculating the drag force remain, which involves density, speed, and frontal area in addition to the drag coefficient inside the tube. All the formulas are shown in the next sections.

Electromechanics For the study of the EDS, the works consulted are Frankl (2018), Abdelrahman (2018) and Lever (1998). The EDS used for this high-speed transportation system is very similar to the one used for maglev vehicles, although in maglev vehicles wheels are needed at low speeds because there is not enough induction magnetic field to levitate. The traditional EDS can be modeled as a LIM (linear induction motor) for levitation and as an LSM (linear synchronous motor) for propulsion. In order to eliminate the need for wheels, the LIM is replaced by an LSM when applying EDS to the high-speed transportation system, where the rotor will be mounted on the pod (short rotor) and the stator on the tube (Abdelrahman, 2018). These two expressions are taken from this work (Eqs. 11 – 12). The article by Flankl (2018) contains explanations and formulae for levitation and the drag force generated by the EDS operation. Below the formulae can be found, although expressed a little differently (Eqs. 13 – 15). Furthermore, the next equations from Lever (1998) have been used in the analysis (Eqs. 16 – 18), where the number three indicates the number of phases

of the motor:

$$\eta_{EDS} = \frac{F_x v}{F_x v + 3I_1^2 R} \tag{14}$$

$$cos\varphi = \frac{F_x v + 3I_1^2 R}{3V_1 I_1} \tag{15}$$

$$F_z = m_{tot} g \tag{16}$$

$$F_{D_{EDS}} = C_{D_{EDS}} F_z \tag{17}$$

$$P_{D_{EDS}} = F_{D_{EDS}} v \tag{18}$$

$$P = 3V_1 I_1 \cos\varphi \tag{19}$$

$$3E_1 I_1 cos\gamma_0 = F_x v \tag{20}$$

$$Q = 3V_1 I_1 \sin\varphi \tag{21}$$

$$\tag{22}$$

The electrical model for EDS is shown in Lever (1998). This model is based on the LSM, which can be seen as a rotary synchronous motor rolled out flat. Subsequently, a resistance and a reactance are used at the stator (on the left). At the model air gap, electric power is equated to mechanical. On the right, a damper and a spring are joined to represent mechanical losses. However, for a first parameter estimation, it is preferable to remove the damper and the spring, and to consider that all active power losses occur in the stator resistance (figure 2 (c)). Other electrotechnical equations have been used to analyze the EDS. Furthermore, to calculate the thrust required and the power input at the end of acceleration (maximum losses), Newton's second law has been applied.

Thermodynamics Lastly, to derive the Kantrowitz limit main expression, three basic thermodynamics equations were utilized: Mass flow conservation, Mach number definition, sound speed in an ideal gas, ideal gas law, and isentropic relations for pressure and temperature. The subscript 1 represents the air state variables or associated ones before the air flows into the annulus and the subscript 2 represents the contrary. According to Mattingly (1996), most of the aforementioned formulae may be consulted. The main expression to analyze the Kantrowitz limit phenomenon is derived by combining Eqs, 19 – 25 (subscripts for $i = 1, 2$). The complete process can be found in Bar-Meir's work (2013) and its outcome is Eq. 26:

$$\dot{m}_i = \rho_i A_i v_i \tag{23}$$

$$\dot{m}_i = \text{constant} \tag{24}$$

$$v_i = a_{s_i} M_i \tag{25}$$

$$a_{s_i} = \sqrt{\gamma R T_i} \tag{26}$$

$$\rho_i = \frac{p_i}{R T_i} \tag{27}$$

$$\frac{p_{0_t}}{p_i} = \left(1 + \left(\frac{\gamma - 1}{2}\right) M_i\right)^{\frac{\gamma}{\gamma - 1}} \tag{28}$$

$$\frac{T_{0_t}}{T_i} = \left(1 + \left(\frac{\gamma - 1}{2}\right) M_i\right)^{\frac{\gamma}{\gamma - 1}} \tag{29}$$

$$\dot{m}_{cc_{\max}} = A_{cc} \frac{p_{0_t}}{\sqrt{T_{0_t}}} \sqrt{\frac{\gamma}{R}} \left(1 + \left(\frac{\gamma - 1}{2}\right)\right)^{\left(\frac{\gamma+1}{2(\gamma-1)}\right)} \tag{30}$$

Note: $\rho_1 = \rho_t$, $A_1 = A_t$, $A_2 = A_{cc}$, $v_1 = v$. See also figure 2.

3.3 Final equation block

The final block of the model is the energy consumption block. This block comes from Pellicer (2019) and relies on the results of the rest of the blocks (figure 1). So the aerodynamics (Eqs. 1 – 10), electromechanics (Eqs. 11 – 18) and thermodynamics (Eqs. 19 – 26) blocks are interrelated through the auxiliary equation blocks (Eqs. 27 – 57, presented in Table 4 in Annex 1). The final block equations are:

3.4 Software choice

Once all the equations have been obtained, it is necessary to process them in an equation solver program. Due to the large number of equations and relations that had to be implemented, only software capable of processing the entire volume of data in an agile way has been considered. After considering several options (Mathematica, Matlab and Engineering Equation Solver), Engineering Equation Solver (Klein, 1993) has been chosen as it is used in other models that involve thermodynamical equations (Tirmizi, 2012). The specific version the results were obtained with is Engineering Equation Solver Professional V9.457-3D (EES). The chosen program, besides solving equations, can create parametric tables and graphs derived from those equations.

3.5 Simulation procedures

The objective is to analyze the capacity of this transport system and compare different alternatives based on their efficiency. However, there is a lot of input

Table 1: Final equation block.

Block	Equation	Left –side variable [SI unit]	Variable definition	Equation number
Energy consumption	$E_{ac} = \left(\frac{m_{tot} a_1 \frac{v}{2} + \bar{P}_{av}}{\eta_{EDS}} \right) t_{ac}$	E_{ac} [J]	Energy consumed during acceleration	58
	$E_{gen} = -\eta_{EDS} \left(m_{tot} a_2 \frac{v}{2} - \bar{P}_{av} \right) t_{dec}$	E_{gen} [J]	Energy generated during deceleration	59
	$\bar{P}_{av} = \frac{P_D}{4} + \frac{P_{D_{EDS}}}{3}$	$\bar{P}_{av}$ [W]	Mean power dissipated by running resistance	60
	$E_v = \frac{P_{av}}{\eta_{EDS}} t_v$	E_v [J]	Energy consumed throughout the travel at the speed v	61
	$E_{bat} = \frac{m_{Li^+} e_{bat} \frac{t_{tot}}{t_{des}}}{\eta_{bat}}$	E_{bat} [J]	Energy consumed by the batteries	62
	$E'_t = \frac{E_{ac} + E_{gen} + E_v + E_{bat}}{L_t}$	E'_t [J · m^{-1}]	Total energy consumed per unit length	63
	$e'_t = \frac{E'_t}{m_{carga} \sum_{i=1}^{i=n_{cont}} f_i}$	e'_t [J · m^{-1} · kg^{-1}]	Total energy per unit length and payload mass	64
	$I_e = \frac{E_{ac} + E_{gen} + E_v + E_{bat}}{m_{carga} \sum_{i=1}^{i=n_{cont}} f_i}$	I_e [J · kg^{-1}]	Energy consumption per payload mass (energy index)	65

data to enter before getting the results, that is, the final values of all the output variables involved.

First, input data are chosen. They may come from different sources: references, calculations, and optimizations with the aid of EES tables and graphs in most cases. Then, they are entered in the program.

Once those data have been selected and entered, the number of containers, speed and tube length can be chosen. The choice of these essential factors that are based on the less important factors that we have just selected and introduced is what this work focuses on because they lead to the results. All these results will be obtained for a single vehicle using one of the two tubes, which will be optimized. This vehicle enters the tube, travels through it, and leaves it at the exact instant that a new vehicle begins its journey. After optimization, the results will be extrapolated to a regular transport flow, including the dispatch frequency of the vehicles.

Starting with the number of containers, the most interesting plot to choose is the $I_E - I_C^{-1}$ plot (several curves, one for each number). When selecting it, two factors are key:

1. I_E or in other words, specific energy consumption to payload, must be the lowest possible.
2. I_C or cargo throughput per unit time must be the highest possible. However, its inverse is used on the plot so that optimal points will fall around the lower-left corner. Seen from another perspective, it can be stated that it is important to minimize the time required to send the payload.

In order to obtain one curve instead of one point with coordinates (I_C^{-1}, I_E) for every number of containers, these two basic variables could be altered: a) Speed, which is a relevant factor, as both I_E and I_C strongly depend on it, so a

range of speed values is included as input to make the plot. Were the range not included, then the outcome would be one point with coordinates (I_C^{-1}, I_E) for every number of containers. The range for a high-speed transportation system without a compressor is 700-1,000 km/h, as will be demonstrated later. b) Tube length. As defined in the beginning, it can take one out of three discrete values: 500, 750 or 1,000 km. I_E and I_C also depend on this to a great extent. Speed is chosen because the $I_E - I_C^{-1}$ curves as a function of speed will be helpful when selecting it afterward. Choosing the tube length would not have been useful later because the consumption per unit length would not have been represented. This leads to the choice of speed. $I_E - I_C^{-1}$ curves are used for this, but Kantrowitz limit results are crucial inasmuch as aerodynamics play a huge role. The speed chosen must comply with the following requirements: Working conditions under the Kantrowitz limit while keeping the lowest possible D_t, low I_E and high I_C (or low I_C^{-1}, its counterpart). Plus, it should leave maglev speeds behind by a sufficient margin. The most suitable graph for presenting Kantrowitz limit results is the $D_t - v$ curve. This way, the speed selected will be the one that optimizes I_E, I_c and D_t.

After this, the tube length is selected out of the three figures available. This time, I_E is no longer useful on its own. This is because I_E is energy divided by mass, being E_v the factor escalating linearly with L_t (through t_v according to equations 26 – 33 and 53). Were I_E utilized, then 500 km would be optimal for minimizing both I_E and I_c^{-1}, but energy per unit distance would not even have been considered. Energy per unit distance is relevant because it contributes to determine operation costs. With that said, the unknown e_t' is chosen instead of I_E, resulting in $e_t' - I_c^{-1}$ curves. e_t' may be seen as the combination of I_E and E_t' and the optimal length will be the one that minimizes both of them, this being interpreted as pursuing low transportation costs and low operation costs.

Finally, the optimal values for the number of containers, speed and length are introduced. Once the program has compiled everything, the window with the final values will appear on the screen, arranged in alphabetical order.

3.6 Input data

Firstly, 20' aluminum Dry Van containers have the following characteristics: 6.058 m ($\simeq$20') for length (L_{cont}), 2.438 m for width, 2.591 m for height, 2,180 kg for tare (m_{tare}), 28,300 kg for maximum load (m_{carga}). According to the width and height of the container, the parameter D_{cont} is 3.558 m, using Pythagoras' theorem. After setting the dimensions of the specified container, the rest of the input variables are given values:

1. $a_1 = a_2 = 14.72 m/s^2$ (1.5 g). This is because cargo withstands higher accelerations than passengers as there are not any discomfort issues.
2. c_i and g are constants and the former is null (there is not any wind flowing inside the tube).
3. e_{bat}, R, γ and η_{bat} were extracted from references.
4. The rest are optimal (Pellicer, 2019).

Table 2: Table 2. Input variables with their respective units and their values to their right.

Variable	Value
a_1 $[^m/_{s^2}]$	14.72
a_2 $[^m/_{s^2}]$	14.72
C_{DEDS} $[\phi]$	3·10-3
C_{Dext} $[\phi]$	0.60
c_i $[^m/_s]$	0 (const.)
D_{cont} $[m]$	3.558
e_{bat} $[Wh/kg]$	225
g $[ms2]$	9.81 (const.)
L_{cont} $[m]$	6.058
m_{carga} $[kg]$	28,300
m_{EB} $[kg]$	800*
m'_{EDS} $[^{kg}/_m]$	32
m'_{est} $[kg/m]$	500
m_{Li^+} $[kg]$	400*
m_{tara} $[kg]$	2,180
p_t $[Pa]$	250
R $[^J/_{kg \cdot K}]$	287
R_1 $[\Omega]$	8
T_t $[^oC]$	20
γ $[\phi]$	1.40
γ_o $[^o]$	15
δ_{lc} $[m]$	0.04
δ_{rc} $[m]$	0.05
η_{bat} $[p.\ u.]$	0.90
η_{EDS} $[p.\ u.]$	0.73
τ [%]	30
φ $[^o]$	30

Note: $m_{Li^+} = 350$ kg for $n_{cont} = 1$ and 50 kg is added per each additional container. $m_{EB} = 750$ kg for $n_{cont} = 1$ and 250 kg is added per each additional container. 350 and 750 kg have been used to start the series.

4 Results

4.1 $I_E - I_C^{-1}$ curves

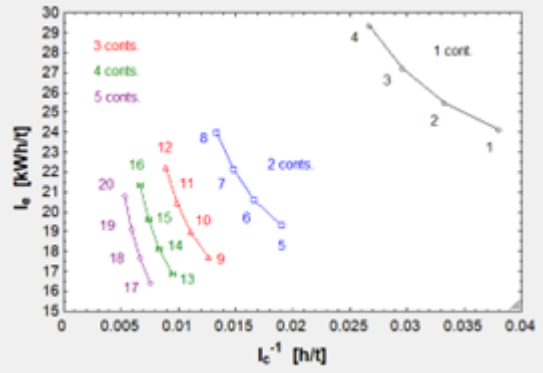


Fig. 3: $I_E - I_C^{-1}$ curves for $L_t = 750$ km and for 1 – 5 containers (abbreviated as cont. or conts.). Results from Table 5 (annexed).

In conclusion, when increasing n_{cont} there is an improvement in both I_E and I_C, which is clearly smaller after every increment. When adding one container for the first time, payload (associated with capacity) grows by roughly 30 t. This is a 100 % growth, from 30 to 60 t. When adding one container again, payload grows by roughly 30 t with respect to the initial 60. This is a 50 % increase. The next time there is a 33 % increase (30/90) and finally 25 % (30/120). This results in a slowing-pace increase in I_C (the contrary for I_C^{-1}).

Besides this, the dead weight also grows increment by increment: m_{Li^+} and m_{EB} grow as established in table 5, m'_{est} and m'_{EDS} multiply a longer length ($n_{cont} L_c$ according to equation 46) and m_{tara} n_{cont} according to the same formula. This and the slowing-pace improvement in capacity explains the slowing-pace decrement in I_E, which is mainly governed by the ratio $m_{tot}/(m_{carga} \sum_{i=1}^{i=n_{cont}} f_i)$ (the difference between the numerator and denominator is the deadweight) and by losses independent from m_{tot} (chiefly $P_D t_v$ and E_{bat}) divided by payload. In the end, n_{cont} is set to 5 because the improvement from 5 to 6 will be predictably tinier and over dimensioning the system is undesirable.

4.2 $D_t - v$ curve

Analyzing figure 4, it can be deduced that the zone of interest goes from 700 to 800 km/h (D_t around 9 m), these beings the reasons: 9 m is suitable considering

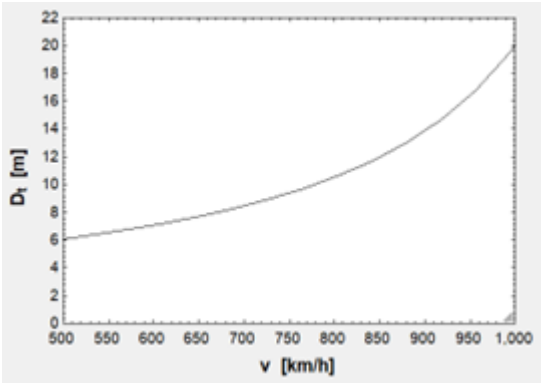


Fig. 4: $D_t - v$ curve. Results from Table 6 (annexed).

that D_c is 3.658 m, so that blockage will be small (0.16 or 16 % at 728 km/h according to table 6); speeds below 700 are near maglev speeds and speeds above 800 yield a D_t rising at a higher rate.

The relevant information provided by figure 3 concerning v is that the ends of any speed range should be avoided: Lowest speeds yield a low I_E, but low I_C (or high I_C^{-1}). By contrast, the highest speeds imply the contrary. This means that the optimal speed will be near the center of the speed interval. With this being said, v is chosen as 750 km/h.

4.3 $e'_t - I_C^{-1}$ curves

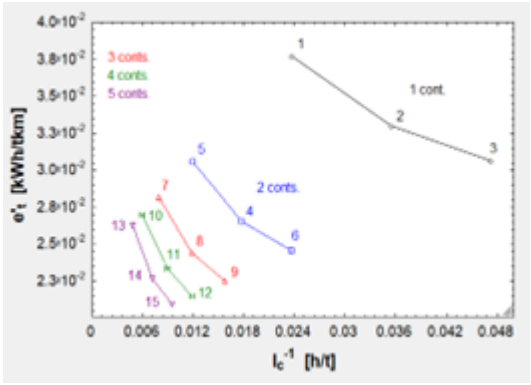


Fig. 5: $e'_t - I_C^{(} - 1)$ curves at $v = 750$ km/h and for 1 – 5 conts. Results from Table 7 (annexed).

In contrast to the $I_E - I_C^{-1}$ curves, now e'_t has replaced I_E. It must be noted that e'_t can be calculated as E'_t divided by m_{tot} or I_E divided by L_t. This means that all of the tendencies observed before are still valid. Now there are two additional tendencies, explained next.

In the first place, E'_t decays as L_t augments, as table 7 proves. This is due to the fact that accelerating the vehicle requires the provision of a high amount of kinetic energy and this energy is better used for longer routes. Secondly, I_C worsens as L_t grows. It is simple to understand this by reviewing equation 49: As L_t grows, t_{tot} does too and I_C decreases (or I_C^{-1} increases). Shorter routes allow a higher throughput because for the same period of time more containers can be dispatched.

After having seen the different trends involved, it can be concluded that the best option is $L_t = 750$ km. 750 km (point/run 14 above in figure 5 and table 7) is the only one that optimizes e'_t (associated with both I_E and E'_t) and I_C. 500 km (point/run 13) improves I_C and its counterpart but worsens e'_t, while 1,000 km (point/run 15) has the contrary effect.

5 Conclusions

Through the mathematical modeling of a novel high-speed transport system based on the use of vacuum tubes, the most convenient design has been obtained which allows an effective freight transport operation, which is also efficient in terms of energy. This effective freight transport operation complies with all of the technical requirements and with all the limitations of the physical problem. The model allows taking into account all the equations involved by the electromechanical, aerodynamic, and thermodynamic laws present in the definition of the problem. By introducing boundary conditions and starting hypotheses, the model allows an analysis of parametric variation to be carried out. In the case presented, the optimal number of containers that can be transported at high speed with the lowest possible energy consumption can be obtained as a result, in a technically feasible model. As a continuation of the research work, the next steps to be carried out will consist of the consideration of solving the problem with the restrictions and difficulties that come with using a tube with different curvatures as infrastructure, and with the existence of slopes along the route.

References

1. Abdelrahman, A.S., Sayeed, J.M., & Youssef, M.Z. (2018, 23-27 September). Hyperloop Transportation System: Control, and Drive System Design [Paper presentation]. IEEE Energy Conversion Congress and Exposition, Portland, OR, USA.
2. Bar-Meir, G. (2013). Fundamentals of Compressible Fluid Mechanics, version 0.4.9.8. Resource document. http://www.bigbook.or.kr/bbs/data/file/bo11/1535291005_vkqPUAtN_Fundamentals_of_Compressible_Fluid_Mechanics_Genick_Bar-Meir.pdf Accessed 1 January 2022.

3. Biswas, M.M., Azim M.S., Saha, T.K., et al. (2013). Towards Implementation of Smart Grid: An Updated Review on Electrical Energy Storage Systems. Smart Grid and Renewable Energy, 4(1), 122-132.
4. Çengel, YA, & Boles, MA. (2014). Thermodynamics: An engineering approach, 8th ed. New York: McGraw-Hill.
5. Flankl, M, Wellerdieck, T, Tüysüz, et al. (2018). Scaling laws for electrodynamic suspension in high-speed transportation. IET Electric Power Applications, 12(1), 357-364.
6. Klein, S.A. (1993). Development and integration of an equation-solving program for engineering thermodynamics courses. Computer Applications in Engineering Education, 1(3), 265-275.
7. Lever, J.H. (1998). Technical Assessment of Maglev System Concepts. Resource document. https://rosap.ntl.bts.gov/view/dot/42287/dot_42287_DS1.pdfF Accessed 1 January 2022.
8. Mattingly, J.D. (1996). Elements of Gas Turbine Propulsion, 1st ed. New York: McGraw-Hill.
9. Melis, M.J., de Matías, I., Alonso, J.M., et al. (2021). Diseño de túneles para trenes de alta velocidad. Rozamiento tren-aire-túnel y ondas de presión. Revista de obras públicas, 148(3415), 27-44.
10. Pellicer, D. S., & Larrodé, E (2019). Conceptual development, analysis and simulation of the transport capacity of a freight transport vehicle in vacuum tubes at high speed (Hyperloop concept). (Bachelor thesis, University of Zaragoza, Zaragoza, Spain). Retrieved from http://deposita.unizar.es/record/48766 Accessed 1 January 2022.
11. Schlichting, H, et al. (2017) Boundary-Layer Theory, 9th ed. New York: Springer.
12. Tirmizi, S.A., Gandhidasan, P., & Zubair, S.M. Performance analysis of a chilled water system with various pumping schemes. Applied Energy, 100(1), 238-248.

Annex 1

Table 3. List of abbreviations.

Abbreviation	Definition	Unit (SI)
A_c	Pod cross-sectional area	m^2
A_{cc}	Annulus area	m^2
A_f	Frontal area projected on a plane normal to the tube	m^2
A_t	Tube cross-sectional area	m^2
a_1	Acceleration	$m \cdot s^{-2}$
a_2	Deceleration	$m \cdot s^{-2}$
a_s	Sound speed	$m \cdot s^{-1}$
$C_{D_{EDS}}$	EDS drag coefficient	ϕ
$C_{D_{ext}}$	Drag coefficient outside the tube	ϕ
C_{D_t}	Drag coefficient inside the tube	ϕ
c_i	Wind speed induced inside the tube	$m \cdot s^{-1}$
D_c	Capsule diameter	m
D_{carga}	Diameter needed to fit the cargo	m
D_{cont}	Diameter of the circumference surrounding one container	m
D_{desp}	Displacement diameter	m
D_{movto}	Momentum diameter	m
D_t	Tube diameter	m
E_1	Phase voltage at the stator after losses	V
E_{ac}	Energy consumed during acceleration	J
E_{bat}	Energy consumed by the batteries	J
E_{gen}	Energy generated during deceleration	J
E'_t	Total energy consumed per unit length	$J \cdot m^{-1}$
E_v	Energy consumed throughout the travel at the speed *v*	J
e_{bat}	Battery stored energy per unit mass	$J \cdot kg^{-1}$
e'_t	Total energy per unit length and payload mass	$J \cdot m^{-1} \cdot kg^{-1}$
F_D	Drag force	N
$F_{D_{EDS}}$	EDS drag force	N
F_x	Propulsion force (along x axis)	N
F_z	Levitation force (along z axis)	N
f_i	Filling factor of each container (for $i = 1, 2, \dots, n_{cont}$)	ϕ
g	Gravity acceleration	$m \cdot s^{-2}$
I_1	Stator line current	A
I_c	Transport capacity per unit time (capacity index)	$kg \cdot s^{-1}$
I_e	Energy consumption per payload mass (energy index)	$J \cdot kg^{-1}$
L_{ac}	Acceleration length	m
L_c	Length of one capsule	m
L_{cont}	Length of one container	m
L_{dec}	Deceleration length	m
L_t	Tube length (same as the route one)	m
L_v	Travel length at the speed *v*	m
M	Mach number	ϕ
m_{carga}	Maximum cargo of one container	kg
$\dot{m}_{cc}$	Mass flow through the annulus	$kg \cdot s^{-1}$
m_{EB}	Emergency brakes mass	kg
m'_{EDS}	EDS magnets mass per unit length	$kg \cdot m^{-1}$
m'_{est}	Structural mass per unit length	$kg \cdot m^{-1}$
m_{Li^+}	Batteries mass	kg
$\dot{m}_t$	Mass flow through the tube (relative to vehicle)	$kg \cdot s^{-1}$
m_{tara}	Tare of one container	kg
m_{tot}	Vehicle total mass	kg
n_{cont}	Number of containers transported	ϕ
P_1	Input power to EDS	W
P_{av}	Power dissipated by running resistance	W
P_D	Power dissipated by aerodynamic drag	W
$P_{D_{EDS}}$	Power dissipated by EDS drag	W
P_x	Power really used for propulsion	W
p_t	Pressure inside the tube	Pa
p_{o_t}	Total pressure inside the tube	Pa
R	Constant for a certain ideal gas	$J \cdot kg^{-1} \cdot K^{-1}$
R_{av}	Vehicle running resistance	N
R_1	Stator resistance	Ω
T_f	Tunnel factor	ϕ
T_t	Temperature inside the tube	K
T_{o_t}	Total temperature inside the tube	K
t_{ac}	Acceleration time	s
t_{dec}	Deceleration time	s
t_{des}	Batteries discharge time	s
t_{tot}	Total route time	s
t_v	Travel time at the speed *v*	s
V_1	Phase input voltage to the stator	V
v	Vehicle operating speed	$m \cdot s^{-1}$
X_1	Stator reactance	Ω

β	Blockage ratio	ϕ
γ	Adiabatic index	ϕ
γ_o	Angle between E_1 e I_1	rad
Δ_1	Displacement section	m^2
Δ_2	Momentum section	m^2
δ^*	Boundary layer displacement thickness	m
δ_{lc}	Pod longitudinal thickness	m
δ_{rc}	Pod radial thickness	m
η_{bat}	Battery charging efficiency	ϕ (p. u.)
η_{EDS}	EDS efficiency	ϕ (p. u.)
θ	Boundary layer momentum thickness	m
ρ_t	Density inside the tube	$kg \cdot m^{-3}$
τ	Percentage of battery duration over travel time	ϕ (%)
φ	EDS power angle	rad

Table 4. Auxiliary equation blocks.

Block	Equation	Left – side variable [SI unit]	Variable definition	Equation number
Kantrowitz limit	$A_{cc} = A_t - A_c$	A_{cc} [m^2]	Annulus area	27
	$A_t = \frac{\pi}{4} D_t^2$	A_t [m^2]	Tube cross-sectional area	28
	$A_c = \frac{\pi}{4} D_c^2$	A_c [m^2]	Pod cross-sectional area	29
	$D_c = D_{carga} + 2\delta_{rc}$	D_c [m^2]	Capsule diameter	30
	$\dot{m}_t = \dot{m}_{cc_{máx}}$	$\dot{m}_t$ [$kg \cdot s^{-1}$]	Mass flow through the tube (relative to vehicle)	31
Aerodynamic drag	$F_D = \frac{1}{2}\rho_t v^2 A_f T_f C_{D_{ext}}$	F_D [N]	Drag force	32
	$P_D = F_D v$	P_D [W]	Power dissipated by aerodynamic drag	33
	$A_f = A_c$	A_f [m^2]	Frontal area projected on a plane normal to the tube	34
	$\beta = \frac{A_c}{A_t}$	β [ϕ]	Blockage ratio	35
	$\Delta_2 = \frac{\pi}{4}(D_{movto}^2 - D_c^2)$	Δ_2 [m^2]	Momentum section	36
	$D_{movto} = D_c + 2\theta$	D_{movto} [m]	Momentum diameter	37
	$D_{desp} = D_c + 2\delta^*$	D_{desp} [m]	Displacement diameter	38
Batteries	$t_{des} = \left(1 + \frac{\tau}{100}\right) t_{tot}$	t_{des} [s]	Batteries discharge time	39
Kinematics	$t_{ac} = \frac{v}{a_1}$	t_{ac} [s]	Acceleration time	40
	$t_{dec} = \frac{v}{a_2}$	t_{dec} [s]	Deceleration time	41
	$\bar{v} = \frac{\frac{v}{2}(t_{ac} + t_{dec}) + v t_v}{t_{ac} + t_{dec} + t_v}$	$\bar{v}$ [$m \cdot s^{-1}$]	Mean speed of the vehicle	42
	$t_{tot} = \frac{L_t}{\bar{v}}$	t_{tot} [s]	Total route time	43
	$t_v = t_{tot} - t_{ac} - t_{dec}$	t_v [s]	Travel time at the speed v	44
	$L_{ac} = \frac{v^2}{2a_1}$	L_{ac} [m]	Acceleration length	45
	$L_{dec} = \frac{v^2}{2a_2}$	L_{dec} [m]	Deceleration length	46
	$L_v = L_t - L_{ac} - L_{dec}$	L_v [m]	Travel length at the speed v	47
Levitation and propulsion	$F_x = m_{tot} a_1 + R_{av}$	F_x [N]	Propulsion force (along x axis)	48
	$P_x = F_x v$	P_x [W]	Power really used for propulsion	49
	$R_{av} = F_D + F_{D_{EDS}}$	R_{av} [N]	Vehicle running resistance	50
	$P_{av} = R_{av} v$	P_{av} [W]	Power dissipated by running resistance	51
	$\varphi = \sin^{-1}\left(\frac{X_1 I_1^2 + E_1 I_1 sin\gamma_o}{V_1 I_1}\right)$	φ [rad]	EDS power angle	52
	$P_1 = \frac{F_x v}{\eta_{EDS}}$	P_1 [W]	Input power to EDS	53
Transportation	$m_{tot} = n_{cont} L_c (m'_{est} + m'_{EDS}) + m_{Li^+} + m_{EB} + m_{carga} \sum_{i=1}^{i=n_{cont}} f_i + n_{cont} m_{tara}$	m_{tot} [kg]	Vehicle total mass	54
	$D_{carga} = D_{cont}$	D_{carga} [m]	Diameter needed to fit the cargo	55
	$L_c = L_{cont} + 2\delta_{lc}$	L_c [m]	Length of one capsule	56
	$I_c = \frac{m_{carga} \sum_{i=1}^{i=n_{cont}} f_i}{t_{tot}}$ (Note 1)	I_c [$kg \cdot s^{-1}$]	Transport capacity per unit time (capacity index)	57

[1] This is not the traditional capacity equation. This equation (3.57) has been specifically engineered for this problem. It assumes that only one vehicle is using the tube at a time, the one which is to be optimized.

The following table contains the unknowns that were given values to obtain the graphs. v is given four values: 700, 800, 900 and 1,000 km/h. The rest of the input data was compiled by the program as well. It should be noted that all of the power systems are predesigned for the maximum possible payload ($f_i = 1$ for i = 1, 2, …, 5) because it is the worst-case scenario for the EDS and the power system:

Table 5. Unknowns given values to obtain the plot and output variables, bolded. m_{tot} is for consultation and I_C, I_E for the curves.

Run	n_{cont} [ϕ]	f_1 [ϕ]	f_2 [ϕ]	f_3 [ϕ]	f_4 [ϕ]	f_5 [ϕ]	v [km/h]	m_{Li^+} [kg]	m_{EB} [kg]	$\mathbf{m_{tot}}$ [**kg**]	$\mathbf{I_C}$ [**t/h**]	$\mathbf{I_E}$ [**kWh/t**]	$\mathbf{I_C^{-1}}$ [**h/t**]
1	1	1	0	0	0	0	700	350	750	34,845	26.32	24.11	3.80·10-2
2	1	1	0	0	0	0	800	350	750	34,845	30.05	25.46	3.33·10-2
3	1	1	0	0	0	0	900	350	750	34,845	33.77	27.19	2.96·10-2
4	1	1	0	0	0	0	1,000	350	750	34,845	37.47	29.36	2.67·10-2
5	2	1	1	0	0	0	700	400	1,000	68,891	52.65	19.29	1.90·10-2
6	2	1	1	0	0	0	800	400	1,000	68,891	60.10	20.58	1.66·10-2
7	2	1	1	0	0	0	900	400	1,000	68,891	67.54	22.15	1.48·10-2
8	2	1	1	0	0	0	1,000	400	1,000	68,891	74.94	24.01	1.33·10-2
9	3	1	1	1	0	0	700	450	1,250	102,936	78.97	17.68	1.27·10-2
10	3	1	1	1	0	0	800	450	1,250	102,936	90.16	18.95	1.11·10-2
11	3	1	1	1	0	0	900	450	1,250	102,936	101.31	20.46	9.87·10-3
12	3	1	1	1	0	0	1,000	450	1,250	102,936	112.41	22.23	8.90·10-3
13	4	1	1	1	1	0	700	500	1,500	136,986	105.29	16.88	9.50·10-3
14	4	1	1	1	1	0	800	500	1,500	136,986	120.21	18.14	8.32·10-3
15	4	1	1	1	1	0	900	500	1,500	136,986	135.08	19.62	7.40·10-3
16	4	1	1	1	1	0	1,000	500	1,500	136,986	149.89	21.33	6.67·10-3
17	5	1	1	1	1	1	700	550	1,750	171,027	131.62	16.40	7.60·10-3
18	5	1	1	1	1	1	800	550	1,750	171,027	150.26	17.65	6.66·10-3
19	5	1	1	1	1	1	900	550	1,750	171,027	168.84	19.12	5.92·10-3
20	5	1	1	1	1	1	1,000	550	1,750	171,027	187.36	20.80	5.34·10-3

In the table below, only the unknown v is given values. This is because the rest of the values are either constants or optimized ones. The lower limit is 500 km/h, a speed reachable by state-of-the-art maglevs or even high-speed vehicles. The upper one is 1,222 km/h, around the 1,220 km/h proposed by Pellicer (2019). At 20 ºC and with $\gamma = 1.40$ and $R = 287 \frac{J}{kg \cdot K}$, $a_s = 1{,}235.53 km/h$ (by means of equation 1), which is slightly superior to 1,222 km/h and means that even if the speed were that, the vehicle would not break the sound barrier and the first hypothesis would still be true:

Table 6. At left, values given to v. At right, the output values for the variables M, D_t and β, bolded.

Run	v [km/h]	**M** [**ϕ**]	**D_t** [**m**]	**β** [**ϕ**]
1	500	0.40	6.05	3.65·10-1
2	538	0.44	6.40	3.27·10-1
3	576	0.47	6.78	2.91·10-1
4	614	0.50	7.22	2.57·10-1
5	652	0.53	7.71	2.25·10-1
6	690	0.56	8.28	1.95·10-1
7	728	0.59	8.94	1.68·10-1
8	766	0.62	9.70	1.42·10-1
9	804	0.65	10.61	1.19·10-1
10	842	0.68	11.69	9.78·10-2
11	880	0.71	13.02	7.90·10-2
12	918	0.74	14.66	6.23·10-2
13	956	0.77	16.76	4.76·10-2
14	994	0.80	19.52	3.51·10-2
15	1,032	0.84	23.33	2.46·10-2

16	1,070	0.87	28.89	1.60·10-2
17	1,108	0.90	37.78	9.37·10-3
18	1,146	0.93	54.24	4.55·10-3
19	1,184	0.96	95.00	1.48·10-3
20	1,222	0.99	364.88	1.01·10-4

The following table is a variation of table 5. Here, the L_t column has substituted the v column and there are five fewer runs because L_t adopts three values for each number of containers (15 rows in total):

Table 7. Input columns, similar to those of table 5 and output columns (bolded). $E_t^{'}$ and I_E are for reference and $e_t^{'}$ and I_C serve to elaborate the curves.

Run	n_{cont} [ϕ]	f_1 [ϕ]	f_2 [ϕ]	f_3 [ϕ]	f_4 [ϕ]	f_5	L_t [km]	m_{Li^+} [kg]	m_{EB} [kg]	$\mathbf{E_t'}$	$\mathbf{I_E}$	$\mathbf{e_t'}$	$\mathbf{I_C}$ [**t/h**]	$\mathbf{I_C^{-1}}$ [**h/t**]
1	1	1	0	0	0	0	500	350	750	1.07	18.86	3.77·10-2	42.20	2.37·10-2
2	1	1	0	0	0	0	750	350	750	0.93	24.74	$3.30 \cdot 10^{-2}$	28.19	3.55·10-2
3	1	1	0	0	0	0	1,000	350	750	0.87	30.62	$3.06 \cdot 10^{-2}$	21.16	4.73·10-2
4	2	1	1	0	0	0	500	400	1,000	1.73	15.28	$3.06 \cdot 10^{-2}$	84.40	1.18·10-2
5	2	1	1	0	0	0	750	400	1,000	1.50	19.90	$2.65 \cdot 10^{-2}$	56.38	1.77·10-2
6	2	1	1	0	0	0	1,000	400	1,000	1.39	24.53	$2.45 \cdot 10^{-2}$	42.33	2.36·10-2
7	3	1	1	1	0	0	500	450	1,250	2.39	14.08	$2.82 \cdot 10^{-2}$	126.60	7.90·10-3
8	3	1	1	1	0	0	750	450	1,250	2.07	18.29	$2.44 \cdot 10^{-2}$	84.57	1.18·10-2
9	3	1	1	1	0	0	1,000	450	1,250	1.91	22.50	$2.25 \cdot 10^{-2}$	63.49	1.58·10-2
10	4	1	1	1	1	0	500	500	1,500	3.05	13.49	$2.70 \cdot 10^{-2}$	168.80	5.92·10-3
11	4	1	1	1	1	0	750	500	1,500	2.64	17.48	$2.33 \cdot 10^{-2}$	112.76	8.87·10-3
12	4	1	1	1	1	0	1,000	500	1,500	2.43	21.48	$2.15 \cdot 10^{-2}$	84.65	1.18·10-2
13	5	1	1	1	1	1	500	550	1,750	3.72	13.13	$2.63 \cdot 10^{-2}$	211.01	4.74·10-3
14	5	1	1	1	1	1	750	550	1,750	3.21	17.00	$2.27 \cdot 10^{-2}$	140.95	7.09·10-3
15	5	1	1	1	1	1	1,000	550	1,750	2.95	20.87	$2.09 \cdot 10^{-2}$	105.81	9.45·10-3

Three-blind validation of a Deep Learning model to obtain the dense tissue area in digital mammographies

Francisco Javier Pérez-Benito, Andrés Larroza, Juan-Carlos Perez-Cortes, and Rafael Llobet

Instituto Tecnológico de la Informática, Universitat Politècnica de València, Cami de vera, S/N, 46022 València, Spain
fjperez@iti.es

Abstract. Although mortality due to breast cancer has significantly decreased in recent years, the incidence continues to increase, making screening programs crucial. In these programs, mammograms are collected from women over 40-45 years of age, which are then assessed by radiologists. The most relevant image biomarker of the risk of developing breast cancer is dense tissue, but its evaluation is time-consuming and has a subjective component. Approaches based on Machine Learning would reduce the analysis time; still, the subjective component challenges automatic methods to deal with the imperfect supervision problem. Moreover, different mammographic devices and acquisition protocols may influence the image and its analysis. In such a way, a conscious validation of automated technologies becomes critical.

Keywords: Deep Learning, Breast Cancer, Digital mammographies, Dense Tissue segmentation

1 Introduction

Breast Cancer (BC) is among the most frequent in women worldwide [1, 2]. In recent decades, mortality has decreased, but exceptions are still observed in countries without early detection policies [3]. Among others, the decrease in fertility rates, the increment of the use of hormonal menopausal therapy or oral contraceptives, and the reduction of breastfeeding time are factors that influence the increasing incidence rates in developed countries [4]. This way, screening programs remain vital in the fight against this pathology.

Breasts are mainly composed of fat and dense tissue, and breast density is consistently associated with the risk of developing breast cancer [5], [6]. Breast density assessment in screening programs is the front line in the war against the disease. The evaluation of mammographies is, beyond time-consuming [7], an imperfect supervision problem from a data science view because of the lack of a ground-truth [8,9] what directly challenges artificial intelligence approaches to deal with this task. Another thing that challenges these approaches is the variability of the images depending on the acquisition device and protocols [10].

In our previous work [11], we designed a model (CM-YNet) to deal with these two concerns. The model uses different specialists' opinions to build the ground truth to train the system. Besides inferring the dense tissue in a mask-based approach, replicating the working of DM-Scan [12], a parametric model branch estimates the values needed to interact with the dense tissue mask. After two conventional validations with promising results, this document shows the results of an experiment involving three radiologists. We hypothesize that one specialist will agree with our model as much as agrees with two other specialists in a blind way.

2 Background

The objective is to validate a model exhaustively to segment the dense tissue in digital mammographies. The main challenges of this task are the differences in the images among acquisition devices and the existence of inter and intra-reader variability [13, 14]. To understand the scope of this communication, this section describes: (1) the datasets used for each of the stages of the technology to be validated and (2) a summary of our previous work covering the main features and results of the model (CM-YNet) [15].

2.1 Datasets

Three datasets were considered in the framework of this work. The first dataset was used to train and make a conventional validation of the model. The second was used for external validation, and the images originated from another location. Finally, the third dataset was used to carry out the three-blind validation that gives its name to this communication.

- **D1.** A total of 3340 images from cranio-caudal and mediolateral oblique projections of the breasts of 1785 women were used to train and validate the model. The mammograms were acquired from 11 centers of the Valencian screening program, covering devices from 6 different vendors. Two radiologists segmented all these images.
- **D2.** A set of 381 images from the cranio-caudal projection of different women. Only cranio-caudal projection was used to reduce variability in the validation. The images were randomly extracted by our partner *Institut Hospital del Mar d'Investigacions Mèdiques (IMIM)*. The acquisition period of the images was selected to obtain 283 out of 381 mammograms from old devices with lower quality making the segmentation task more challenging. Two specialists also segmented these mammographies.
- **D3.** A new set of 500 cranio-caudal projections were extracted from IMIM to carry out the exhaustive validation of the model. To ensure the model is validated covering a wide range of devices (and then image quality), the extraction period was fixed to 2011-2021. Two radiologists also segmented these mammograms.

Figure 1 shows the distribution of the D3 images according to the year and device vendor of acquisition. Our previous work [11] details the datasets D1 and D2.

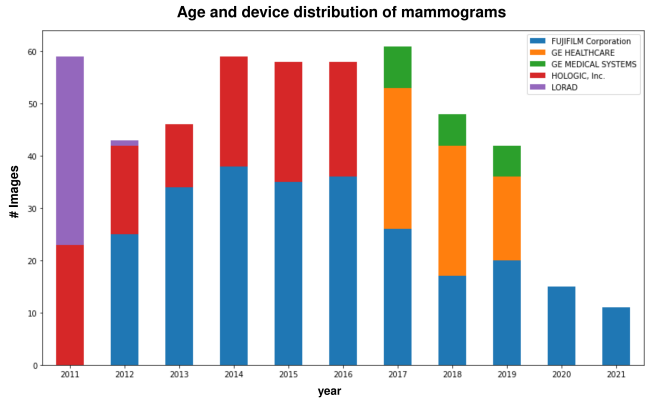


Fig. 1: Distribution of the mammograms according the year and acquisition vendor.

2.2 CM-YNet

The CM-YNet [11], whose architecture is shown in Figure 2, is a Deep Learning (DL) model that automatically infers dense tissue in digital mammographies.

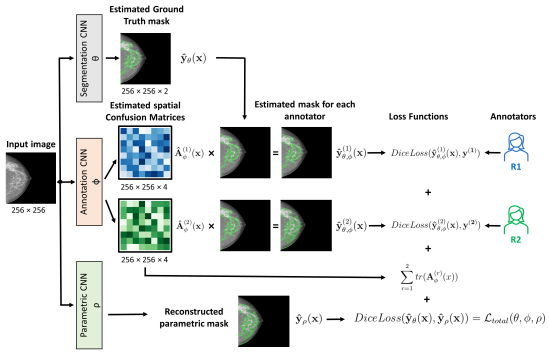


Fig. 2: CM-YNet architecture. The first branch models the dense tissue ground truth using the confusion matrix of the different labelers. The last branch models a parametric approach allowing potential users to interact with the mask.

To measure the model's goodness, we use the DICE score [15], which ranges between 0 and 1 and measures the matching between two segmentation masks. The more similar the two masks are, the closer to 1 is the DICE score. Since this is an "imperfect supervision problem" with the opinion of two specialists, we compare the DICE between the two specialists' masks to the DICE of the output of our model against the mask of the more similar specialist (there is no way to state which radiologist's mask is better). The main contributions and results are the following:

1. The images were processed to reduce variability among different acquisition devices. Besides, each image was segmented by two experienced radiologists.
2. The model's training used 70% of the D1 images, covering 11 different acquisition centers and six different vendors. The results in the rest of the D1 images suggested the generalization capability of CM-YNet. See Table 1.
3. An external validation of the model was carried out with the images of D2. These images were acquired from another location, and then the acquisition devices and configuration differed from those of D1. The results in Table 1 confirm the model's performance.
4. Given the existence of variability in the specialists' opinion, the last branch of the model tries to simulate a parametric approach [16] that would enable the option of interacting with the segmentation area.

Dataset	**Center Id**	**#images**	**R1 vs. R2**	**CM-YNet (param.)**	**CM-YNet (mask)**
D1	01	96	079 ± 0.16	0.75 ± 0.19	**0.81 ± 0.11**
D1	02	96	0.79 ± 0.14	0.81 ± 0.15	**0.83 ± 0.13**
D1	04	34	0.75 ± 0.17	0.74 ± 0.20	**0.83 ± 0.08***
D1	05	80	0.64 ± 0.17	0.81 ± 0.16	**0.84 ± 0.10**
D1	07	14	**0.88 ± 0.15**	0.73 ± 0.18	0.82 ± 0.14
D1	10	156	0.77 ± 0.16	0.79 ± 0.15	**0.85 ± 0.10***
D1	11	140	0.82 ± 0.12	0.84 ± 0.10	**0.87 ± 0.07**
D1	13	60	0.78 ± 0.12	0.82 ± 0.13	**0.86 ± 0.11**
D1	18	20	0.74 ± 0.14	0.80 ± 0.13	**0.86 ± 0.08***
D1	20	90	0.78 ± 0.16	0.78 ± 0.16	**0.83 ± 0.12***
D1	99	58	0.79 ± 0.13	0.83 ± 0.13	**0.89 ± 0.09***
D2	21	98	0.76 ± 0.14	0.82 ± 0.17	**0.86 ± 0.10**
D2	22	283	0.71 ± 0.22	0.72 ± 0.19	**0.74 ± 0.19***
Total	1225	0.76 ± 0.17	0.73 ± 0.22	0.78 ± 0.17	**0.82 ± 0.14**

Table 1: Comparison of DICE scores according to the different acquisition centers. Images from the centers 21 and 22 are from the independent IMIM dataset in which id 22 correspond to images acquired with old devices. The highest value for each row is highlighted in bold. * The difference between CM-YNet (mask) and ECNN is statistically significant ($p < 0.001$).

The results show that the model performs well for both the parametric and mask branches. The model obtains a DICE in most devices even higher than that obtained between the two specialists. The interpretation could be that the radiologists usually agree more with the CM-YNet segmentation than the other radiologist. Since the problem lacks a ground truth, the next step would be to verify that a specialist agrees with CM-YNet as much as with another specialist, which introduces the concept of **three-blind validation** that is detailed in the following Section.

3 Three-blind validation

Once the model was trained and validated using a partition of D1, the performance of the model was tested using an external dataset, D2. The promising results motivated us to design an exhaustive validation in which three radiologists would collaborate, the three-blind validation. The first step was the extraction of a new dataset consisting of 500 left cranio-caudal images. These images were labeled by two specialists (L1 and L2) and by CM-YNet, obtaining a total of 1500 segmentations; moreover, 300 segmentations were duplicated (1800 segmentations) to measure the intra-reader variability of a third experienced radiologist (L3) assessing if agrees or disagrees with each of the randomly presented segmentations.

Figure 3 shows a diagram of the experiment. This experiment allows us to identify if our model is as good a labeler as any other specialist, making CM-YNet target of becoming a powerful tool in the day-to-day of the practitioners of the screening programs.

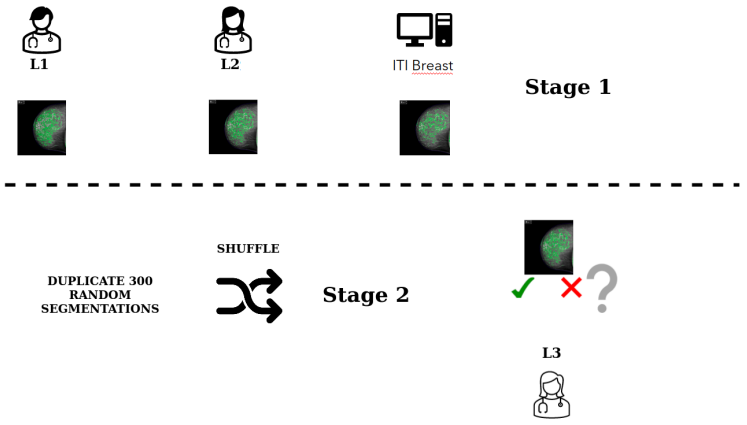


Fig. 3: Stage 1 corresponds to the labeling stage, and Stage 2 is the "blind-validation".

3.1 Results and discussion

The 300 duplicated images to analyze the intra-observer variability demonstrate the consistency of L3, as can be seen in the confusion matrix of Figure 4. L3 only showed inconsistency in 23 out of the 300 segmentations, thus the results of the validation can be considered robust.

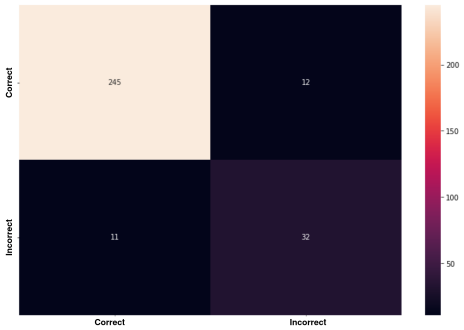


Fig. 4: The confusion matrix shows the L3 agreement of the 300 duplicated segmentations. The inconsistency rate was 7.67%.

Figure 5 shows the percentage of concordance between L3 and the labelers: L1, L2 and CM-YNet. The agreement of L3 with L1 and L2 is similar, and significantly better than the obtained by CM-YNet. It suggests that even though the DICE score, in mean, of the mask of a specialist segmentation and CM-YNet segmentation is very high, there must be some artifact in the automatic model segmentations that makes L3 disagree more with CM-YNet than with the other specialists.

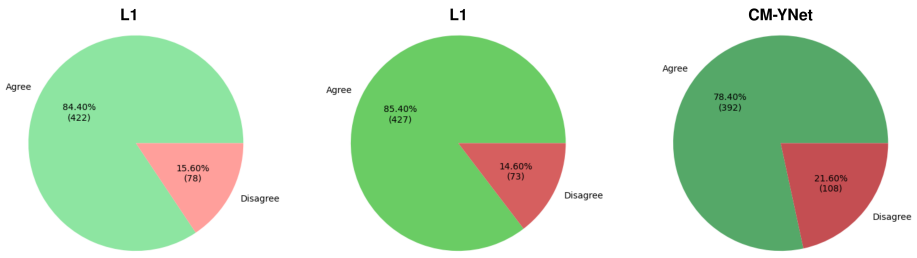


Fig. 5: Comparison of L3 agreement rate with L1, L2 and CM-YNet labelers.

The Breast Imaging Reporting and Data System (BIRADS) [17] of the American College of Radiology tries to standardize the way mammographies are read. The BIRADS proposes categorizing density into four classes, from almost fatty breasts to extremely dense ones. Although the last recommendations suggest not only the quantity of dense tissue but also other features such as shape or position influence the risk of developing breast cancer, there is no doubt that almost fatty breasts and those extremely dense are the easiest to read [18]. In this sense, we expected L3 to disagree more with our model in the images that are more difficult to read. We analyze the DICE score of our model against L1 and L2, splitting the results by the quartile of the density read by the specialists as seen in Table 2.

	L1 vs L2	**L1 vs. CM-YNet**	**L2 vs. CM-YNet**	**Closest vs. CM-YNet**
PDL1-Q1	0.688 ± 0.208	0.470 ± 0.159	0.578 ± 0.191	0.593 ± 0.180
PDL1-IQR	0.799 ± 0.132	0.749 ± 0.107	0.779 ± 0.105	0.807 ± 0.087
PDL1-Q4	0.873 ± 0.087	0.873 ± 0.069	0.837 ± 0.092	0.887 ± 0.062
PDL2-Q1	0.754 ± 0.195	0.511 ± 0.180	0.542 ± 0.170	0.575 ± 0.166
PDL2-IQR	0.798 ± 0.147	0.741 ± 0.138	0.810 ± 0.082	0.823 ± 0.079
PDL2-Q4	0.809 ± 0.141	0.849 ± 0.096	0.813 ± 0.096	0.873 ± 0.067

Table 2: Comparison of the DICE score of the 500 validation images split by the DICE score quartile of CM-YNet against the labelers L1 and L2. PDLi means *Percent Density reader Li*, Q1 represents the first quartile, IQR is the interquartile range, and Q4 is the fourth quartile.

Contrary to the hypothesis, the model has a lower DICE in the less dense breasts, which might suggest that some kind of artifact in these images is influencing the results. After a visual review of the mammograms, we hypothesize that the identification of the breast could be improved by eliminating areas such as the pectoral muscle, axilla, or abdomen, which, due to their composition, are brighter and could be considered dense tissue by the automatic model as can be seen in Figure 6.

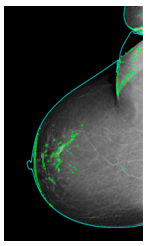


Fig. 6: Visual impact of considering external artifacts in dense tissue segmentation.

Under this hypothesis, we propose improving the identification of the breast, which is an image preprocessing before applying the CM-YNet model [11]. This preprocessing does not contemplate the existence of pectoral muscle and armpit in cranio-caudal images nor considers the presence of the abdomen in any breast projection. Several approaches have arisen to deal with this task since the CM-YNet article's publication, among them the "Segment Anything" Model [19, 20] stands out. After the correct automatic breast detection in mammographies, the three-blind validation should be repeated to assess the improvement in the results, making CM-YNet an excellent candidate for decision support in breast cancer screening programs.

4 Conclusions

The incidence of breast cancer, the existence of screening programs to control the pathology, and the lack of standardization in mammogram reading make important the emergence of tools to help practitioners in their work. The CM-YNet model provides a robust way to quantify the dense tissue in digital mammographies, with the added value of being possible to modify the segmentation thanks to its parametric branch. The model obtained excellent results in two independent conventional validations. What motivated us to carry out a more exhaustive verification, the "three-blind validation", to answer the question, "Is our model as good a labeler as a radiologist?". The poorer performance of CM-YNet in the images with the less dense breasts has suggested that it may not be a problem in the segmentation of the dense tissue but in the appearance of bright elements in the mammogram that have nothing to do with the breast tissue. The previous suggests a need to improve breast detection by excluding areas of no interest and to re-conduct the validation outlined in this document.

Acknowledgements

The authors of this work like to thank Marta Román, Javier Louro, and the rest of the partners from Hospital del Mar for their willingness to work, obtaining different datasets and clinical analysis of the results.

References

1. G. N. Sharma, R. Dave, J. Sanadya, P. Sharma, and K. Sharma, "Various types and management of breast cancer: an overview," Journal of advanced pharmaceutical technology & research, vol. 1, no. 2, p. 109, 2010.
2. E. J. Watkins, "Overview of breast cancer," Journal of the American Academy of PAs, vol. 32, no. 10, pp. 13–17, 2019.
3. C. Wojtyla, P. Bertuccio, A. Wojtyla, and C. La Vecchia, "European trends in breast cancer mortality, 1980–2017 and predictions to 2025," European Journal of Cancer, vol. 152, pp. 4–17, 2021.

4. S. Lei, R. Zheng, S. Zhang, S. Wang, R. Chen, K. Sun, H. Zeng, J. Zhou, and W. Wei, "Global patterns of breast cancer incidence and mortality: A populationbased cancer registry data analysis from 2000 to 2020," Cancer Communications, vol. 41, no. 11, pp. 1183–1194, 2021.
5. C. M. Vachon, K. R. Brandt, K. Ghosh, C. G. Scott, S. D. Maloney, M. J. Carston, V. S. Pankratz, and T. A. Sellers, "Mammographic breast density as a general marker of breast cancer risk," Cancer Epidemiology Biomarkers & Prevention, vol. 16, no. 1, pp. 43–49, 2007.
6. N. F. Boyd, L. J. Martin, M. Bronskill, M. J. Yaffe, N. Duric, and S. Minkin, "Breast tissue composition and susceptibility to breast cancer," Journal of the National Cancer Institute, vol. 102, no. 16, pp. 1224–1237, 2010.
7. C. K. Kuhl, "The changing world of breast cancer: a radiologist's perspective," Investigative radiology, vol. 50, no. 9, p. 615, 2015.
8. E. F. Conant, D. Li, S. Gavenonis, P. R. Bakic, A.-K. Carton, C. Zhang, A. D. Maidment, and D. Kontos, "A comparative study of the inter-reader variability of breast percent density estimation in digital mammography: potential effect of reader's training and clinical experience," in Digital Mammography: 10th International Workshop, IWDM 2010, Girona, Catalonia, Spain, June 16-18, 2010. Proceedings 10, pp. 114–120, Springer, 2010.
9. K. Pesce, M. Tajerian, M. Chico, M. Swiecicki, B. Boietti, M. Frangella, and S. Benitez, "Interobserver and intraobserver variability in determining breast density according to the fifth edition of the bi-rads® atlas," Radiología (English Edition), vol. 62, no. 6, pp. 481–486, 2020.
10. C. F. de Vries, S. J. Colosimo, R. T. Staff, J. A. Dymiter, J. Yearsley, D. Dinneen, M. Boyle, D. J. Harrison, L. A. Anderson, G. Lip, et al., "Impact of different mammography systems on artificial intelligence performance in breast cancer screening," Radiology: Artificial Intelligence, vol. 5, no. 3, p. e220146, 2023.
11. A. Larroza, F. J. Pérez-Benito, J.-C. Perez-Cortes, M. Román, M. Pollán, B. Pérez-Gómez, D. Salas-Trejo, M. Casals, and R. Llobet, "Breast dense tissue segmentation with noisy labels: A hybrid threshold-based and mask-based approach," Diagnostics, vol. 12, no. 8, p. 1822, 2022.
12. M. Pollán, R. Llobet, J. Miranda-García, J. Antón, M. Casals, I. Martínez, C. Palop, F. Ruiz-Perales, C. Sánchez-Contador, C. Vidal, et al., "Validation of dmscan, a computer-assisted tool to assess mammographic density in full-field digital mammograms," Springerplus, vol. 2, no. 1, pp. 1–13, 2013.
13. A. Redondo, M. Comas, F. Macia, F. Ferrer, C. Murta-Nascimento, M. Maristany, E. Molins, M. Sala, and X. Castells, "Inter-and intraradiologist variability in the bi-rads assessment and breast density categories for screening mammograms," The British journal of radiology, vol. 85, no. 1019, pp. 1465–1470, 2012.
14. Z. Gandomkar, S. Siviengphanom, M. Suleiman, D. Wong, W. Reed, E. U. Ekpo, D. Xu, S. J. Lewis, K. K. Evans, J. M. Wolfe, et al., "Reliability of radiologists' first impression when interpreting a screening mammogram," Plos one, vol. 18, no. 4, p. e0284605, 2023.
15. A. Carass, S. Roy, A. Gherman, J. C. Reinhold, A. Jesson, T. Arbel, O. Maier, H. Handels, M. Ghafoorian, B. Platel, et al., "Evaluating white matter lesion segmentations with refined sørensen-dice analysis," Scientific reports, vol. 10, no. 1, p. 8242, 2020.
16. F. J. Pérez-Benito, F. Signol, J.-C. Perez-Cortes, A. Fuster-Baggetto, M. Pollan, B. Perez-Gómez, D. Salas-Trejo, M. Casals, I. Martínez, and R. LLobet, "A deep learning system to obtain the optimal parameters for a threshold-based breast and

dense tissue segmentation," Computer Methods and Programs in Biomedicine, vol. 195, p. 105668, 2020.
17. E. F. Conant, B. L. Sprague, and D. Kontos, "Beyond bi-rads density: a call for quantification in the breast imaging clinic," Radiology, vol. 286, no. 2, pp. 401–404, 2018.
18. L. E. Philpotts, "Advancing artificial intelligence to meet breast imaging needs," 2022.
19. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., "Segment anything," arXiv preprint arXiv:2304.02643, 2023.
20. Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, et al., "Segment anything model for medical images?," arXiv preprint arXiv:2304.14660, 2023.

Scatter and random correction of PET list-mode data using machine learning approaches

Joan Prats-Climent[1,†], Filomeno Sánchez[†], Hector Espinós-Morató[♭], Antonio Javier González[†], María José Rodríguez-Álvarez[†]

(†) Instituto de Instrumentación para Imagen Molecular,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(♭) Universidad Europea de Valencia,
Paseo de la Alameda 7, 46010 València, Spain

Abstract. A novel approach is presented in this study, proposing a method for effectively performing scatter and random correction of list-mode data within a simulated PET ring system, based on machine learning algorithms. Using positional and energetic information from both photons stored in the detector we are able to classify and discard non-true coincidences in order to enhance image quality. Two machine learning algorithms have been tested, based on decision trees and neural networks respectively, and lead to high accuracies in the classification task. A significant reduction in the number of scattered and random coincidences is confirmed. To assess the improvement in image quality, several image quality metrics were investigated. Root Mean Square Error (RMSE) exhibited a decrease, while Peak Signal-to-Noise Ratio (PSNR) showed an increase in both corrected images. These findings collectively indicate higher image quality and improved image contrast resulting from the correction process.

Keywords: Machine learning, Positron Emission Tomography, Scatter correction, Random correction.

1 Introduction

Positron emission tomography (PET) is a medical imaging technique that can be used to visualize the metabolic and biological function of the human body. This imaging modality is commonly used to diagnose and monitor the treatment of cancer, heart disease, and neurological disorders, among other medical uses. The performance of quantitative measurements with this technique requires a careful understanding of the imaging process and all sources of error that distort the true physiological information under study. One such source of error in PET image reconstruction is the presence of scattered and random coincidences, leading to a loss of image contrast and inaccurate quantification of activity. To address this

[1] joapracl@i3m.upv.es
Funded by PID2019-107790RB-C2 MCIN/ AEI/10.13039/501100011033

issue and enhance the overall image quality, it is essential to identify and remove these scattered and random coincidences. We propose the use of machine learning algorithms for this task, using positional and energetic information from both photons in the coincidence as input data. Once the algorithms are trained and tested on new data, we are able to identify each event and discard the randoms and scattered ones before the reconstruction of the list-mode dataset to improve image quality.

2 Materials and Methods

2.1 Simulation

In our work, we simulated a PET scanner based on the real system SIEMENS Biograph 6 TruePoint [1] using the software GATE [2] (*GEANT4 Application for Tomographic Emission*), a toolkit for Monte Carlo simulations in medical physics. This detector consists of three block rings of 421 mm scanner radius, with each ring consisting of 48 blocks with 4 x 4 x 20 mm^3 LSO crystals resulting in a 162 mm axial and 585 mm transaxial field-of-view (FOV) respectively. A Jaszczak-like phantom is placed in the center of the FOV, with six inserted spheres with different diameters in a cylinder as hot regions. An energy window of 350-650 keV and a time coincidence window of 5 ns were chosen.

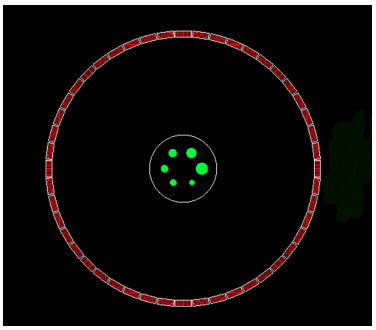


Fig. 1: PET System Simulation: Siemens Biograph TruePoint with Jaszczak phantom.

2.2 Supervised Machine Learning: Classification

We propose a new approach for performing the correction of scattered and random coincidendes in the detector. Using energy and interaction position of both photons in the coincidence obtained in the simulation as input features, we apply Machine Learning algorithms to discard non-true coincidences and enhance

signal from background. In this framework, we perform a supervised binary classification, where we label each coincidence as 'True", if it is a signal coincidence, or "False", if it is a scatter or random coincidence.

The data were split into 80% for training and 20% for testing. Information regarding to the interaction position in x, y and z axes and energy from both photons given by the simulation are used as input data. The label of each coincidence (True or False) used in the training stage is obtained from GATE and will be predicted in the testing stage. Once the algorithms were trained and tested, they are applied to a new simulation with the same characteristics.

In order to classify data coming from simulation and based in previous studies [3] on the feasibility of machine learning algorithms for this task, we have implemented two machine learning algorithms using the Python libraries XGBoost [4] and Scikit-learn [5], which we refer as:

- Extreme Gradient Boosting (**XGB**): A decision tree-based ensemble algorithm.
- Neural Network (**NN**): A feedforward artificial neural network with 2 hidden layers (input-hidden-hidden-output).

2.3 Image Reconstruction

After the successful training and testing of the algorithms, they are subsequently applied to a new simulation with identical characteristics to ensure consistency and reliability of the correction process. Upon correcting the new dataset, the images are reconstructed using the CASToR (*Customizable and Advanced Software for Tomographic Reconstruction*) software [6] with an Maximum Likelihood Expectation Maximization (MLEM) algorithm. To validate our correction method, several metrics are used in the assesment of the image quality and are calculated comparing the measured and corrected reconstructed images to the ideal image (benchmark using only true coincidences and excluding all scattered and random coincidences). The chosen metrics are defined as:

- Root mean squared error (**RMSE**):

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(u(n) - u_{true}(n))^2} \tag{1}$$

- Peak signal-to-noise ratio expressed in dB (**PSNR**):

$$PSNR = 10 \cdot \log_{10}\left(\frac{max(u_{true}(n))^2}{\sum_{n=1}^{N}(u(n) - u_{true}(n))^2}\right) \tag{2}$$

where u denotes reconstructed image, u_{true} denotes ideal image, $max(u_{true})$ denotes the maximum value of intensity of the ideal image. This metrics were

calculated on a volume-of-interest (VOI) defined by the voxels occupied by the ideal image, with n indicating the voxel index and N the total number of voxels.

RMSE approaches to zero when reconstructed and ideal image are more identical with respect to voxel difference inside the VOI. Consequently, PSNR increases beacuse RMSE decreases. In the absence of noise, both images are equal, and thus the RMSE is zero and the PSNR is infinite. Then a lower RMSE and higher PSNR generally indicates that the reconstruction is of higher quality.

3 Results

Once trained, our models were used to classify the coincidences in the new samples and discard coincidences labeled as False, in order to reduce scatter and random fraction and enhance the true coincidence rate in our dataset. Table 1 presents the results, showcasing a substantial reduction in the number of scattered and random coincidences, while the majority of signal events were preserved. However, a minor portion of true coincidences was misclassified.

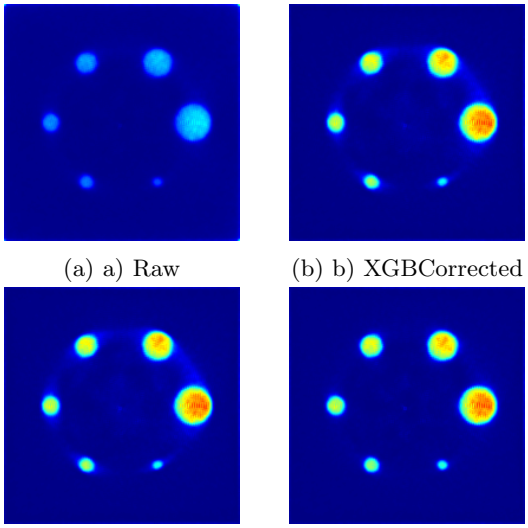
Algorithm		True	Scatter	Random
	Raw Counts	4112118	2937868	316324
	% of dataset	(55.82%)	(39.88%)	(4.29%)
XGB	Corrected	3931065	913719	25817
	% of dataset	(80.71%)	(18.76%)	(0.53%)
NN	Corrected	3948644	924378	26141
	% of dataset	(80.60%)	(18.87%)	(0.53%)

Table 1: Scatter and random reduction for the proposed algorithms.

Once the datasets are corrected, the measured, corrected and ideal images are reconstructed (Fig. 2) and the quality metrics are calculated using the ideal image as reference. Values for the quality metrics are shown in Table 2. The results demonstrate that the corrected images exhibit a notable decrease in RMSE and an increase in PSNR compared to the raw image. This suggests that the corrected images possess higher quality and improved image contrast.

4 Conclusion

We have successfully validated a method for scatter and random correction of PET images prior to image reconstruction. This innovative approach relies on binary classification through supervised machine learning algorithms applied to list-mode data. In our experimental study, both the XGBoost (XGB) and Neural



(a) a) Raw (b) b) XGBCorrected

(c) c) NNCorrected (d) d) True

Fig. 2: Reconstructed central slices of the Jaszczak phantom for iteration 24 of a) Raw measured image, b) XGB corrected image, c) NN corrected image and d) True image.

Network (NN) algorithms demonstrated similar performance, effectively achieving a substantial reduction in the number of scattered coincidences (from 40% to 19%) and random (from 4% to 0.5%) coincidences. Improvement of image quality has been assesed with decreasing RMSE and increasing PSNR of both corrected images with respect to the measured one.

The proposed algorithms successfully contribute to reducing scatter and random coincidences, thereby enhancing the accuracy and reliability of image reconstruction in PET, offering potential advancements in nuclear medicine for improved diagnostic capabilities and research outcomes. Machine learning techniques show up as a powerful alternative in performing PET image corrections. At the moment this algorithm has been tested only at simulation level, with next step being application to real PET data, and compare with state-of-the-art correction algorithms.

References

1. F. A. Kotasidis et al., 'Single scan parameterization of space-variant point spread functions in image space via a printed array: the impact for two PET/CT scanners', Phys. Med. Biol., vol. 56, pp. 2917-2942, 2011.
2. S. Jan et al., 'GATE: a simulation toolkit for PET and SPECT', Phys. Med. Biol., vol. 49, pp. 4543-4561, 2004.

Image	RMSE	PSNR [dB]
Raw image	4.4 x 10^{-5}	29.1
XGB image	2.5 x 10^{-5}	34.1
NN image	2.6 x 10^{-5}	33.8

Table 2: RMSE and PSNR of raw and corrected images using ideal image as reference.

3. J. Prats et al., 'PET scatter correction using machine learning techniques', 2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Manchester, UK, pp. 1-3, 2019.
4. T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
5. F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python', J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.
6. T. Merlin et al., 'CASToR: a generic data organization and processing code framework for multi-modal and multi-dimensional tomographic reconstruction', Phys. Med. Bio., vol. 63, no. 18, pp. 5505, 2018.

The Simplified Double PN approximation for the Neutron Transport equation

A. Vidal-Ferràndiz[1], A. Carreño[1], D. Ginestar[3], and G. Verdú[1]

[1] Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València,
Camí de Vera, s/n. 46022 València, Spain.
[2] Instituto Universitario de Seguridad Industrial, Radiofísica y Medioambiental (ISIRYM),
Universitat Politècnica de València,
Camí de Vera, s/n. 46022 València, Spain.

Abstract. The behavior of the neutrons inside a nuclear reactor is determined by the neutron transport equation. This equation demands high computational resources to solve it without drastic approximations. In this work, the Simplified Double P_N (SDP_N) approximation is derived. This approximation reduce the neutron transport equation to a system of diffusion-like equations. Preliminary numerical results suggest that when strong spatial heterogeneities are present, SDP_N equations obtain more accurate results than the diffusion approximation or the Simplified P_N equation of the same order with less computational time.

Keywords: Neutron Tansport,Double PN equations

1 Introduction

The behavior of a nuclear reactor is described by the neutron distributions in the reactor as a function of position, energy, and time. Their behavior is modeled by the neutron transport equation [3].

$$\begin{aligned}\frac{1}{\upsilon}\frac{\partial \Psi}{\partial t}(\vec{r},\, E,\, \vec{\Omega},\, t) = &- \vec{\Omega}\cdot\vec{\nabla}\,\Psi(\vec{r},\, E,\, \vec{\Omega},\, t) - \Sigma_t(\vec{r},\, E,\, t)\,\Psi(\vec{r},\, E, \vec{\Omega},\, t)\\ &+ \int_0^\infty \mathrm{d}E' \int_{(4\pi)} \mathrm{d}\Omega'\, \Sigma_s(\vec{r},\, E' \to E,\, \vec{\Omega}'\cdot\vec{\Omega},\, t)\,\Psi(\vec{r}, E', \vec{\Omega}', t)\\ &+ \frac{1}{k_{eff}}\,\frac{\chi^p(E)}{4\pi}\int_0^\infty \nu\Sigma_f(\vec{r},\, E',\, t)\,\Phi(\vec{r},\, E',\, t)\mathrm{d}E'\end{aligned} \tag{1}$$

For an entire reactor core without spatial homogenization, the solution of the neutron transport equation require enormous computational resources. In this way, approximations to these equations are must be used. The neutron diffusion approximation provides accurate enough results at a reasonable computational

cost for traditional nuclear reactors. More accurate approximations as the discrete ordinates, method of characteristics, spherical harmonics, etc; have to be considered when novel fuel types and setups are analyzed.

A less computational demanding approximation is the a simplified formulation for the P_N equations, known as the Simplified P_N equations (SP_N). This equation relies on the expansion of the angular flux in terms of Legendre polynomials, where the neutron flux moments are assumed continuous. However, the presence of heterogeneous materials give rise to discontinuous angular fluxes and motivate the introduction of the Double P_N approximation. In this work, the Simplified Double spherical harmonics equations is studied.

The rest of this work is organized as follows. Firstly, the $\mathrm{P_N}$ equations and the Simplified $\mathrm{P_N}$ equation are explained. Secondly, the Double $\mathrm{P_N}$ and its simplifications are introduced. Then, preliminary numerical results on a highly heterogeneous problem are presented. Finally the main conclusion of the paper are summarized.

2 $\mathbf{P_N}$ equations

To develop the spherical harmonics approximation ($\mathrm{P_N}$ equations), our starting point is the neutron transport equation for a slab geometry,

$$\mu\frac{d}{dx}\psi_g(x,\mu)+\Sigma_{tg}(x)\psi_g(x,\mu)-\sum_{g'=1}^{G}\int_0^{2\pi}d\varphi'\int_{-1}^{+1}\Sigma_s^{g\to g'}(x,\mu_0)\,\psi(x,\mu_0)\,d\mu' = \frac{1}{2}\frac{\chi_g}{k_{\mathrm{eff}}}\sum_{g'=1}^{G}\nu_{g'}\Sigma_{fg'}(x)\int_{-1}^{+1}\Sigma_{g'}(x,\mu')\,d\mu' \quad (2)$$

We assume that the angular neutron flux can be expanded.

$$\psi_g(x,\mu)=\sum_{l=0}^{\infty}\phi_{l,g}(x)\left(\frac{2l+1}{2}\right)P_l(\mu)\ . \quad (3)$$

Then, the neutron transport equation:

$$\sum_{l=0}^{\infty}\frac{d}{dx}\phi_{l,g}(x)\frac{1}{2}\left((l+1)P_{l+1}(\mu)+lP_{l-1}(\mu)\right) +\Sigma_{tg}(x)\sum_{l=0}^{\infty}\phi_{l,g}(x)\left(\frac{2l+1}{2}\right)P_l(\mu) = \sum_{g'=1}^{G}\sum_{l=0}^{\infty}\left(\frac{2l+1}{2}\right)\Sigma_{sl}^{g\to g'}(x)\,P_l(\mu)\phi_{l,g}(x)+\frac{\chi_g}{2k_{\mathrm{eff}}}\sum_{g'=1}^{G}\nu_g\Sigma_{fg'}(x)\phi_{0,g'}(x)\ . \quad (4)$$

where we have used the orthogonality relations for the Legendre polynomials

$$\int_{-1}^{+1}P_m(\mu)P_n(\mu)\,d\mu=\frac{2}{2n+1}\delta_{m,n}, \quad (5)$$

Multiplying equation (4) by $P_k(\mu)$, integrating from -1 to $+1$ and using the orthogonality relation, we obtain the P_N equations:

$$\frac{k}{(2k+1)}\frac{d}{dx}\phi_{k-1,g}(x) + \frac{k+1}{(2k+1)}\frac{d}{dx}\phi_{k+1,g}(x) + \Sigma_{t,g}(x)\phi_{k,g}(x)$$
$$= \sum_{g'=1}^{\infty} \Sigma_{sk}^{g\to g'}(x)\,\phi_{k,g'}(x) + \frac{\chi_g}{k_{\text{eff}}}\sum_{g'=1}^{G}\nu_{g'}\Sigma_{fg'}(x)\phi_{0,g'}(x)\delta_{0,k}\ . \tag{6}$$

This set of equations is composed of $N+1$ equations with $N+2$ unknowns. This is usually solved, imposing the condition $\frac{d}{dx}\phi_{N+1,g}=0$ as a cloure equation.

In multidimensional geometries, the derivatives with respect to x are substituted by a gradient operator (Simplified $\mathrm{P_N}$). In this way, a system of diffusive-like equations are obtained. The theoretical basis for this ad-hoc approximation equations were provided by Brantley and Larsen (2000) [1], showing that these equations are high-order asymptotic solutions of the transport equation when diffusion theory is the leading-order approximation, as it is the case for nuclear systems.

3 Double $\mathbf{P_N}$ equations

To improve the convergence in problem where the flux becomes discontinuous, two sets of polynomials can be introduced, leading to the Double $\mathrm{P_N}$ ($\mathrm{DP_N}$) equations.

$$P_l^+(\mu) = \begin{cases} P_l(2\mu-1) & \text{if } \mu \geq 0 \\ 0 & \text{if } \mu < 0 \end{cases}\ , \quad P_l^-(\mu) = \begin{cases} P_l(2\mu+1) & \text{if } \mu < 0 \\ 0 & \text{if } \mu \geq 0 \end{cases}\ , \tag{7}$$

This expansion allows treating the flux separately in each half-space $0 \leq \mu \leq 1$ and $-1 \leq \mu < 0$.

To obtain the classical Double P_N equations, it is assumed that

$$\psi_{l,g}(x,\mu) = \sum_{l=0}^{\infty}(2l+1)\left(\phi_{l,g}^+(x)P_l^+(\mu) + \phi_{l,g}^-(x)P_{l,g}^-(\mu)\right).$$

Again, by using the orthogonality relationships in equation (4) and integrating μ between 0 and 1, we obtain,

$$\sum_{l=0}^{\infty}\frac{d}{dx}\phi_{l,g}(x)\frac{1}{2}\left((l+1)C_{k,l+1}^+ + lC_{k,l-1}^+\right) + \Sigma_{tg}(x)\sum_{l=0}^{\infty}\phi_{l,g}(x)\left(\frac{2l+1}{2}\right)C_{k,l}^+$$
$$= \sum_{g'=1}^{G}\sum_{l=0}^{\infty}\left(\frac{2l+1}{2}\right)\Sigma_{sl}^{g\to g'}(x)\,\phi_{l,g'}(x)C_{k,l}^+ + \frac{\chi_g}{2k_{\text{eff}}}\sum_{g'=1}^{G}\nu_g\Sigma_{fg'}(x)\phi_{0,g'}(x)C_{k,0}^+ \tag{8}$$

where

$$C_{k,l}^+ = \int_0^{+1} P_k^+(\mu)P_l(\mu)\,d\mu\ . \tag{9}$$

Now, we integrate μ between -1 and 0 and obtain the Double P_N equations.

$$\sum_{l=0}^{\infty} \frac{d}{dx}\phi_{l,g}(x)\frac{1}{2}\left((l+1)C^{-}_{k,l+1} + lC^{-}_{k,l-1}\right) + \Sigma_{tg}(x)\sum_{l=0}^{\infty}\phi_{l,g}(x)\left(\frac{2l+1}{2}\right)C^{-}_{k,l}$$
$$= \sum_{g'=1}^{G}\sum_{l=0}^{\infty}\left(\frac{2l+1}{2}\right)\Sigma_{sl}^{g\to g'}(x)\,\phi_{l,g'}(x)C^{+}_{k,l} + \frac{\chi_g}{2k_{\text{eff}}}\sum_{g'=1}^{G}\nu_g\Sigma_{fg'}(x)\phi_{0,g'}(x)C^{-}_{k,0} \quad (10)$$

where

$$C^{-}_{k,l} = \int_{-1}^{0} P^{-}_{k}(\mu)P_l(\mu)\,d\mu\ . \quad (11)$$

We can truncate neutron flux expansion at $l = 3$ to obtain the Double P_1 as:

$$\psi_g(x,\mu) = \sum_{l=0}^{3}\phi_{l,g}(x)\left(\frac{2l+1}{2}\right)P_l(x)\ , \quad (12)$$

and we obtain:

$$\frac{1}{4}\frac{d}{dx}\phi^0 + \frac{1}{2}\frac{d}{dx}\phi^1 + \frac{5}{16}\frac{d}{dx}\phi^2 + \frac{1}{2}\boldsymbol{\Sigma^0}\phi^0 + \frac{3}{4}\boldsymbol{\Sigma^1}\phi^1 - \frac{7}{16}\boldsymbol{\Sigma^3}\phi^3 = \frac{1}{2k_{\text{eff}}}\mathbf{F}\phi^0 \quad (13)$$
$$-\frac{1}{4}\frac{d}{dx}\phi^0 + \frac{1}{2}\frac{d}{dx}\phi^1 - \frac{5}{16}\frac{d}{dx}\phi^2 + \frac{1}{2}\boldsymbol{\Sigma^0}\phi^0 - \frac{3}{4}\boldsymbol{\Sigma^1}\phi^1 + \frac{7}{16}\boldsymbol{\Sigma^3} = \frac{1}{2k_{\text{eff}}}\mathbf{F}\phi^0 \quad (14)$$
$$\frac{1}{12}\frac{d}{dx}\phi^0 + \frac{1}{4}\frac{d}{dx}\phi^1 + \frac{17}{48}\frac{d}{dx}\phi^2 + \frac{7}{24}\frac{d}{dx}\phi^3 + \frac{1}{4}\boldsymbol{\Sigma^1}\phi^1 + \frac{5}{8}\boldsymbol{\Sigma^2}\phi^2 + \frac{7}{16}\boldsymbol{\Sigma^3}\phi^3 = 0 \quad (15)$$
$$\frac{1}{12}\frac{d}{dx}\phi^0 - \frac{1}{4}\frac{d}{dx}\phi^1 + \frac{17}{48}\frac{d}{dx}\phi^2 - \frac{7}{24}\frac{d}{dx}\phi^3 + \frac{1}{4}\boldsymbol{\Sigma^1}\phi^1 - \frac{5}{8}\boldsymbol{\Sigma^2}\phi^2 + \frac{7}{16}\boldsymbol{\Sigma^3}\phi^3 = 0 \quad (16)$$

The odd moments can be isolated in terms of even moments, and substituted back

$$-\frac{1}{3}\frac{d}{dx}\left(\boldsymbol{\Sigma^1}\right)^{-1}\left(\frac{d\phi^0}{dx} + 2\frac{d\phi^2}{dx}\right) + \boldsymbol{\Sigma^0}\phi^0 = \frac{1}{k_{\text{eff}}}\mathbf{F}\phi^0 \quad (17)$$
$$-\frac{2}{15}\frac{d}{dx}\left(\boldsymbol{\Sigma^1}\right)^{-1}\left(\frac{d\phi^0}{dx} + 2\frac{d\phi^2}{dx}\right) - \frac{1}{5}\frac{d}{dx}\left(\boldsymbol{\Sigma^3}\right)^{-1}\left(\frac{d\phi^2}{dx}\right) + \boldsymbol{\Sigma^2}\phi^2 = 0 \quad (18)$$

This problem can be transformed into a system of diffusion-like equations by introducing the following linear change of variables,

$$U^1 = \phi^0 + 2\,\phi^2, \qquad U^2 = 3\phi^2, \quad (19)$$

The obtained system can be rewritten as

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathbb{D}\frac{\mathrm{d}}{\mathrm{d}x}U\right) + \mathbb{A}U = \frac{1}{\lambda}\mathbb{F}U\,, \quad (20)$$

where

$$\mathbb{D} = \begin{pmatrix} \frac{1}{3}\left(\boldsymbol{\Sigma^1}\right)^{-1} & 0 \\ 0 & \frac{1}{7}\left(\boldsymbol{\Sigma^3}\right)^{-1} \end{pmatrix}, \quad \mathbb{A}_{ij} = \sum_{m=1}^{2}\mathbf{c}_{ij}^{(\mathbf{m})}\boldsymbol{\Sigma^m}, \quad \mathbb{F}_{ij} = \mathbf{c}_{ij}^{(\mathbf{1})}\mathbf{F}, \quad (21)$$

$$\mathbf{c^{(1)}} = \begin{pmatrix} 1 & -\frac{2}{3} \\ -\frac{6}{7} & \frac{4}{7} \end{pmatrix}, \quad \mathbf{c^{(2)}} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{5}{7} \end{pmatrix}. \tag{22}$$

To maintain the structure of diffusion-like equations in multidimensional domains, we substitute the derivative with respect to x are substituted by a nabla operator as it was done with the Simplified PN equations Then, we get a system of diffusion-like equations that can use numerical solvers optimized for the neutron diffusion equation.

If we truncate the flux expansion at 6th element, we get the DP_2 equations.

$$\psi_g(x,\mu) = \sum_{l=0}^{5} \phi_{l,g}(x) \left(\frac{2l+1}{2}\right) P_l(x) \ , \tag{23}$$

Isolating the odd moments and making some simplifications, we obtain

$$-\frac{1}{3}\frac{d}{dx}\left(\boldsymbol{\Sigma^1}\right)^{-1}\left(\frac{d\phi^0}{dx} + 2\frac{d\phi^2}{dx}\right) + \Sigma^0\phi^0 = \frac{1}{k_{\text{eff}}}\mathbf{F}\phi^0 \tag{24}$$

$$-\frac{2}{15}\frac{d}{dx}\left(\boldsymbol{\Sigma^1}\right)^{-1}\left(\frac{d\phi^0}{dx} + 2\frac{d\phi^2}{dx}\right) - \frac{3}{35}\frac{d}{dx}\left(\boldsymbol{\Sigma^3}\right)^{-1}\left(3\frac{d\phi^2}{dx} + 4\frac{d\phi^4}{dx}\right) + \Sigma^2\phi^2 = 0 \tag{25}$$

$$-\frac{4}{63}\frac{d}{dx}\left(\boldsymbol{\Sigma^3}\right)^{-1}\left(3\frac{d\phi^0}{dx} + 4\frac{d\phi^2}{dx}\right) - \frac{1}{36}\frac{d}{dx}\left(\boldsymbol{\Sigma^5}\right)^{-1}\left(5\frac{d\phi^4}{dx}\right) + \Sigma^2\phi^2 = 0 \tag{26}$$

For the SDP_2 equations, the change of variables is

$$U^1 = \phi^0 + 2\phi^2 \,, \quad U^2 = 3\phi^2 + 4\phi^4 \,, \quad U^3 = 5\phi^4 \,. \tag{27}$$

To obtain a system of diffusion like equations:

$$-\vec{\nabla}\left(\mathbb{D}\vec{\nabla}U\right) + \mathbb{A}U = \frac{1}{\lambda}\mathbb{F}U \,, \tag{28}$$

The elements of the system are given by

$$\begin{gathered} \mathbb{A}_{ij} = \sum_{m=1}^{3} \mathbf{c}_{ij}^{(\mathbf{m})} \boldsymbol{\Sigma^m}, \quad \mathbb{F}_{ij} = \boldsymbol{c}_{ij}^{(\mathbf{1})} \boldsymbol{F} \,, \\ \mathbb{D} = \begin{pmatrix} \frac{1}{3}\left(\boldsymbol{\Sigma^1}\right)^{-1} & 0 & 0 \\ 0 & \frac{1}{7}\left(\boldsymbol{\Sigma^3}\right)^{-1} & 0 \\ 0 & 0 & \frac{1}{11}\left(\boldsymbol{\Sigma^5}\right)^{-1} \end{pmatrix} \end{gathered} \tag{29}$$

$$\boldsymbol{c^{(1)}} = \begin{pmatrix} 1 & -\frac{2}{3} & \frac{8}{15} \\ -\frac{2}{3} & \frac{4}{9} & -\frac{16}{45} \\ -\frac{32}{33} & -\frac{64}{99} & \frac{256}{495} \end{pmatrix}, \quad \boldsymbol{c^{(2)}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{5}{9} & -\frac{4}{9} \\ 0 & -\frac{80}{99} & \frac{64}{99} \end{pmatrix}, \quad \boldsymbol{c^{(3)}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{36}{55} \end{pmatrix}. \tag{30}$$

4 Numerical Results

We consider a two-dimensional one energy group nuclear system. The geometry of this system is provided in figure 1 and it is extracted form [1]. This problem possesses a strong spatial heterogeneity that makes that the diffusion approximation gives inaccurate results. This problem has been solved using the SDP_N and SP_N approximations using the same finite element method discretization.

Numerical results for this problem are given in Table 1. Figure 2 show the scalar neutron flux along the line $y = 4.5$ cm. We can see the SDP3. In this figure, we also compare the results obtained in [2].

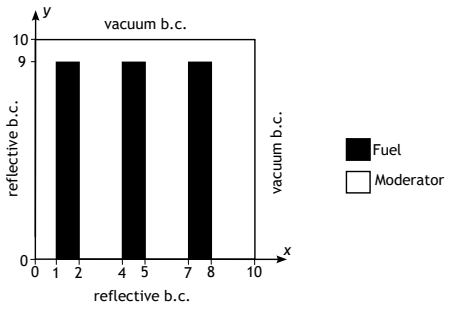
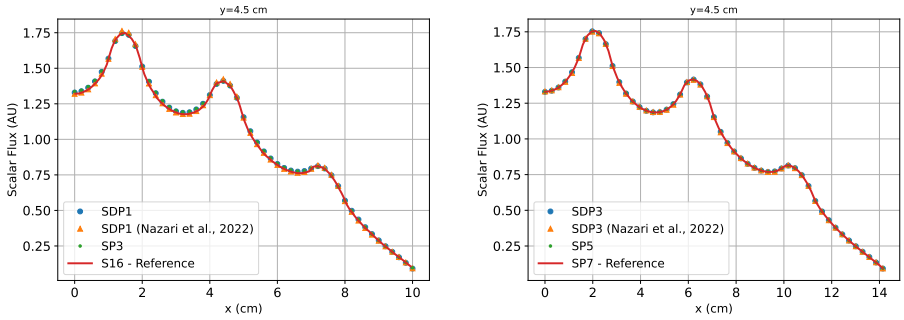


Fig. 1: Geometry of the two-dimensional one energy group problem [1].

Table 1: Numerical results for one-group problem.

Approximation	$k_{\mathbf{eff}}$	$\Delta k_{\mathbf{eff}}$ (pcm)	**RMS** (%)
SP_1	0.776 80	2674	3.80
SP_3	0.799 04	450	0.63
SP_5	0.802 80	74	0.13
SP_7	0.803 54	0	0.00
SDP_1	0.801 61	193	0.51
SDP_2	0.803 73	19	0.04
SDP_3	0.804 02	48	0.10
S_{16}	0.806 13		



(a) SPD1 and SP3 comparison (b) SPD2 and SP5 comparison

Fig. 2: Scalar flux along $y = 4.5$ cm for one-group eigenvalue problem

5 Conclusion

In this work, the Simplified Double P_N (SDP_N) approximation to the neutron transport equation is derived. These equations are leads to a system of diffusion-like equations. In the problem studied where strong spatial heterogeneities are present, SDP_N equations obtain more accurate results than the diffusion approximation or the Simplified P_N equation of the same order.

Future works will be devoted to compare the SDP_N equation in more realistic nuclear reactor configurations. This work is part of the open source software `FEMFFUSION` [4] developed at Universitat Politècnica de València. This software can be freely downloaded at (https://www.femffusion.webs.upv.es/).

References

1. Patrick S. Brantley and Edward W Larsen. The simplified P3 approximation. Nuclear Science and Engineering, 134(1):1–21, 2000.
2. M. Nazari, A Zolfaghari, and M Abbasi. The simplified-DPN approximation of the neutron transport equation. arXiv preprint arXiv:2212.02140, 2022.
3. Weston M. Stacey. Nuclear Reactor Physics. Wiley, Weinheim, Germany, February 2007.
4. A. Vidal-Ferràndiz, A. Carreño, D. Ginestar, and G. Verdú. FEMFFUSION: A finite element method code for the neutron diffusion equation. https://www.femffusion.webs.upv.es, 2023. Accessed: 2023-07-30.

Part II

Uncertainty Quantification and Modelling

Full probabilistic analysis of a stochastic embedded beam

J.-C. Cortés, E. López-Navarro*, J.-V. Romero, and M.-D. Roselló

Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València,
Camino de Vera, s/n, 46022, València, Spain,
*ellona1@doctor.upv.es,
WWW home page: http://munqu.webs.upv.es/

Abstract. In this contribution, we deal with the probabilistic analysis of the deflection of an embedded beam based on the Euler-Bernoulli's theory. We consider that all parameters of our model (the moment of inertia and the Young's modulus) are independent random variables. Furthermore, the load acting vertically over the beam is assumed to be described by the Brownian Bridge process. The aim of this work is to obtain the first probability density function using the Random Variable Transformation method. The theoretical findings will be illustrated with numerical simulations.

Keywords: stochastic embedded beam, Brownian Bridge process, deflection, Random Variable Transformation method

1 Introduction

The Euler-Bernoulli's theory is a key in structural engineering used to predict the behaviour of beams under loading, enabling engineers to understand how external forces influence the beam's deformation along its length. However, many factors, such as variability in materials and environmental conditions, can affect the accuracy of predictions based on this theory. This contribution will address the extension of the Euler-Bernoulli's theory to beams by incorporating randomness to the model. In this way, the analysis becomes more realistic to deal with real-world scenarios.

We will focus on the calculation of the probability density function (pdf) of the beam deflection using the Random Variable Transformation method (RVT), (see [1, page 25]). The computation of the pdf is highly relevant as it allows us to determining all the one-dimensional moments and the probability that the solution lies on a certain interval of specific interest. In the setting of structural mechanics, it permits calcuating the probability that the deflection of a beam lies within a certain safety interval or analyze failure probabilities.

The static deflection of a beam can be mathematically described by the following fourth-order differential equation [2]

$$\frac{\mathrm{d}^4 Y(x)}{\mathrm{d}x^4} = \frac{1}{EI} P(x), \quad 0 < x < l, \tag{1}$$

where $Y(x)$ represents the deflection curve of the beam, EI is the flexural rigidity being E the Young's modulus of elasticity and I the moment of inertia. We will consider that these two parameters are independent random variables. Here, l is the length of the beam and $P(x)$ is the force acting vertically on the beam at the spatial point x, which is described by $P(x) = P_0 + B(x)$, where P_0 is a deterministic parameter and $B(x)$ is the Brownian Bridge process defined by means of the Wiener process [3], $W(x)$, as

$$B(x) = W(x) - \frac{x}{l}W(x), \quad 0 \leq x \leq l. \tag{2}$$

In this work, specifically, we are going to study the deflection of a beam embedded at both ends. This can be written via the following boundary conditions

$$\begin{aligned} &Y(0) = 0, Y(l) = 0, \quad \text{(null deflection in the embeds)}, \\ &Y'(0) = 0, Y'(l) = 0, \quad \text{(null slope in the embeds)}. \end{aligned} \tag{3}$$

In order to perform the stochastic analysis, we will take advantage of the Karhunen-Loève expansion of the Brownian Bridge process

$$B(x) = \sum_{j=1}^{\infty} \frac{\sqrt{2l}}{j\pi} \sin\left(\frac{j\pi}{l}x\right) \xi_j, \tag{4}$$

where ξ_j are independent and indentically distributed Gaussian random variables, $\xi_j \sim \mathrm{N}(0,1)$.

2 Computing the first probability density function

In order to obtain the pdf, first we need to consider the approximation of $B(x)$ obtained by truncating at N its Karhunen-Loève expansion defined in (4). So, the model is approximated via the following differential equation

$$\begin{cases} \dfrac{\mathrm{d}^4 Y(x)}{\mathrm{d}x^4} = \dfrac{1}{EI}\left(P_0 + \sum_{j=1}^{N} \frac{\sqrt{2l}}{j\pi} \sin\left(\frac{j\pi}{l}x\right) \xi_j\right), \quad 0 < x < l, \\ Y(0) = 0, Y(l) = 0, Y'(0) = 0, Y'(l) = 0, \end{cases} \tag{5}$$

Second, we need to compute the stochastic solution of model (5). Integrating four times the equation and using the boundary conditions, we obtain

$$\begin{aligned} Y(x) = &\frac{1}{\pi^5 EI}\left(\frac{\pi^5}{24}P_0(l-x)^2x^2 + l^{3/2}\sqrt{2}\left(\pi(l-x)x\left((-l+x)\sum_{j=1}^{N}\frac{1}{j^4}\xi_j\right.\right.\right. \\ &\left.\left.\left. + x\sum_{j=1}^{N}\frac{1}{j^4}\cos(j\pi)\xi_j\right) + l^3\sum_{j=1}^{N}\frac{1}{j^5}\sin\left(\frac{j\pi}{l}x\right)\xi_j\right)\right). \end{aligned} \tag{6}$$

Now, we fix $0 < x < l$, and we apply the RVT method. In short, the RVT method permits obtaining the pdf of a random vector $\mathbf{V}$ that results from mapping

another random vector $\mathbf{U}$ whose pdf is known. So, taking $\mathbf{U} = (E, I, \xi_1, \ldots, \xi_N)$ to obtain the pdf of $\mathbf{V} = (V_1, V_2, \ldots, V_{N+2})$, defined by the mapping $\mathbf{r} : \mathbb{R}^{N+2} \to \mathbb{R}^{N+2}$, whose components are defined by

$$\begin{array}{rcll} v_1 & = & r_1(e,i,\xi_1,\ldots,\xi_N) & = Z(x;e,i,\xi_1,\ldots,\xi_N), \\ v_2 & = & r_2(e,i,\xi_1,\ldots,\xi_N) & = i, \\ v_3 & = & r_3(e,i,\xi_1,\ldots,\xi_N) & = \xi_1, \\ \vdots & \vdots & \vdots & \vdots\ \vdots \\ v_{N+2} & = & r_{N+2}(e,i,\xi_1,\ldots,\xi_N) & = \xi_N. \end{array}$$

where $Z(x;e,i,\xi_1,\ldots,\xi_N)$ is given by

$$\begin{aligned} Z(x;e,i,\xi_1,\ldots,\xi_N) = & \frac{1}{\pi^5 ei}\left(\frac{\pi^5}{24}P_0(l-x)^2x^2 + l^{3/2}\sqrt{2}\left(\pi(l-x)x\left((-l+x)\sum_{j=1}^{N}\frac{1}{j^4}\xi_j\right.\right.\right. \\ & \left.\left.\left. + x\sum_{j=1}^{N}\frac{1}{j^4}\cos(j\pi)\xi_j\right) + l^3\sum_{j=1}^{N}\frac{1}{j^5}\sin\left(\frac{j\pi}{l}x\right)\xi_j\right)\right). \end{aligned}$$

Notice that $Z(x;e,i,\xi_1,\ldots,\xi_N)$, i.e., the first component of $\mathbf{V}$, is the solution given by (6).

The inverse transformation or $\mathbf{r}$ is given by $\mathbf{s} : \mathbb{R}^{N+2} \to \mathbb{R}^{N+2}$, whose components are defined by

$$\begin{array}{rcll} e & = & s_1(v_1,v_2,v_3,\ldots,v_{N+2}) & = \widehat{Z}(x;v_1,v_2,v_3,\ldots,v_{N+2}), \\ i & = & s_2(v_1,v_2,v_3,\ldots,v_{N+2}) & = v_2, \\ \xi_1 & = & s_3(v_1,v_2,v_3,\ldots,v_{N+2}) & = v_3, \\ \vdots & \vdots & \vdots & \vdots\ \vdots \\ \xi_{N+2} & = & s_{N+2}(v_1,v_2,v_3,\ldots,v_{N+2}) & = v_{N+2}, \end{array}$$

where $\widehat{Z}(x;v_1,v_2,v_3,\ldots,v_{N+2})$ is given by

$$\begin{aligned} \widehat{Z}(x;v_1,v_2,v_3,\ldots,v_{N+2}) = & \frac{1}{\pi^5 v_1 v_2}\left(\frac{\pi^5}{24}P(l-x)^2x^2\right. \\ & + l^{3/2}\sqrt{2}\left(\pi(l-x)x\left((-l+x)\sum_{j=1}^{N}\frac{1}{j^4}v_{j+2}\right.\right. \\ & \left.\left. + x\sum_{j=1}^{N}\frac{1}{j^4}\cos(j\pi)v_{j+2}\right) + l^3\sum_{j=1}^{N}\frac{1}{j^5}\sin\left(\frac{j\pi}{l}x\right)v_{j+2}\right)\Bigg). \end{aligned}$$

The absolute value of the Jacobian is determined by

$$|\mathcal{J}| = \left|\frac{\partial s_1(v_1,v_2,v_3,\ldots,v_{N+2})}{\partial v_1}\right| = \left|-\frac{1}{v_1}\widehat{Z}(x;v_1,v_2,\ldots,v_{N+2})\right|,$$

which is different from zero with probability 1 (w.p. 1), since the parameters are continuous random variables.

Once we have all the ingredients to apply the RVT method, we can compute the pdf of the random vector $\mathbf{V}$

$$f_{V_1,V_2,\ldots,V_{N+2}}(v_1,v_2,v_3,\ldots,v_{N+2}) = f_{E,I,\xi_1,\ldots\xi_N}\left(\widehat{Z}(x;v_1,v_2,\ldots,v_{N+2}),v_2,v_3,\ldots,v_{N+2}\right) \cdot \left|-\frac{1}{v_1}\widehat{Z}(x;v_1,v_2,\ldots,v_{N+2})\right|.$$

Since E, I, and ξ_1, …, ξ_N are independent random variables, so

$$\begin{aligned} f_{V_1,V_2,\ldots,V_{N+2}}(v_1,v_2,v_3,\ldots,v_{N+2}) &= f_E\left(\widehat{Z}(x;v_1,v_2,\ldots,v_{N+2})\right) f_I(v_2) \\ &\cdot f_{\xi_1}(v_3)\cdots f_{\xi_N}(v_{N+2})\left|-\frac{1}{v_1}\widehat{Z}(x;v_1,v_2,\ldots,v_{N+2})\right|. \end{aligned} \tag{7}$$

As we have noticed before, the model solution (6) corresponds to the first component V_1, then, we have to marginalize respect to $V_2 = I$, $V_3 = \xi_1$, …, $V_{N+2} = \xi_N$, in order to obtain the 1-pdf of the stochastic solution

$$\begin{aligned} f^N_{Y(x)}(y) = \int_{\mathbb{R}^{N+1}} & f_E\left(\frac{1}{\pi^5 yi}\left(\frac{\pi^5}{24}P(l-x)^2x^2\right.\right. \\ & +l^{3/2}\sqrt{2}\left(\pi(l-x)x\left((-l+x)\sum_{j=1}^{N}\frac{1}{j^4}\xi_j + x\sum_{j=1}^{N}\frac{1}{j^4}\cos(j\pi)\xi_j\right)\right. \\ & \left.\left.\left.+l^3\sum_{j=1}^{N}\frac{1}{j^5}\sin\left(\frac{j\pi}{l}x\right)\xi_j\right)\right)\right) f_I(i) f_{\xi_1}(\xi_1)\ldots f_{\xi_N}(\xi_N)\left|-\frac{1}{\pi^5 y^2 i}\left(\frac{\pi^5}{24}P(l-x)^2x^2\right.\right. \\ & +l^{3/2}\sqrt{2}\left(\pi(l-x)x\left((-l+x)\sum_{j=1}^{N}\frac{1}{j^4}\xi_j + x\sum_{j=1}^{N}\frac{1}{j^4}\cos(j\pi)\xi_j\right)\right. \\ & \left.\left.\left.+l^3\sum_{j=1}^{N}\frac{1}{j^5}\sin\left(\frac{j\pi}{l}x\right)\xi_j\right)\right)\right| \mathrm{d}\, i\, \mathrm{d}\,\xi_1\ldots\mathrm{d}\,\xi_N. \end{aligned} \tag{8}$$

This latter expression can be rewritten in terms of the expectation operator, that is computationally more efficient to obtain the pdf, as it allows us to use Monte Carlo simulations.

$$\begin{aligned} f^N_{Y(x)}(y) = \mathbb{E}_{I,\xi_1,\ldots,\xi_N} & \left[f_E\left(\frac{1}{\pi^5 yi}\left(\frac{\pi^5}{24}P(l-x)^2x^2\right.\right.\right. \\ & +l^{3/2}\sqrt{2}\left(\pi(l-x)x\left((-l+x)\sum_{j=1}^{N}\frac{1}{j^4}\xi_j + x\sum_{j=1}^{N}\frac{1}{j^4}\cos(j\pi)\xi_j\right)\right. \\ & \left.\left.\left.+l^3\sum_{j=1}^{N}\frac{1}{j^5}\sin\left(\frac{j\pi}{l}x\right)\xi_j\right)\right)\right)\left|-\frac{1}{\pi^5 y^2 i}\left(\frac{\pi^5}{24}P(l-x)^2x^2\right.\right. \\ & +l^{3/2}\sqrt{2}\left(\pi(l-x)x\left((-l+x)\sum_{j=1}^{N}\frac{1}{j^4}\xi_j + x\sum_{j=1}^{N}\frac{1}{j^4}\cos(j\pi)\xi_j\right)\right. \\ & \left.\left.\left.+l^3\sum_{j=1}^{N}\frac{1}{j^5}\sin\left(\frac{j\pi}{l}x\right)\xi_j\right)\right)\right| \Bigg]. \end{aligned} \tag{9}$$

3 Numerical example

In this section, we show a numerical example using the previous theoretical findings. Let us consider the following deterministic parameters: the length of the beam, $l = 10\text{m}$, and the nominal value of the load, $P_0 = 700\text{N}$. And for the random parameters, we assume that the Young's modulus, E, and the moment of inertia, I, have both Gaussian distribution. More specifically, $E \sim \text{N}(210 \cdot 10^9, 105 \cdot 10^8)\text{N/m}^2$ and $I \sim \text{N}(33740 \cdot 10^{-8}, 6748 \cdot 10^{-6})\text{m}^4$. As we have seen before, the random variables ξ_j, obtained from the Karhunen-Loève expansion, follow a Gaussian dristribution, $\xi_j \sim \text{N}(0, 1)$.

In Figure 1, we show the graphical representation of the 1-pdf of the deflection of the beam, $f^N_{Y(x)}(y)$, at different spatial points, $x \in \{1, \ldots, 9\}$, with the truncation order of the Karhunen-Loève expansion $N = 5$. We can observe, that the variance increases in the middle of the beam as expected.

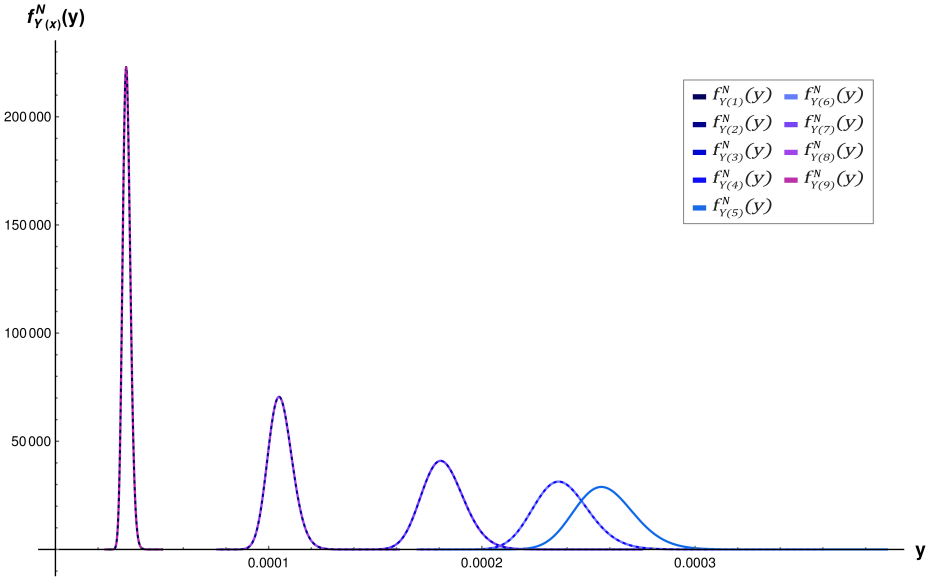


Fig. 1: 1-pdf of the deflection of the beam, $f^N_{Y(x)}(y)$, at different spatial points $x \in \{1, \ldots, 9\}$, with the truncation order $N = 5$.

In Figure 2, we show the mean and 95% confidence intervals of the deflection, $Y(x)$. Again, we can observe that the variability increases in the middle of the beam.

In Table 1, we show the mean of the deflection, $Y(x)$ at the spatial point $x = 5$ for different values of the truncation order of the Karhunen-Loève expansion, $N \in \{1, 5, 10, 50\}$. We can observe, that the approximations are very similar with $N = 1$, matching up to 8 decimal numbers.

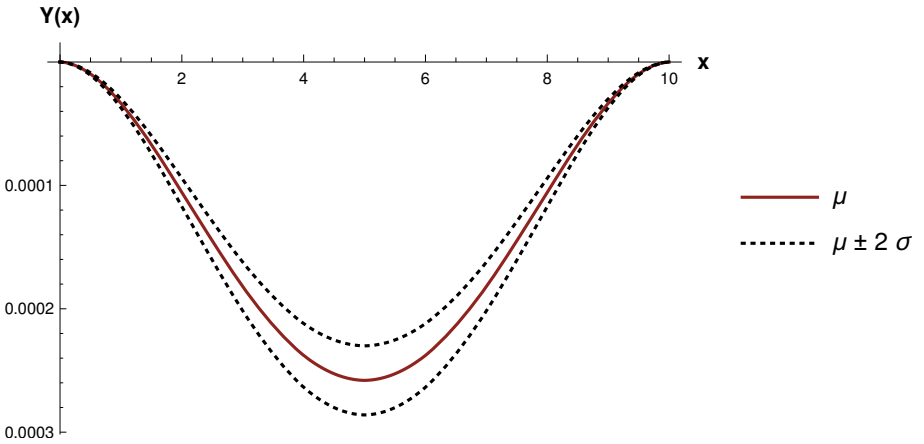


Fig. 2: Mean and confidence intervals of the deflection, $Y(x)$.

N	1	5	10	50
Mean	0.00025803277	0.00025803452	0.00025803360	0.00025803338

Table 1: Mean of the deflection, $Y(x)$, at the spatial point $x = 5$ for different values of the truncation order N.

4 Conclusions

In this contribution, we have applied the Random Variable Transformation method to obtain the first probability density function of the deflection of an embedded beam at both ends. We consider that the moment of inertia and the Young's modulus are independent random variables. And we use the Brownian Bridge process to describe the load acting vertically over the beam. To represent this stochastic process, we have taken advantage of the Karhunen-Loève expansion. In this way, we have obtained a full probabilistic description of the deflection of an embedded beam.

In future works, we will study other important characteristics of the beam, such as the bending moment and the shear force, and we will carry out an in-depth study of convergence of the approximation obtained by truncating the Karhunen-Loève expansion.

Acknowledgements

This work has been supported by the grant PID2020-115270GB-I00 funded by MCIN/AEI/10.13039/501100011033 and the grant AICO/2021/302 (Generalitat Valenciana).

References

1. Soong, T., Random Differential Equations in Science and Engineering. New York, Academic Press, 1973.
2. Öchsner, A., Classical Beam Theories of Structural Mechanics. Springer, 2021.
3. Lord, G., Powell, C., Shardlow, T., An Introduction to Computational Stochastic PDEs, in: Cambridge Texts in Applied Mathematics, Cambridge University Press, New York, 2014.
4. Cortés, JC., López-Navarro, E., Romero, JV., Roselló, MD. Probabilistic analysis of a cantilever beam subjected to random loads via probability density functions. *Comp. Appl. Math.* 42, 42, 2023. https://doi.org/10.1007/s40314-023-02194-0
5. Malkiel, N., Rabinovitch, O., Elishakoff, I. Exact solutions for stochastic Bernoulli–Euler beams under deterministic loading. *Acta Mechanica* 232(6), 2201–2224, 2021.

A comprehensive study of the random hyperlogistic differential equation combining theoretical insights and simulation analysis

Juan Carlos Cortés[1], Ana Navarro-Quiles[2], and Sorina Madalina Sferle[1]

[1] Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain,
jccortes@imm.upv.es,
[2] Department of Statistics and Operational Research, Universitat de València, Dr. Moliner 50, 46100, Burjassot, Spain

Abstract. We deal with the probabilistic analysis of a full randomization of the hyperlogistic differential equation. By assuming that all parameters of this equation as well as the initial condition are absolutely continuous random variables with a joint density, we obtain, under very general hypotheses, a semi-explicit expression of the first finite distribution of the solution, which is a stochastic process. Furthermore, we determine the probability density function of the inflection point, which plays a key role in the distinctive mathematical properties the hyperlogistic model has with respect to other classical growth models. The mathematical findings are illustrated by means of examples where different distributions are assigned to model parameters.

Keywords: uncertainty quantification, hyperlogistic differential equation, stochastic analysis, random variable transformation technique, growth processes

1 Introduction and Motivation

Growth processes are ubiquitous in nature and occur at a variety of scales, from individual organisms to populations, economies and social systems. The study of theses processes is essential for understanding the patterns, mechanisms and dynamics of how systems develop and change over time, and is vital to fields such as biology [1], medicine [2] and ecology [3], among others. Logistic-type differential equations play a key role in the mathematical modeling of the above processes in many settings, such as population dynamics, spread and control of infectious diseases, chemical reactions, nucleation theory, financial systems, urban planning and transportation systems, etc. The formulation of such equations is based on the choice of the function that defines the right-hand side of equation. The logistic equation, in particular, is a classic example of a differential equation used to model population growth and other similar phenomena; however, in the literature we can find other equations also widely used for the same purpose,

such as the Gompertz, Richards or Von-Bertalanffy equations (see [4], [5] and [6] for original documents, respectively).

The solution of logistic models is usually represented by a sigmoidal curve, which is divided into three phases, the lag phase, the exponential phase and the stationary phase. The horizontal asymptote around the upper boundary value is a distinctive property of this curve and represents the carrying capacity. The concept of an inflection point is fundamental in the study of curves, as it provides relevant information about the changing behavior of a function or the rate of change of a quantity represented by the curve. The exponential phase usually represents the period of fastest growth, so the highest growth rate occurs at the inflection point.

In the realm of population dynamics and growth processes, the logistic growth model, initially introduced by Verhulst in 1838, assumes a pivotal role. Unlike earlier models, which did not take into account critical environmental constraints to population growth, this mathematical framework offers a more realistic representation of how populations evolve over time. A key aspect of this model is that the population at the inflection point is exactly half of the carrying capacity. This causes an undesirable constraint on the shape of the curve, since it is symmetric with respect to that point. That is why several models have been developed as parameterized and extended versions, which provide a relaxation of the limitations of this model.

In the present work, we will introduce the hyperlogistic equation proposed by Turner et al. [7]. Their theory is based on three postulates, thus obtaining a generic growth function from which several of the known growth curves, such as the Verhulst, Gompertz, Von-Bertalanffy or Richards curve, can be obtained as special cases, but they also obtained several new forms, one of which is the one we have mentioned and which we will study in the present work.

The first postulate follows in the tradition of the pioneering work of Quetelet [8], Verhulst [9], Pearl and Reed [10], and Lotka [11] and establishes that the growth rate, say $\frac{\mathrm{d}\,x(t)}{\mathrm{d}\,t}$, is proportional to a monotonically increasing function, which indicates the generalized distance from the origin to the current population, $x(t)$, and to another monotonically decreasing function, which indicates the generalized distance from the current population to the final population, $(k - x(t))$. The second postulate restricts these functions to strictly positive power functions. Finally, the third one is based on some restrictions that the exponents must obey, thus this postulate limits the model to a mathematically manageable set and ensures the obtainment of the common special cases. In this way, the resulting initial value problem (IVP) has the following form

$$\begin{cases} \frac{\mathrm{d}\,x(t)}{\mathrm{d}\,t} = \frac{\beta}{k} x(t)^{1-p}(k - x(t))^{1+p}, \\ x(t_0) = x_0, \end{cases} \tag{1}$$

where $x_0 > 0$ is the initial population at initial time $t_0 \geq 0$, $\frac{\beta}{k}$ is the proportionality constant, being $\beta > 0$ the intrinsic growth constant and $k > 0$ the carrying capacity, and p the shape parameter, where $p \in (0, 1)$ so that its solution is a sigmoidal function. This parameter reveals the importance and the difference of

this equation, since its variation makes it possible to place the inflection point of the sigmoidal curve at any value between the minimum and the carrying capacity. Here the population at the inflection point, x_{PI}, is given by

$$x_{\mathrm{PI}} = \frac{k(1-p)}{2}. \tag{2}$$

The IVP (1) admits the following analytic solution

$$x(t) = k - \frac{k}{1 + \left(p\beta(t-t_0) + \left(\frac{k}{x_0} - 1\right)^{-p}\right)^{\frac{1}{p}}}, \tag{3}$$

which depends on the parameters β, k, p and x_0. In many studies, these are considered deterministic. However, it is important to note that growth is a highly regulated and complex process, involving a delicate balance of genetic, cellular, environmental factors, such as climate variability, resource availability, disease outbreaks, and demographic factors, such as birth, mortality or fertility rates. Therefore, they contain an intrinsic uncertainty that would otherwise not be taken into account. There are many types of strategies to cope with this type of approach, and one of them is to consider uncertainties through random variables with regular sampling behavior. As a result, they can help to account for randomness, natural variability and the influence of stochastic events on population dynamics. Thus, random parameters are able to better explain the complexity of the real world. Consequently, the hyperlogistic equation (1) becomes a random differential equation (RDE) with a random initial condition as follows

$$\begin{cases} \frac{\mathrm{d}\,x(t,\omega)}{\mathrm{d}\,t} = \frac{\beta(\omega)}{k(\omega)} x(t,\omega)^{1-p(\omega)} (k(\omega) - x(t,\omega))^{1+p(\omega)}, \\ x(t_0) = x_0(\omega), \end{cases} \tag{4}$$

where $\omega \in \Omega$ indicates the sample dependence for random variables, which will be omitted hereafter for simplicity. Its corresponding deterministic solution (3) is now a recast stochastic process as the following

$$x(t,\omega) = k(\omega) - \frac{k(\omega)}{1 + \left(p(\omega)\beta(\omega)(t-t_0) + \left(\frac{k(\omega)}{x_0(\omega)} - 1\right)^{-p(\omega)}\right)^{\frac{1}{p(\omega)}}}, \qquad \omega \in \Omega, \tag{5}$$

where $\beta(\omega)$, $k(\omega)$, $p(\omega)$ and $x_0(\omega)$ are absolutely continuous random variables defined in a common complete probability space (Ω, F_Ω, P). For the sake of generality in our subsequent development, we shall assume that the model parameters, $\beta(\omega)$, $k(\omega)$, $p(\omega)$ and $x_0(\omega)$ have a joint probability density function (PDF), say $f_0(\beta, k, p, x_0)$. The particularity of these RDEs is that much more information can be obtained from the model apart from its solution, specifically, by determining its first probability density function (1-PDF) one can calculate significant information for the model such as the mean, variance, confidence

intervals among other statistical characteristics. In this way we would have a complete probabilistic description of the stochastic solution for each time instant.

The layout of this document is as follows. In Section 2, we will determine the 1-PDF of the stochastic solution (5). In Section 3, we will illustrate the developed theoretical results by means of examples with synthetic data. Finally, conclusions are drawn in Section 4.

2 Determining PDFs

In this section we will see how to determine the 1-PDF of the stochastic solution (5) using the Random Variable Transformation (RVT) technique. The version of this technique that we will implement in this work is found in [12], and is stated as follows.

Theorem 1. *Let* $\mathbf{U} = (U_1, \ldots, U_n)$ *and* $\mathbf{W} = (W_1, \ldots, W_n)$ *be* n*-dimensional random vectors. Let* $\mathbf{r} : \mathbb{R}^n \to \mathbb{R}^n$ *be a one-to-one transformation of* $\mathbf{U}$ *into* $\mathbf{W}$*, i.e.,* $\mathbf{W} = \mathbf{r}(\mathbf{U})$*. Assume that* $\mathbf{r}$ *is continuous in* $\mathbf{U}$ *and has continuous partial derivatives with respect to* $\mathbf{U}$*. Then, if* $f_{\mathbf{U}}(\mathbf{u})$ *denotes the known joint PDF of vector* $\mathbf{U}$*, and* $\mathbf{s} = \mathbf{r}^{-1}$ *represents the inverse mapping of* $\mathbf{r}$*, the joint PDF of vector* $\mathbf{W}$ *is given by*

$$f_{\mathbf{W}}(\mathbf{w}) = f_{\mathbf{U}}\left(\mathbf{s}(\mathbf{w})\right) |J_n| \,,$$

where $|J_n|$ *is the Jacobian, which is defined by*

$$J_n = \det \begin{pmatrix} \dfrac{\partial s_1(\mathbf{w})}{\partial W_1} & \cdots & \dfrac{\partial s_n(\mathbf{w})}{\partial W_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial s_1(\mathbf{w})}{\partial W_n} & \cdots & \dfrac{\partial s_n(\mathbf{w})}{\partial W_n} \end{pmatrix}.$$

In addition, as shall be seen later, this technique will allow us to calculate the PDF of other random quantities of interest to our problem, such as the inflection point. It should be noted that this technique has also been applied to other problems with randomness (see [13–15]).

2.1 Determining the 1-PDF of the stochastic solution

Fixing $t > 0$, we apply RVT technique with the following identification

$$\mathbf{U} = (\beta, k, p, x_0), \qquad f_{\mathbf{U}}(\mathbf{u}) = f_0(\beta, k, p, x_0),$$

$$\mathbf{W} = (W_1, W_2, W_3, W_4) = \mathbf{r}(\beta, k, p, x_0),$$

where $\mathbf{r} : \mathbb{R}^4 \to \mathbb{R}^4$ is the bijective mapping, the components of which are defined by

$$\begin{aligned}
W_1 &= r_1(\beta, k, p, v_0) = \beta,\\
W_2 &= r_2(\beta, k, p, v_0) = k,\\
W_3 &= r_3(\beta, k, p, v_0) = p,\\
W_4 &= r_4(\beta, k, p, v_0) = x(t) = k - \frac{k}{1 + \left(p\beta(t - t_0) + \left(\frac{k}{x_0} - 1\right)^{-p}\right)^{\frac{1}{p}}},
\end{aligned}$$

and its inverse $\mathbf{s} : \mathbb{R}^4 \to \mathbb{R}^4$ is given by

$$\begin{aligned}
\beta &= s_1(W_1, W_2, W_3, W_4) = W_1,\\
k &= s_2(W_1, W_2, W_3, W_4) = W_2,\\
p &= s_3(W_1, W_2, W_3, W_4) = W_3,\\
x_0 &= s_4(W_1, W_2, W_3, W_4) = \frac{W_2}{1 + \left(\left(\frac{W_4}{W_2 - W_4}\right)^{W_3} - W_3 W_1 (t - t_0)\right)^{-\frac{1}{W_3}}}.
\end{aligned}$$

From this, we obtain the Jacobian

$$J_4 = \frac{W_2^2 \left(\frac{W_4}{W_2 - W_4}\right)^{W_3 - 1} \left(\left(\frac{W_4}{W_2 - W_4}\right)^{W_3} - W_1 W_3 (t - t_0)\right)^{\frac{1}{W_3} - 1}}{(W_2 - W_4)^2 \left(\left(\left(\frac{W_4}{W_2 - W_4}\right)^{W_3} - W_1 W_3 (t - t_0)\right)^{\frac{1}{W_3}} + 1\right)^2}. \tag{6}$$

Thus the joint PDF of the random vector $\mathbf{W}$ is

$$f_{\mathbf{W}}(\mathbf{w}) = f_0 \left(W_1, W_2, W_3, \frac{W_2}{1 + \left(\left(\frac{W_4}{W_2 - W_4}\right)^{W_3} - W_3 W_1 (t - t_0)\right)^{-\frac{1}{W_3}}} \right) |J_4|.$$

Taking into account that $W_4 = x(t)$, and considering an arbitrary $t > 0$ and marginalizing with respect to the other components of $\mathbf{W}$, the 1-PDF is

$$f_1(x, t) = \int_0^1 \int_0^\infty \int_0^\infty f_0 \left(\beta, k, p, \frac{k}{1 + \left(\left(\frac{x}{k - x}\right)^p - p\beta(t - t_0)\right)^{-\frac{1}{p}}} \right) |J_4| \, \mathrm{d}\beta \, \mathrm{d}k \, \mathrm{d}p, \tag{7}$$

where J_4 is given by (6).

2.2 Determining the PDF of the inflection point

Following the same procedure, we obtain the PDF of the inflection point. So, fixing $t > 0$, we apply Theorem 1 using the following identification

$$\mathbf{U} = (k, p), \qquad f_{\mathbf{U}}(\mathbf{u}) = f_0(k, p),$$
$$\mathbf{W} = (W_1, W_2) = \mathbf{r}(k, p),$$

where $\mathbf{r} : \mathbb{R}^2 \to \mathbb{R}^2$ is the bijective mapping, the components of which are defined by

$$W_1 = r_1(k, p) = k,$$
$$W_2 = r_2(k, p) = x_{\mathrm{PI}} = \frac{k(1-p)}{2},$$

and its inverse $\mathbf{s} : \mathbb{R}^2 \to \mathbb{R}^2$ is given by

$$k = s_1(W_1, W_2) = W_1,$$
$$p = s_2(W_1, W_2) = 1 - \frac{2W_2}{W_1},$$

which is well-defined. From this, we obtain the Jacobian

$$|J_2| = \left|\frac{-2}{W_1}\right| = \frac{2}{W_1}.$$

Thus the joint PDF of the random vector $\mathbf{W}$ is

$$f_{\mathbf{W}}(\mathbf{w}) = f_0\left(W_1, 1 - \frac{2W_2}{W_1}\right)\frac{2}{W_1}.$$

Taking into account that $W_2 = x_{\mathrm{PI}}$, and considering an arbitrary $t > 0$ and marginalizing with respect to the other component of $\mathbf{W}$, the PDF is

$$f(x_{\mathrm{PI}}) = \int_0^\infty f_0\left(k, 1 - \frac{2x_{\mathrm{PI}}}{k}\right)\frac{2}{k}\,\mathrm{d}\,k. \tag{8}$$

3 Numerical examples

In this section we will apply the theoretical results obtained previously by means of numerical examples using synthetic data. The aim is to obtain the 1-PDF of the stochastic solution for each time instant, and the PDF of the inflection point.

To apply the theoretical results, it is first necessary to establish a suitable joint PDF of the model parameters. In practice, with real data, this step is crucial, since we have to find appropriate distributions that best capture the uncertainty of the data. However, as in this paper we only want to illustrate the developed theory, we will assume some arbitrary distributions based on the positivity and boundedness of the random variables.

Let us assume that all parameters are independent, so f_0 is the product of the marginals, $f_0(\beta, k, p, x_0) = f_B(\beta)f_K(k)f_P(p)f_{X_0}(x_0)$.

3.1 Example 1

In the first case we assume uniform distributions for all random variables. To determine the parameter on which these distributions depend on, we set the mean and variance to some values as follows,

$$\begin{aligned}
\mathrm{E}(\beta) = 0.2, \quad \mathrm{Var}(\beta) = 0.001 \quad &\rightarrow \quad \beta \sim Unif(0.1452, 0.2547), \\
\mathrm{E}(k) = 5, \quad \mathrm{Var}(k) = 0.04 \quad &\rightarrow \quad k \sim Unif(4.6535, 5.3464), \\
\mathrm{E}(p) = 0.2, \quad \mathrm{Var}(p) = 0.0002 \quad &\rightarrow \quad p \sim Unif(0.1755, 0.2244), \\
\mathrm{E}(x_0) = 0.02, \quad \mathrm{Var}(x_0) = 0.0001 \quad &\rightarrow \quad x_0 \sim Unif(0.0026, 0.0373).
\end{aligned}$$

Applying expression (7), the resulting 1-PDF is shown on the left panel of Figure 1, where it is observed that the mass density moves over time towards the carrying capacity, whose PDF resembles a uniform one. In addition, the PDF of the inflection point is shown on the right panel of Figure 1, whose mean is around 2, thus indicating the population at the inflection point.

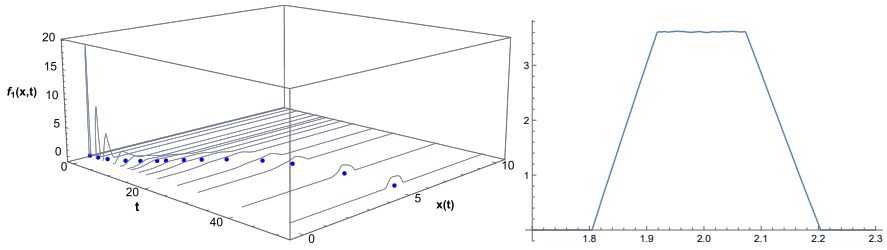


Fig. 1: Left: 1-PDF representation of the stochastic process (5) of the random differential equation (4) at different time instants. It has been calculating from expression (7). Right: PDF of the inflection point (2). It has been calculating from expression (8).

The behavior of the 1-PDF is best seen in the Figure 2. On the left panel, we observe that the mean has a sigmoidal shape, as expected, tending to the carrying capacity. On the right panel, we can say that in the lag phase there is not much variability, then in the exponential phase it grows until it reaches a maximum. This is because in this phase several external factors contribute to the growth. And, in the stationary phase, it goes down again and stabilizes.

3.2 Example 2

In the second case we assume the following distributions,

$$\mathrm{E}(\beta) = 0.6, \quad \mathrm{Var}(\beta) = 0.01 \quad \rightarrow \quad \beta \sim Beta(13.8, 9.2),$$

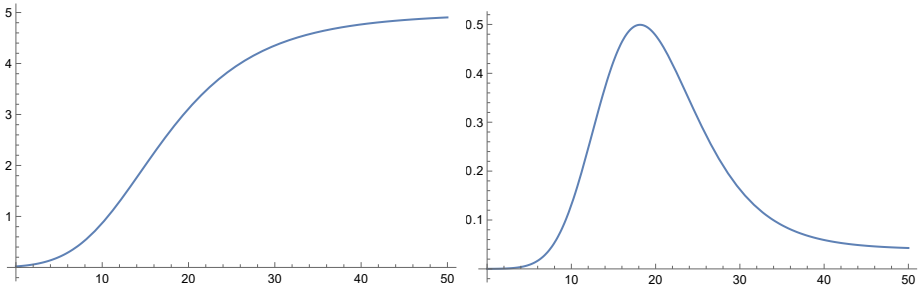


Fig. 2: Left: Expectation of the stochastic process (5) obtained using the 1-PDF (7). Right: Variance of the stochastic process (5) obtained using the 1-PDF (7).

$$\mathrm{E}(k) = 15, \quad \mathrm{Var}(k) = 0.5 \quad \rightarrow \quad k \sim LogNormal(2.7069, 0.0471),$$
$$\mathrm{E}(p) = 0.3, \quad \mathrm{Var}(p) = 0.002 \quad \rightarrow \quad p \sim Unif(0.2225, 0.3775),$$
$$\mathrm{E}(x_0) = 0.05, \quad \mathrm{Var}(x_0) = 0.0001 \quad \rightarrow \quad x_0 \sim Unif(0.0327, 0.0.067).$$

The interpretation of the results obtained is similar to that of the previous

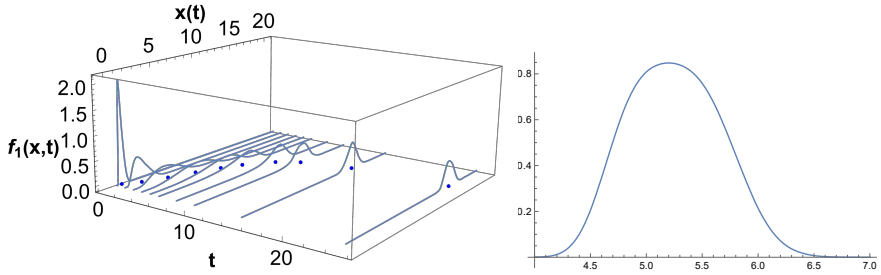


Fig. 3: Left: 1-PDF representation of the stochastic process (5) of the random differential equation (4) at different time instants. Right: PDF of the inflection point (2).

example, except that now the shape of the 1-PDF for each time appears to be Gaussian.

We point out that, depending on the values of β and p, the sigmoidal curve changes. In this case, the growth phase begins earlier and its slope is greater, reaching the carrying capacity earlier.

4 Conclusions

In this work we have presented a full stochastic analysis of the random hyperlogistic differential equation assuming that all its parameters and the initial con-

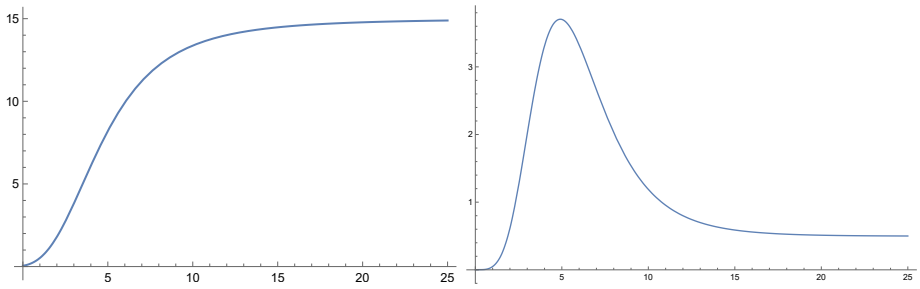


Fig. 4: Left: Expectation of the stochastic process (5) obtained using the 1-PDF (7). Right: Variance of the stochastic process (5) obtained using the 1-PDF (7).

dition are absolutely continuous random variables with arbitrary distributions. For this purpose, the first finite distribution of the solution has been obtained via the Random Variable Transformation technique under general assumptions on the random parameters. Finally, the theoretical findings have been illustrated by means of numerical examples.

Acknowledgements

This work has been supported by the grant PID2020-115270GB–I00 granted by MCIN/AEI/10.13039/501100011033 and the grant PRE2021-101090, granted by MCIN/AEI/10.13039/501100011033 and by FSE+.

References

1. Peleg, M., Corradini, M.G.: Microbial growth curves: what the models tell us and what they cannot. Crit. Rev. Food Sci. Nutr. 51, 917–945 (2011). doi:10.1080/10408398.2011.570463
2. Marusic, M., Bajzer, Z., Freyer, J.P., Vuk-Pavlovic, S.: Analysis of growth of multicellular tumour spheroids by mathematical models. Cell Prolif. 27, 73–94 (1994). doi:10.1111/j.1365-2184.1994.tb01407.x
3. Petter, G., Kreft, H., Ong, Y., Zotz, G., Cabral, J. S.: Modelling the long-term dynamics of tropical forests: From leaf traits to whole-tree growth patterns. Ecol. Modell. 460, 109735 (2021). doi:10.1016/j.ecolmodel.2021.109735
4. Gompertz, B.: On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philos. Trans. Royal Soc. 115, 513–585 (1825). doi:10.1098/rspl.1815.0271
5. Richards, F. J.: A flexible growth function for empirical use. J. Exp. Bot. 10, 290–301 (1959). doi:10.1093/jxb/10.2.290
6. Von-Bertalanffy, L.: Quantitative laws in metabolism and growth. Q. Rev. Biol. 32, 217–231 (1957). doi:10.1086/401873

7. Turner Jr, M.E., Bradley Jr, E.L., Kirk, K.A., Pruitt, K.M.: A theory of growth. Math. Biosci. 29, 367–373 (1976). doi:10.1016/0025-5564(76)90112-7
8. Quetelet, M.A.: A Treatise on Man and the Development of His Faculties. Burt Franklin, New York (1968)
9. Verhulst, P.F.: Notice sur la loi que la population suit dans son accroissement. Corr. Math. Phys. 10, 113–121 (1838)
10. Pearl, R., Reed, L.J.: On the rate of growth of the population of the United States since 1790 and its mathematical representation. Proc. Nat. Acad. Sci. 6, 275–288 (1920). doi:10.1073/pnas.6.6.275
11. Lotka, A.J.: Elements of Mathematical Biology. Dover, New York (1956)
12. Soong, T.T.: Random Differential Equations in Science and Engineering. Academic Press, New York and London (1973)
13. Dorini, F.A., de Castro Cunha, M.C., Dorini, L. B.: A note on the solution to the random Burgers–Riemann problem subject to independent and uniformly distributed initial conditions. Comput. Appl. Math. 42, 64 (2023). doi:10.1007/s40314-023-02207-y
14. Dorini, F.A., Bobko, N., Dorini, L.B. A note on the logistic equation subject to uncertainties in parameters. Comp. Appl. Math. 37, 1496–1506 (2018). doi:10.1007/s40314-016-0409-6
15. Casabán, M.C., Cortés, J.C., Navarro-Quiles, A., Romero, J.V., Roselló, M.D., Villanueva, R.J.: A comprehensive probabilistic solution of random SIS-type epidemiological models using the random variable transformation technique. Commun. Nonlinear Sci. Numer. Simul. 32, 199–210 (2016). doi:10.1016/j.cnsns.2015.08.009

Part III

Maths and Physics

Some Computational Tools in RPS

Màrius Josep Fullana i Alfonso[1], Neus Puchades Colmenero[2], and Josep Vicent Arnau i Córdoba[1]

[1] Institut Universitari de Matemàtica Multidisciplinària, Universitat Politècnica de València, Camí de Vera, s/n, 46022, València.
mfullana@mat.upv.es,
WWW home page: https://imm.webs.upv.es/ca/
[2] Àrea d'Enginyeria, Florida Universitària,
Carrer del Rei En Jaume I 2, Catarroja, València. 46470, Spain.

Abstract. Following the line started in [1], some interesting numerical and computational techniques used in part of our research in Relativitisc Positioning Systems (RPS) are here explained (see [2–4]). As remarked in [1] (with other algorithms we have built), we also think the algorithms described in this paper can be applied to other areas in the fields we work and can also be extended in Science and Technology in general. One of our innovations is the application of HEALPix mollweide maps in the determination of the positioning errors. This technique is applied in the analysis of Cosmic Microwave Background (CMB) anisotropies and we have extended it to RPS.

Keywords: Methods: numerical, General Relativity, CMB, Relativistic Positionig Systems

1 RPS

Simulations of the Galileo ESA and GPS NASA Satellites Constellations were described in [5]. That description was made in GRT (General Relativity Theory) solving the timelike geodesics of those satellites. Circular trajectories with Schwarzschild metric, which have the same centre as the Earth, were used. Those orbits are named the nominal orbits. Then a description of the null geodesics of photons emitted from such satellites were simulated in Minkowskian space-time. The light trajectories from emission to reception inertial coordinates were represented by the analytical solution of [6]. Numerical algorithms were implemented to performe such calculations. The relativistic positioning of an event was there computed. The so-called emission coordinate region and co-region, the bifurcation problem (double localization) in the positioning of the receiver satellite, among other research were then pointed out for such satellites and a detailed discussion was there performed.

In [7] those algorithms were applied to compute positioning errors using RPS. Those errors are due to the uncertities in the description of the satellite world lines. In that paper, a new approach was considered. The satellite orbits are

not circles. They have perturbations due to multiple causes. Such perturbations must be taken into account to better describe the satellite orbits. Statistical perturbations of the nominal orbits were simulated in [7]. The formula from [6] was computed for four satellites described with nominal orbits and with statistically perturbed orbits. The difference of positioning with both orbital descriptions was defined as the U-error, Δ_d.

Further on, the null geodesics of the signal photons were computed in a more realistic way. That computation was compared with the movement in Minkoskian space-time (see [6]). That difference was called the S-error. In [8] the S-errors were compared with the U-errors. For the same conditions, the S-errors values were much smaller than the U-errors. As a conclusion, the hypothesis that the Earth's gravitational field produces negligible effects on photons can be used in a large region surrounding Earth. This assumption simplifies numerical computations. In our works, this conjecture is used.

Recently, our previous calculations have been improved (see [2]). The perturbed orbits of the satellites have been directly computed considering a GRT space-time metric from the beginning. No linear perturbation have been used. This metric took into account the gravitational effects of the Earth, the Moon and the Sun, and also the Earth oblateness. Moreover, in a future more perturbations could be considered. However, the order of magnitude of every contribution depends on the distance to the Earth and this fact should be considered in the metrics. A GRT metric was used from the first step to describe the space-time. The time-like satellite geodesics equations were calculated. A study of the satellite orbits in these new metrics was first made. Once this study was completed, the algorithm presented in [7] was developed, using a new analysis of the U-errors inside a great region surrounding the Earth. This analysis was performed comparing a great deal of positions given with the Schwarzschild metric and the new metric introduced there. The computations made in [2] improved the previous ones.

So far, we have made a summary of our research in RPS. **One of the important novelties we have introduced in all this work respect to other works in the literature is the use of HEALPix (Hierarchical Equal Area iso-Latitude Pixelisation) mollweide maps (see [9]) to describe the positioning errors. This technique represents an advance respect previous calculations in RPS and an extension of the work made in CMB research to another field in Astrophysics**. Now we will make a brief description of this generalisation.

2 HEALPix mollweide maps applied to RPS

HEALPix mollweide maps were first used in the study of CMB sky and its anisotropies (see [9–11]). The first probe launched to study the CMB sky was COBE from NASA in 1992. There have been a lot of studies related with this spacial mission. Some of the main papers related to the use of HEALPix and

COBE results are [12–14]. A much more precise satellite was launched from ESA in 2009, it is the PLANCK mission (see for instance [15]).

We have been working on the study of CMB anisotropies for a long time from the first nineties of last century. First using Lemaître-Tolman-Bondi metric to describe the effect of a great inhomogeneity on the CMB. That way we computed the CMB anistropies caused by great voids and Great Attractor structures. See for instance [16, 17]. Afterwards we used N-body codes (our own codes and the Hydra Consortium codes) to simulate regions of the Universe and describe the effect on CMB. See for example [18–20]. In this research we used HEALPix mollweide maps. An example of a map computed using mollweide techiques can be found in figure 4 of [18]. HEALPix mollweide maps represent the values of a physical quantity arriving from a celestial spherical surface centered in the observer. Those techniques are applied to CMB measurements. When we started our RPS computations we decided to use those numerical algorithms to the determination of positioning. Then we computed the positioning in a spherical surface. This was a novelty in satellite positioning and an improvement respect previous works.

Let us know show the HEALPix technique. This technique describes a hierarchical pixelisation of the celestial sphere. All pixels have the same area and those pixels with the same latitude have the same shape. It is designed to efficiently support:

i) Pixel location operations.

ii) The use of the hierarchical structure to change the resolution.

The HEALPix package contains optimized IDL and Fortran 90 subroutines.

HEALPix pixelization meets the following requirements:

1) **Hierarchical structure:** this is recognized as essential for very large databases. Figure 1 shows a hierarchical partition, with a quadrilateral tree structure, in which the pixels are easily numbered on a binary basis. By increasing the resolution, each pixel is subdivided into four child pixels (see Figure 1). Each one of these child pixels inherits all bits from the parent pixel and are distinguished between them by two new bits located to the right that are always 00, 01, 10, 11. In the lowest resolution the celestial sphere is divided into 12 patches, which are then hierarchically subdivided according to the above. Actually these are 12 patches, with four curvilinear sides, covering the entire celestial sphere.

2) **Equal areas for all pixels:** if the areas of the pixels are all the same, they cannot all have the same shape, as can be seen in Figures 1 and 2 of [9]. The pixels are more elongated in the polar areas.

3) **Distribution in iso-latitude:** there are sets of pixels with identical shape covering the parallels of the celestial sphere.

A distribution of pixels with the previous properties can be seen in Figure 1 of [9]. With this type of pixelization it is easy to search for a pixel of known coordinates. The 12 large pixels that define the lowest resolution are distributed in three rings around the Poles and the Equator in the way that can be seen in the upper left sphere of Figure 1 of [9].

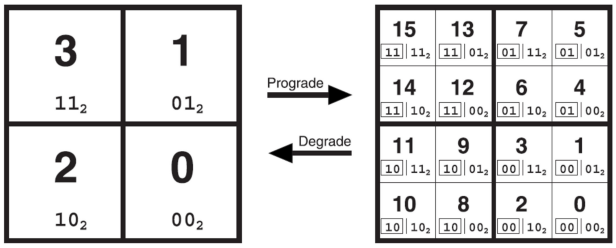


Fig. 1: .
Quadrilateral hierarchical tree structure. On the left is shown a coarsely pixelated patch, consisting of four binary base numbered pixels. On the right hand side the resolution of this patch is increased, so that each pixel is subdivided into four child pixels.

The number of pixels is $12 \times N_{side}^2$, where the free parameter N_{side} takes natural even values and indicates the resolution of the grid. For $N_{side} = 1$ we are at the lowest level of resolution and every time we divide the pixels into four parts, the parameter N_{side} is multiplied by 2; in this way, the total number of pixels is given by $N_{pix} = 12 \times N_{side}^2 = 12, 48, 192, 768$.

In our studies of the emission region, we have chosen $N_{side} = 16$, which is equivalent to considering 3072 pixels; while we use $N_{side} = 32$ (12288 pixels) in the case of the co-region. The angular area of each pixel is $\approx 13.43 \times 13.43\ deg^2$ for $N_{side} = 16$, whereas it is $\approx 3.36 \times 3.36\ deg^2$ if $N_{side} = 32$. Therefore, such area is very close to 64 (16) times the mean angular area of the full Moon for $N_{side} = 16$ ($N_{side} = 32$).

There are a total of $4 \times N_{side} - 1$ (parallel) rings over the entire celestial sphere. The pixel centres are equidistantly located on these rings. Three different zones on the sky are distinguished: one equatorial and two polar. The equatorial zone is the region bounded by the two parallels that pass through the upper and lower vertices lower than those base pixels of minimum resolution that have their centre in the equator (see Figure 1 of [9]). This zone separates the two polar zones. The rings located in the equatorial zone, are divided into the same number of pixels: $N_{eq} = 4 \times N_{side}$. The rest of the rings, located in the polar zones, contain a number of varied pixels, which grow from ring to ring as the distance to the objects increases through the poles.

In the equatorial region the equation of the curves that define the borders of the pixels is: $\cos\theta = a \pm b\phi$, while in polar caps it is: $\cos\theta = a + b/\phi^2$.

The geometric properties that characterize HEALPix allow us to number pixels in two different ways, as illustrated in Figure 2 of [9].

The two numbering schemes supported by HEALPix are called: RING and NESTED. Both schemes transform the two-dimensional (2D) distribution of discrete area elements (pixels) on the sphere, in a one-dimensional distribution, which is essential for calculations involving data sets with a number of very large total pixels.

The most convenient numbering system will be chosen depending on the type of numerical problem to solve. The numbering schemes for pixels are:

a) The RING scheme, in which pixels are numbered increasing, from the North Pole to the South Pole along each ring (iso-latitude).

b) The NESTED scheme, in which pixels are numbered in each of the 12 previously mentioned tree structures. Before we assigned a number in base 2 to each of the pixels (see our Figure 1) and the NESTED number of the pixel is the one that results when writing it in base 10.

We use RING numbering for all figures.

Finally, we visualize the pixelized sphere using the mollweide projection, in which divides the sphere into two parts: the frontal hemisphere is projected in the central part of the figure and the opposite hemisphere is represented in the lateral parts.

3 Results and Conclusions

In this paper, we have show how to extend the use of HEALPix mollweide maps from CMB maps to RPS maps. This is an example of how to use numerical tools created in some field of research to another ones. In [1], we described the use of instruments in numerical integration applied to a problem in Cosmology and the way they could be extended in other areas. Continuing with the field of research iniciated in [1], we now indicate another extension of a mathematical technique, as the mapping of the sky, to other similar physical processes.

Using HEALPix mollweide maps in RPS problems allows the determination of maximum and minimum positioning errors and other physical quantities in a spherical surface. Moreover, maximum and minimum values give us an idea of the variations and the precision of positioning with the use of different techniques and the contributions in the determination of this positioning too. The shape of HEALPix mollwide maps can be seen in our papers commented in the bibliography. For the shakness of briefness, we do not repeat the figures, but the reader can compare them with the CMB maps and observe their similarities and differences. Also the use we make of this kind of maps and the advances that suppose to generate them in RPS. The spherical surface we use can be centered in any place of the Earth. It can be the centre of it, any Earth surface position and even an observer away from our planet. The advantatge of this method is that one can determine the positioning in any observer in a wide range of places. Moreover, with the use of GRT from the beginning, the contribution of any gravitational term to the positioning can be determined.

Just to give an example of the advances that HEALPix method give us, we remark some useful results. In [2], the positioning errors values, Δ_d, are almost of the same order of magnitude as those of the perturbed satellite orbits (orbital perturbation effect). This is the same conclusion as in [7] although now a better accuracy is obtained. Here the highest Δ_d values correspond to having the maximum radial distance deviations of the satellite for the case of the four chosen satellites. Therefore, the value Δ_d directly depends on the satellite-Earth-Moon-Sun relative spatial configuration and it does so for each of the four satellites considered. Almost the same HEALPix maps are recovered after a Galileo satellite orbital period. This is because the relative spatial configuration among satellite-Moon-Sun-Earth does not nearly change after 14.2 h (periodic effect), as the Moon and Sun hardly move after a Galileo orbital period. Here the Δ_d values are smaller than the ones obtained with the statistical procedure used in [7]. This fact is due to a more realistic representation of the satellite orbital perturbations by the use of metrics instead of statistical deviations. Therefore a more accurate computation of the U-errors is performed and so a more precise calculation of the user's positioning can be achieved. This represents an advance in our computations respect our previous works.

Another interesting improvement in projection with the use of HEALPix is to create HEALPix maps, but computing the positioning error Δ_d on the geoid, instead of on the spherical surface of radius centred in the Earth. A better precision of orbits could be achieved. Data from Galileo Constellation, and other

constellations, could be compared. These results should also be interesting for geodesic treatment.

To summaraise, the perturbations computed using metrics improve our previous works based on statistical methods as:

1) A better description of the real satellite world lines is made.

2) The effect of each perturbing contribution in the satellite world lines and their combinations is studied. So, the orbits of the satellites are described depending on the terms considered.

3) Therefore, the contribution of each effect on the user's positioning can also be studied. One can add other gravitational contributions as those from Venus or Jupiter, and so on.

4) The value of the U-errors is now smaller.

5) That means a more precise computation of the user's positioning.

In summary, tools applied in the analysis of great observational data from CMB are extended to RPS: such as HEALPix representations, and also computational methods for solving the ordinary differential equations systems. They properly work in both cases. So, what we learn is that some mathematical and numerical tools working in some physical problems should be extended to other fields.

Now we are working on the analysis and description of the numerical advances made in other parts of our research. Such as other mathematical and numerical tools applied in the analysis of great observational data from CMB and extended to RPS, the numerical integration of the ordinary differential equations systems, N-Body treatment on CMB anisotropies, or new numerical applications on EMRIs and gravitational waves (a new field we have started). This work will be presented in the future.

Acknowledgments

We would like to acknowledge our great debt to Professor Diego Pascual Sáez Milán who was the pioneer of this research and worked in its theoretical and numerical approach. We worked together for a very long time in much of our common projects. He left us five years ago. There are no words to describe our gratitude, both for his scientific and his human teachings. We also acknowledge Dr. J.A. Morales-Lladosa for all his help. Dr. Pacôme Delva should also be mentioned for the same reason. This work has been supported by the Spanish Ministerio de Ciencia, Innovación y Universidades and the Fondo Europeo de Desarrollo Regional, Projects PID2019-109753GB-C21 and PID2019-109753GB-C22, the Generalitat Valenciana Project AICO/2020/125 and the Universitat de València Special Action Project UV-INVAE19-1197312.

References

1. Arnau i Córdoba, J.V., Fullana i Alfonso, M.J.: Resolution of Initial Value Problems of Ordinary Differential Equations Systems. Mathematics 10(4), 593 (1-27) (2022). https://doi.org/10.3390/math10040593

2. Puchades Colemenero, N., Arnau i Córdoba, J.V., Fullana i Alfonso, M.J.: Relativistic positioning: including the influence of the gravitational action of the Sun and the Moon and the Earth's oblateness on Galileo satellites. Astrophys. Space Sci. 366, 66, 19pp. (2021). doi:10.1007/s10509-021-03973-z
3. Fullana i Alfonso, M.J., Arnau i Córdoba, J.V., Puchades Colmenero, N.: Some advances in Relativistic Positioning Systems. In Modelling for Engineering & Human Behaviour 2021. València, July 14th-16th, 2021. Edited by: I.U. de Matemàtica Multidisciplinària, Universitat Politècnica de València. J.R. Torregrosa, J.C. Cortés, J. A. Hervàs, A. Vidal-Ferràndiz and E. López-Navarro, 2021. *ISBN: 978-84-09-36287-5*
4. Fullana i Alfonso, M.J., Sáez Milán, D.P., Arnau i Córdoba, J.V., Puchades Colmenero, N.: Some Improvements on Relativistic Positioning Systems. Applied Mathematics and Nonlinear Sciences 3(1), 161-166 (2018). doi:10.21042/AMNS.2018.1.00012
5. Puchades, N., Sáez, D.: Relativistic positioning: Four-dimensional numerical approach in Minkowski space-time. Astrophys. Space Sci. 341, 631-643 (2012). doi:10.1007/s10509-012-1135-1
6. Coll, B., Ferrando, J.J., Morales–Lladosa, J.A.: Positioning systems in Minkowski spacetime: from emission to inertial coordinates. Class. Quantum Grav. 27, 065013, 17pp. (2010). doi:10.1088/0264-9381/27/6/065013
7. Puchades, N., Sáez, D.: Relativistic positioning: Errors due to uncertainties in the satellite world lines. Astrophys. Space Sci. 352, 307-320 (2014). doi:10.1007/s10509-014-1908-9
8. Puchades, N., Sáez, D.: Approaches to relativistic positioning around Earth and error estimations. Advances in Space Research 57, 499-508 (2016). doiI:10.1016/j.asr.2015.10.031
9. Górski, K.M., Hivon, E., Wandelt, B.D.: In: Banday, A.J., Sheth, R.K., Da Costa, L. (eds.) Proceedings of the MPA/ESO Conference on Evolution of Large Scale Structure, pp. 37-42 (1999). Print partners Ipskamp Enschede. arXiv:astro-ph/9812350
10. HEALPix: Data Analysis, Simulations and Visualization on the Sphere. https://healpix.sourceforge.io/
11. Górski, K. M. et al.: HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. ApJ 622, 759-771 (2005). doi:10.1086/427976
12. Smoot G.F., et al.: Structure in the COBE Differential Microwave Radiometer First-Year Maps. ApJ 396, L1-L5 (1992). doi:10.1086/186504
13. Bennett, C.L., et al.: First-Year Wilkinson Microwave Anisotropy Probe (WMAP)* Observations: Preliminary Maps and Basic Results. ApJS 148, 1-27 (2003). https://doi.org/10.1086/377253
14. Bennett, C.L., et al.: Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results. ApJS 208, 20, 54pp. (2013). 10.1088/0067-0049/208/2/20
15. Akrami, Y., et al.: Planck intermediate results. LVII. Joint Planck LFI and HFI data processing. A&A 643, A42 (2020). https://doi.org/10.1051/0004-6361/202038073
16. Arnau, J. V., Fullana, M. J., Monreal, L., Sáez, D.: On the microwave background anisotropies produced by nonlinear voids. ApJ 402, 359-368 (1992). doi:10.1086/172139
17. Fullana, M.J., Arnau, J,V., Sáez, D.: Looking for the imprints of nonlinear structures on the cosmic microwave background. Vistas in Astronomy 41, Issue 4, 467-492 (1997). doi:10.1016/S0083-6656(98)00032-4

18. Puchades, N., Fullana, M.J., Arnau, J,V., Sáez, D.: On the Rees-Sciama effect: maps and statistics. MNRAS 370, 1849-1858 (2006). doi:10.1111/j.1365-2966.2006.10607.x
19. Fullana, M.J., Arnau, J,V., Thacker, R.J., Couchman, H.M.P., Sáez, D.: Estimating small angular scale CMB anisotropy with high resolution N-body simulations: weak lensing. ApJ 712, 367-379 (2010). arXiv:1001.4991v1
20. Fullana, M.J., Arnau, J,V., Thacker, R.J., Couchman, H.M.P., Sáez, D.: On the estimation and detection of the Rees-Sciama effect. MNRAS 464, Issue 4, 3784-3795 (2017). https://doi.org/10.1093/mnras/stw2615

On space-time in hydrogenoid atoms

Guillem Gómez i Blanch[1] and Màrius Josep Fullana i Alfonso[1]

Institut de Matemàtica Multidisciplinària
Universitat Politècnica de València
Camí de Vera s/n 46022 València, Spain.
e-mail:guigobla@doctor.upv.es; mfullana@mat.upv.es

Abstract. The contribution addresses the relationship between quantum particles and space-time by the study of a case, particularly simple: the hydrogenoid atoms. From the electronic orbit defined in the de Boglie-Bohm theory, we investigate the conditions to be interpreted as a geodesic in a Lorentzian manifold and we find a relation between two components of the metric. We need at least two further relations, so we go to the Quantum Gravity in the Wheeler-DeWitt theory. We use the so-called Einstein's Quantum Field equations and we reach another relation between the three components of the metric. We are working on the definition of additional relations for its complete definition.

Keywords: de Broglie-Bohm theory, General Relativity, Lorentzial metric, Wheeler DeWitt Canonical Quantum Gravity, Einstein's Quantum field equations.

1 From the De Broglie-Bohm theory and the Riemannian manifold to the geodesic condition

The contribution presented is a continuation of some previous works. ([1]), ([2]), ([3]). The starting conjecture is that the different performance of the electron when is free and when is integrated in an atom - where, according to the de Broglie-Bohm (dBB) ([4]) theory, describes circular trajectories when the quantum magnetic number is higher than 0 - could be explained by the fact that the space-time is perturbed and the electron describes on it geodesic trajectories. So the main hypothesis is that the action of the electrical and quantum potentials of the atomic system curve the space-time and the electron trajectory is a geodesic of it. Then, the relationship between the electron movement and the wave function would not only be guiding from the wave to the electron, but a mutual or dialectical relation between they, where, as is told in General Relativity, "space acts to matter, telling it how to move and matter reacts back on space, telling it how to curve"(Wheeler) [5].

1.1 Local relation between the Euclidean space of dBB and the Lorentzian manifold

Let's consider the dBB representation of the electron trajectory in an hydrogenoid atom. The position of the electron can be considered as the image of its position

in a Lorentzian manifold. The tangent metric shows a local identity between both metrics of the manifold and of the tangent space. Then it is possible to relate this metric with the Euclidean metric of a reference space. ([6])

Consider *a Euclidean space of reference*, with origin in the mass center of an hydrogenoid atom. In it, let be a point m (i.e., the electron position) with cylindrical coordinates $x^{'\alpha}$. We include the time coordinate, so we have a four-coordinate system, able to describe the dBB theory. We can write:

$$ds^2 = -d^2x^{'0} + d^2x^{'1} + (x^{'1})^2 d^2x^{'2} + d^2x^{'3} \tag{1}$$

It is possible to establish a relationship between a Euclidean space and a Riemannian manifold *at local scope, so in the neighborhood of a point* following the concepts of *tangent Euclidean metric* and *first-order representation*. Consider a Lorentzian manifold with signature $(-,+,+,+)$, being M_0, of coordinates x^α. We can put into correspondence the point M_0 of the manifold and a point m_0 of the Euclidean space, so m_0 is the *image* of M_0. Every point M in the neighborhood of M_0 can be mapped to a point m in the neighborhood of m_0, using second-degree functions $\Lambda_{(2)}$ of the difference of coordinates $y^\nu - y_0^\mu$ at the point of the manifold:

$$\overrightarrow{m_0 m} = [(y^\mu - y_0^\mu) + \Lambda^\mu_{(2)}(y^\nu - y_0^\nu)]\vec{e}_\mu \tag{2}$$

From (2), passing to the limit, it follows that:

$$\left(\frac{\partial \vec{m}}{\partial y^\mu}\right)_0 = \vec{e}_\mu \tag{3}$$

Thus, the point m in this Euclidean space is defined by the coordinates of the manifold y^μ; the values $y^\mu - y_0^\mu$ act as curvilinear coordinates of the point m.

Let us now consider $\bar{g}_{\mu\nu}$ the metric of the Euclidean space, defined by:

$$d\bar{s}^2 = \bar{g}_{\mu\nu} dy^\mu dy^\nu \tag{4}$$

Then, for $y^\mu = y_0^\mu$ the Euclidean and Riemannian metrics have the same values, and both metrics are said to be *tangent* at this point.

$$g_{\mu\nu} = \bar{g}_{\mu\nu} \tag{5}$$

So, both manifold and tangent Euclidean space have the *same* metric in this point.

The elemental distance between two points in the neighborhood considered in the Riemannian manifold and the Euclidean space is the same, so it is conserved:

$$\overline{m_0 m}^2 = (\bar{g_{\mu\nu}})_0 \, dy^\mu dy^\nu = (g_{\mu\nu})_0 \, dy^\mu dy^\nu = \overline{M_0 M}^2 \tag{6}$$

So ds^2 is conserved between the Lorentzian manifold and the Euclidean space. Now, since cylindrical and axially simmetric space-time allows the following metric's structure ([8]):

$$g_{\mu\nu} = \begin{pmatrix} g_{11} & 0 & 0 & 0 \\ 0 & g_{22} & 0 & g_{20} \\ 0 & 0 & g_{33} & 0 \\ 0 & g_{02} & 0 & g_{00} \end{pmatrix} \tag{7}$$

we can write:

$$ds^2 = -g_{00}d^2x^0 + 2g_{02}dx^0dx^2 + g_{11}d^2x^1 + g_{22}d^2x^2 + g_{33}d^2x^3 \tag{8}$$

The tensors defined *on the point* m in both Euclidean spaces will transform according to the rules of transformations of the coordinates; particularly the ds^2 as invariant, will have the same value in both reference systems.

If $\eta_{\alpha\beta}$ is the orthogonal Euclidean metric of the reference system (e.g. cylindrical) and $x^{'\alpha}$ its coordinates, we can write this equivalence as:

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = \eta_{\alpha\beta}dx^{'\alpha}dx^{'\beta} \tag{9}$$

And this indeed will be the metric of the tangent Euclidean space-time at the point m. Then we can relate the elemental interval in the tangent space and the space in the reference system, with $g_{00} = 1$ as:

$$2g_{02}dx^0dx^2 + g_{22}d^2x^2 = (x^{'1})^2d^2x^{'2} \tag{10}$$

Consider the motion of the electron around the proton in the dBB representation:

$$x^{'1} = \rho_0 \ ; x^{'2} = \phi = \frac{u\hbar t}{m\rho_0^2} \ ; x^{'3} = z \ ; x^{'0} = ct$$

We can make the following **approximate** considerations:

* $dx^0 = dx^{'0} = cdt$, because we work with velocities that are little respect c.
* $dx^1 = dx^{'1} = 0$ because the electron has a circular trajectory in the dBB theory.
* $dx^2 = dx^{'2} = \phi$ because of symmetry considerations.
* $dx^3 = dx^{'3} = z$ because the trajectory is plane in $\mathcal{E}_3$

So we can write dx^α instead of $dx^{'\alpha}$. Then we can establish the **quadrivelocity** as:

$$\begin{aligned} \frac{dx^1}{dt} &= \dot{\rho} = 0 \\ \frac{dx^2}{dt} &= \dot{\phi} = \omega = \frac{u\hbar}{m\rho_0^2} \\ \frac{dx^3}{dt} &= \dot{z} = 0 \\ \frac{dx^0}{dt} &= c \end{aligned} \tag{11}$$

Then the equation (10) simply runs:

$$g_{22} = \rho_0^2 - \frac{2cg_{02}}{\omega} \tag{12}$$

We now introduce the **quantum condition** that all electrons that may belong to the same quantum state corresponding to the same atomic orbital must possess **the same angular or kinetic moment**:

$$m\omega\rho_0^2 = u\hbar \tag{13}$$

where $\hbar$ is the reduced Planck constant, u is the magnetic quantum number and ρ_0 *is the radius of the orbit, while* ρ *is the generic coordinate of space-time*, which matches ρ_0 in the electron's trajectory.

And we introduce the constant: f, the reduced Compton's length:

$$f = \frac{\hbar}{mc} \tag{14}$$

with which we can express, by substituting ω in (12):

$$\boxed{g_{22} = \rho_0^2 \left(1 - \frac{2}{uf} g_{02}\right)} \tag{15}$$

This relation is very important for our purpose; *we will see that it is a particular geodesic of the Lorentzian manifold.*

1.2 Geodesic condition

The figure (1) represents, in a hydrogenoid atom, the composition of the electrostatic force F_E and the quantum force F_Q, derived from the quantum potential, both in a constant phase plane. Their resultant is the centripetal force F_C. But we can consider this movement as a geodesic in a Lorentzian manifold, so the effect of forces is substituted by the geometry constraint.

We introduce the hypothesis that the electron's trajectory according to dBB corresponds to a geodesic in a Lorentzian manifold, as described by a null covariant derivative:

$$\frac{d^2x^\mu}{dt^2} + \Gamma^\mu_{\nu\lambda} \frac{dx^\nu}{dt} \frac{dx^\lambda}{dt} = 0 \tag{16}$$

with parameter the proper time, that we identify as the observator time (speed in the range of $10^{-2}c$). Replacing velocities (11) to (16) we get:

$$\omega^2 \Gamma^\mu_{22} + 2\omega c \Gamma^\mu_{02} + c^2 \Gamma^\mu_{00} = 0 \tag{17}$$

The affine connectors can be expressed with respect to the metric tensor as follows:

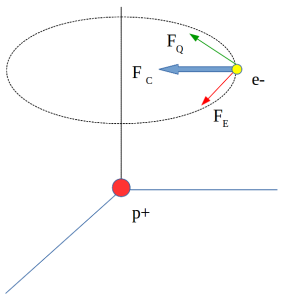


Fig. 1: Forces acting on the hydrogen electron according to dBB theory.

$$\Gamma^{\mu}_{\nu\lambda} = \frac{1}{2} g^{\mu\delta} (\partial_{\mu} g_{\lambda\delta} + \partial_{\lambda} g_{\mu\delta} - \partial_{\delta} g_{\mu\lambda}) \tag{18}$$

We now make the hypothesis that the components of the metric do not depend on the angle, $\phi = x^2$, the $z = x^3$ or the time $ct = x^0$. In the components of the metrics, we will only consider variation with respect to the $\rho = x^1$ coordinate as possible.

Replacing this in the equation that expresses the connector depending on the metric g_{ij} we can write the non-zero connectors and the geodesics' equations:

$$\Gamma^{\mu}_{22} = \frac{1}{2} g^{\mu\delta} (\partial_2 g_{2\delta} + \partial_2 g_{\delta 2} - \partial_{\delta} g_{22}) = -\frac{1}{2} g^{\mu\delta} \partial_{\delta} g_{22} = -\frac{1}{2} g^{\mu 1} \partial_1 g_{22} \tag{19}$$

$$\Gamma^{\mu}_{02} = \frac{1}{2} g^{\mu\delta} (\partial_2 g_{0\delta} + \partial_0 g_{\delta 2} - \partial_{\delta} g_{02}) = -\frac{1}{2} g^{\mu\delta} \partial_{\delta} g_{02} = -\frac{1}{2} g^{\mu 1} \partial_1 g_{02} \tag{20}$$

$$\Gamma^{\mu}_{00} = \frac{1}{2} g^{\mu\delta} (\partial_0 g_{0\delta} + \partial_0 g_{\delta 0} - \partial_{\delta} g_{00}) = -\frac{1}{2} g^{\mu\delta} \partial_{\delta} g_{00} = -\frac{1}{2} g^{\mu 1} \partial_1 g_{00} \tag{21}$$

Recalling that the axially symmetric metrics have the above mentioned structure, that $g^{11} \neq 0$, indicating by $g^{'}$ the total derivatives with respect to x^1 and using previous equations, we come to the by us named "Geodesic dBB theorem":

$$u^2 f^2 g'_{22} + 2uf\rho_0^2 g'_{02} + \rho_0^4 g'_{00} = 0 \tag{22}$$

where ρ_0 is the electron's orbit radius, f is the Compton length and u is the quantum magnetic number. With g_{00} constant, we come to:

$$\boxed{u\, f\, g'_{22} + 2\, \rho_0^2\, g'_{02} = 0} \tag{23}$$

The integration of the differential equation (23) runs:

$$u\, f\, g_{22} + 2\, \rho_0^2\, g_{02} = constant \tag{24}$$

but it must also be accomplished (15), so the constant is defined as $\frac{u f^{\bullet}}{2}$ and we establish the following equation:

$$\boxed{g_{22} = \rho_0^2 \left(1 - \frac{2}{uf} g_{02}\right)} \tag{25}$$

So, this equation (25) is, between the bundle of solutions given by (24), the particular solution that fulfills both conditions of elemental distance conservation and of geodesic, with the approximations and hypotheses made. There is a multiplicity of metrics that accomplish this condition. First, the equation gives a bundle of possible solutions for the function components. Secondly, the component g_{11} does not appear in the equation, so is undetermined.

That indicates that further (at least two) conditions are required to determine the metrics of the Lorentzian manifold in the vicinity of the particle.

2 Quantum gravitation

An avenue to overcome the missing description previously mentioned is to look for the relationship between the energy-moment content and the geometry. It addresses us to the quantum gravity: to the so-called "Einstein's quantum field equations" that have been recently formulated in the frame of the 3+1 and specifically the Wheeler-deWitt theory. ([9]),([10]), ([11]),([12])

We start with the *3+1 model of General Relativity (ADM)*. It describes the space-time as a set of 3-dimensional hypersurfaces, each one labeled by the time. So, time is constant in every hypersurface.

From there, it is possible to generate a Lagrangian and a Hamiltonian formulation that allows for defining the dynamical equations and Einstein's field equations.

The translation to the quantum conception is possible through a quantification, by changing the canonical coordinates to operators. We come then to the Wheeler-DeWitt theory of quantum gravity.

When applied to mass-particles (fermions) it is also possible to derive Einstein's field equations for these quantum objects.

We will try to adopt this development to complete the definition of our previous theory. We will adopt here the signature (+ - - -) and units according to $\hbar = c = 1$. Greek indexes vary from 0 to 3 and Latins from 1 to 3.

2.1 Classical approach

Let a coordinate point $(t, x^i(t))$ be in the Σ_t hypersurface that has $h_{ij}(t)$ 3-dimensional metric and the coordinate point $(t+dt, x^i)$ in the Σ_{t+dt} hypersurface that has $h_{ij}(t+dt)$ metric.

Let the *proper time* between the two events be Ndt, N the ***lapse function***. Let the space coordinates of the point of Σ_{t+dt} be $x^i + N^i dt$ with N^i as functions

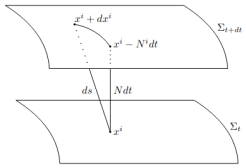


Fig. 2: Two hypersurfaces separated by a time differential (S. Carlip, [4] (2019).

that determine the displacement vector. So, the differential line element (Figure 2):

$$ds^2 = N^2 dt^2 - h_{ij}(dx^i + N^i dt)(dx^j + N^j dt) \tag{26}$$

The action of the system (Einstein-Hilbert S_G plus non-relativistic particle S_M):

$$S_{cl} = S_G + S_M = -\frac{1}{\kappa}\int R\sqrt{-g}d^4x - m\int N dt + \frac{m}{2}\int \frac{h_{ij}}{N}(\dot{X}^i + N^i)(\dot{X}^j + N^j)dt \tag{27}$$

where R is the scalar curvature of the manifold; N and N^i are functions of t; $\kappa = 16\pi G$. The canonical variables are X^i and $H_{ij} = g_{ij}$. Then we can formulate the Lagrangian L and the Lagrangian density $\mathcal{L}$. From it we can formulate the canonical moments:

$$P_i = \partial_{\dot{X}^i} L_{cl} = \frac{m}{N(X)}\left(\dot{X}_i + N_i(X)\right), \quad \pi_{ij} = \partial_{\dot{h}_{ij}}\mathcal{L}_{cl} = -\frac{1}{\kappa}\sqrt{h}(K^{ij} - Kh^{ij}) \tag{28}$$

, and from there the energy-momentum tensor, the energy component of which is T_C^{00}. If we integrate it over the hypersurface Σ we obtain the total energy:

$$\int_\Sigma d^3x T_C^{00} = \int_\Sigma d^3x \frac{m + \frac{1}{2m}P_k(t)P^k(t)}{N^2(t,x)\sqrt{h(t,x)}}\delta(\mathbf{x} - \mathbf{X(t)}) = \frac{m + \frac{1}{2m}P_k(t,X)P^k(t,X)}{N^2(t,X)\sqrt{h(t,X)}} \tag{29}$$

2.2 On Wheeler DeWitt canonical quantization

In Wheeler-DeWitt theory, the "state" of the system is a *functional* of particle coordinates and 3-metric, $\Psi(X, h_{ij})$; the Wheeler-Dewitt equation holds:

$$\hat{\mathcal{H}}\Psi = \hat{\mathcal{H}}_G\Psi + \hat{\mathcal{H}}_M\Psi \tag{30}$$

where:

$$\hat{\mathcal{H}}_G = -\kappa G_{ijkl}\frac{\delta^2}{\delta h_{ij}\delta h_{kl}} + \mathcal{V}(h,x)\ ;\ \hat{\mathcal{H}}_M = \delta(x - X)(m - \frac{\nabla^2}{2m}) \tag{31}$$

with $\mathcal{V}(h,x)$ as effective potential density and G_{ijkl} the DeWitt metrics. Also, there is the diffeomorphic constraint(D_i covariant derivation):

$$\hat{\mathcal{H}}_i\Psi = \hat{\mathcal{H}}_{Gi}\Psi + \hat{\mathcal{H}}_{Mi}\Psi \ ; \ \hat{\mathcal{H}}_{Gi} = 2h_{ik}D_j\frac{\delta}{\delta_{h_{jk}}} \ ; \ \hat{\mathcal{H}}_{Mi} = \delta(x-X)\nabla_i \tag{32}$$

It is important to notice that when writing the wave equation in polar form, $\psi = |\psi|e^{iS}$ and entering the conditions $P_i = \nabla_i S$ and $\pi^{ij} = \frac{\delta S}{\delta h_{ij}}$ we get the particle guide equations:

$$\dot{X}^i = \frac{N(X)}{m}\nabla^i S - N^i(X) \ ; \ \dot{h}_{ij} = 2\kappa N G_{ijkl}\frac{\delta S}{\delta h_{ij}} + D_i N_j + D_j N_i \tag{33}$$

Varying total action (classical and quantum) to the metric $g_{\mu\nu}$ we can derive the Einstein's *Quantum* Field Equations. We get, with the cosmological constant $\Lambda = 0$:

$$G_{\mu\nu} = R_{\mu\nu} - \frac{g_{\mu\nu}}{2}R = 8\pi G(T_{C\mu\nu} + T_{Q\mu\nu}) \tag{34}$$

So we must add to the previously calculated T_C, the "Quantum contribution": the tensor $T_{Q\mu\nu}$. We are mainly interested in the density energy component T_Q^{00}:

$$T_Q^{00} = \frac{1}{N^2(t,x)\sqrt{h(t,x)}}\left(-\kappa G_{ijkl}\frac{1}{|\Psi|}\frac{\delta^2|\Psi|}{\delta h_{ij}\delta h_{kl}} - \delta(\mathbf{x}-\mathbf{X})\frac{\hbar^2\nabla^2|\Psi|}{2m|\Psi|}\right) \tag{35}$$

Integrating it into a 3-d (closed) hypersurface, we obtain the total particle(s)'s energy :

$$\begin{aligned} E_t = \int_\Sigma d^3x\, T^{00} &= \frac{1}{N^2(t,x)\sqrt{h(t,x)}}\left(m + \frac{1}{2m}P_k(t)P^k(t) - \frac{\hbar^2\nabla^2|\Psi|}{2m|\Psi|}\right) \\ &- \kappa\int_\Sigma d^3x G_{ijkl}\frac{1}{|\Psi|}\frac{\delta^2|\Psi|}{\delta h_{ij}\delta h_{kl}} \end{aligned} \tag{36}$$

that can provide an additional equation to define the metric of space-time.

In two-particle systems such as the hydrogen atom, in which one particle is much heavier than the other one, the Schrödinger wave function coincides with that of Wheeler-DeWitt, ([11]), which allows us to use atomic orbitals as such.

The spatial metric $h_{ij} = g_{ij}$ remains constant in time: $\dot{h_{ij}} = 0$. The geodesic hypothesis links the derivatives of the components of the metric g'_{22} and g'_{02}.

As we have advanced previously, the metric $g_{\mu\nu}$ allows the following expression:

$$g_{\mu\nu}dx^\mu dx^\nu = (N^2 - N_k N^k)dt^2 - 2N_k dx^k dt - h_{ij}dx^i dx^j \tag{37}$$

Expressing the differential element ds^2 based on (26) and matching the above equation yields:

$$g_{00}dt^2 = (N^2 - N_k N^k)dt^2 \ ; \ -2h_{02}dx^2dx^0 = -2N_k dx^k dt \tag{38}$$

where the displacement vector has components:

$$N_1 = 0 \ ; \ N_2 = h_{02} \ ; \ N_3 = 0 \tag{39}$$

2.3 Elements of a first approximation to the application to the hydrogenoid atoms II.

We can add a new condition for the metric tensor, concerning also the component g_{11}. In our treatment of the hydrogenoid atom, we can come back to the expression of T^{00} (35), which allows the calculation of the total integral energy on a hypersurface. The $\delta(x - X)$ in the mentioned reference indicates that the integral can be done in a very reduced neighborhood of ρ as the energy of the system is reduced to the nucleus and the electron. This energy must be, for low particle momentum, simply $M - En$. Therefore, we get:

$$\begin{aligned} E_T =& M - E_n = \int_\Sigma dx^3 T^{00} = \\ &= \frac{\sqrt{g_{22}}}{(g_{22} - g_{02}^2)g_{11}} \left[M + \frac{m\left(\frac{u}{m\rho^2} + g_{02}\right)^2}{2(g_{02}^2 - g_{22})} - E_n + \frac{u^2}{2m\rho^2} - \frac{q_e^2}{4\pi\epsilon_0\sqrt{\rho^2 + z^2}} \right] \end{aligned} \tag{40}$$

allowing g_{11} to be defined as a function of g_{22} and g_{02}, previously defined.

$$\begin{aligned} g_{11} =& \\ &= \frac{\sqrt{g_{22}}}{(g_{22} - g_{02}^2)(M - E_n)} \left[M + \frac{m\left(\frac{u}{m\rho^2} + g_{02}\right)^2}{2(g_{02}^2 - g_{22})} - E_n + \frac{u^2}{2m\rho^2} - \frac{q_e^2}{4\pi\epsilon_0\sqrt{\rho^2 + z^2}} \right] \end{aligned} \tag{41}$$

3 Conclusions

The hypothesis that the electrons in the hydrogenoid atoms describe geodesics of Lorentzian manifolds, in the frame of the dBB theory and with the assumed approximations, is found to be coherent with the Riemmannian geometry. A condition is derived that allows to distinguish the only geodesic that fulfills all the requirements.

The transposition to Quantum Gravity to go deeper in the interpretation allows the definition of an additional constraint that relates the total energy of the atomic system with all the components of the metrics.

We are working on the detailed development of the theory, in order to provide a complete form for the metric. So, further work is required to achieve the complete metric definition.

Acknowledgments

This work has been supported by the Spanish Ministerio de Ciencia, Innovación y Universidades and the Fondo Europeo de Desarrollo Regional, Projects PID2019-109753GB-C21 and PID2019-109753GB-C22, the Generalitat Valenciana Project AICO/2020/125 and the Universitat de València Special Action Project UV-INVAE19-1197312.

References

1. G. Gómez Blanch and alii:Space-time geometry in the atomic hydrogenoid system, approach to a dust relativistic model from Causal Quantum Mechanics.Revista Mexicana de Física (4),18-29 (2018).
2. Gómez Blanch and M.J. Fullana Alfonso. On geometrodynamics in atomic stationary states. Revista Mexicana de Física, (65),148-158, (2019)alii:Space-time geometry in the atomic hydrogenoid system, approach to a dust relativistic model from Causal Quantum Mechanics.Revista Mexicana de Física (4),18-29 (2018).
3. G.Gómez Blanch:Tesi doctoral, Varietats Lorentzianes en la representació dels estats estacionaris dels àtoms hidrogenoides en la teoria de de Broglie- Bohm. Uns models heuristics. Universitat Politècnica de València (2021).
4. P.R.Holland: The Quantum theory of Motion.Cambridge University Press. Cambridge (1993).
5. C.W. Misner, K.S.Thorne, J.A. Wheeler: Gravitation. Princenton University Press (2017).
6. A. Lichnerowicz: Elementos de cálculo tensorial. AguilarS.A, Madrid (1968)
7. S. Carlip: General Relativity. Oxford University Press, Oxford, (2019).
8. H. MacCallung, M. Herlt, E. Kramer, D. Stephani: Exact solutions of Einstein's fiels equations. Cambridge University Oress (1980)
9. F.Shojai and M. Golshani:On the general covariance in Bohmian Quantum Gravity. Int.J. Mod.PHys. A13, 2135-2144 and arXiv:gr-qc/9903047, 1-17, (1998)
10. A. Shojai and F.Shojai:Constraint algebra and equations of motion in the Bohmian interpretation of Quantum Gravity. Pramana 58, pages 13-19 (2004).
11. D. Dürr and W. Struyve. Quantum Einstein Equations (2020).
12. C. Maes, K.Meerts and W. Struyve:Wave mechanics for gravity with point particles. Classical and Quantum Gravity 38,175003, 2021

Ballistic coefficient and life estimation for LEO satellites

A. Herrero[1], S. Moll[2] J.A. Moraño[1], and David Soriano[3]

[1] Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, 46022 Valencia, Spain,
{aherrero, jomofer}@mat.upv.es
[2] Departamento de Matemática Aplicada, Universitat Politécnica de València, València, Spain,
sanmollp@mat.upv.es
[3] Escuela Técnica Superior de Ingeniería del Diseño, Universitat Politècnica de València, 46022 Valencia, Spain,
davidsorianopascual@gmail.com

Abstract. In this paper, a semi-analytical orbit propagator is developed to study trajectories of satellites orbiting between 200 and 2000 km of altitude considering the atmospheric drag coefficient, the oblateness of the Earth and the effects of the Sun and the Moon in the orbit. The main goal of the propagator is to predict the decay process of the orbit and the re-entry of the satellite into the atmosphere. In this sense the ballistic coefficient is one of the most important characteristics of the satellite to be considered. This coefficient depends on several factors as the shape of the satellite, its attitude and the temperature and composition of the atmosphere. The implemented algorithm takes into account these factors and their variability in order to solve the nonlinear equations of the spacecraft motion obtaining an optimal balance between precision and computational cost. Moreover, the tool makes a self verification study checking with known data of the chosen satellite.

Keywords: ballisitic coeficient, LEO satellites, life estimation

1 Introduction and Statement of the problem

The number of satellites and objects orbiting around the Earth at altitudes between 200 and 2000 km has increased considerably in recent years. Most of these objects are monitored by means of telescopes, radars and other stations and are available at the NORAD (North American Aerospace Defense Command) website [1]. However, knowing the trajectory of these objects as accurately as possible in order to predict their positions can help us to avoid collisions and other potential hazards.

Low Earth Orbit (LEO) satellites are expected to decay into the Earth atmosphere in a short period of time mainly due to the atmospheric drag. To estimate their lifetime, it is necessary to use prediction tools that solve the corresponding

motion equations. The precision and efficiency of these tools depends on the orbit propagation model and the perturbation forces considered on it, and the accuracy of the numerical integration method. These models consider the Keplerian equations of motion and the dominant perturbation forces which, for LEO satellites, are the atmospheric drag coefficient, the oblateness of the Earth and the effects of the Sun and the Moon.

The atmospheric drag is related with the atmospheric density and the ballistic coefficient of the satellite. While the atmospheric density is independent of the considered satellite and varies with time, solar activity, altitude and geographic location; the ballistic coefficient is a parameter that includes different characteristics of the satellite like its drag coefficient, cross-sectional area and mass. Determining this ballistic coefficient is one of the most important objective of this work, as it enables the development of an accurate propagator. This propagator is designed to predict the re-entry of a Low Earth Orbit (LEO) satellite, while taking into account the most significant perturbation accelerations affecting it.

So, this work develops an integrator of the spacecraft perturbed motion equations using Matlab R2018a software. This integrator includes an iterative algorithm to determine the ballistic coefficient of the satellite under study. Moreover, the implemented propagator is validated by checking the evolution of the solution with some given data of the satellite.

The Keplerian motion equations of a satellite with perturbations on the acceleration, according to the Cowell's formulation [2], is

$$\ddot{\mathbf{r}}(t) = -\frac{\mu}{r^3}\,\mathbf{r} + \mathbf{a}, \tag{1}$$

where $\mu = 398604.4\ \mathrm{km}^3/\mathrm{s}^2$ is the Earth's gravitational parameter, $\mathbf{r}$ represents the position of the satellite relative to the Earth, r is the modulus of the vector position, dots are time derivatives and $\mathbf{a}$ is the perturbing acceleration vector that includes, by addition, the main perturbing sources such as the Earth's oblateness ($\mathbf{a}_J$), the gravitational effects of the Sun and Moon ($\mathbf{a}_m$ and $\mathbf{a}_\odot$, respectively), and atmospheric drag perturbations ($\mathbf{a}_D$), which are primarily significant for a spacecraft orbiting the Earth at lower orbits. That is,

$$\mathbf{a} = \mathbf{a}_J + \mathbf{a}_m + \mathbf{a}_\odot + \mathbf{a}_D. \tag{2}$$

The first perturbation considered is a well-known perturbation that is usually included in the motion equations when carrying out long-term orbital propagation. It considers the Earth as an oblate spheroid and adds the acceleration perturbation corresponding to each zonal harmonic of the planet [3]. The five first harmonics have been implemented in the propagator to have a more accurate model, that is

$$\mathbf{a}_J = \mathbf{a}_{J_2} + \mathbf{a}_{J_3} + \mathbf{a}_{J_4} + \mathbf{a}_{J_5}. \tag{3}$$

Relative to the Sun and Moon effects on the acceleration, a three-body approach has been carried out for each one of them getting that these effects can

be included as [4]:

$$\mathbf{a}_m = \mu_m \left(\frac{\mathbf{r}_{m/s}}{r_{m/s}^3} - \frac{\mathbf{r}_m}{r_m^3} \right) \quad \text{and} \quad \mathbf{a}_\odot = \mu_\odot \left(\frac{\mathbf{r}_{\odot/s}}{r_{\odot/s}^3} - \frac{\mathbf{r}_\odot}{r_\odot^3} \right), \tag{4}$$

where $\mathbf{r}_{m/s}$ and $\mathbf{r}_{\odot/s}$ are the position of the Moon and the Sun relative to the satellite, $\mathbf{r}_m$ and $\mathbf{r}_\odot$ are the position of the Moon and the Sun relative to the Earth, respectively, and $\mu_m = 4903$ km^3/s^2 and $\mu_\odot = 1.32712\dot{4}10^{11}$ km^3/s^2.

So, the implemented integrator will include an addition of all these acceleration effects together with the atmospheric drag perturbation, which is described in the next section.

.

2 Atmospheric drag perturbation

The atmospheric drag perturbation $\mathbf{a_D}$ is mainly obtained from the atmospheric density and the ballistic coefficient of the satellite as [2]:

$$\mathbf{a_D} = -\frac{1}{2}\beta\ \rho\ v_{rel}^2 \frac{\mathbf{v}_{rel}}{|\mathbf{v}_{rel}|} \tag{5}$$

where ρ is the atmospheric density, $\mathbf{v}_{rel}$ is the relative velocity vector of the satellite and β is the ballistic coefficient, which depends on the characteristics of the spacecraft. The more realistic the density and ballistic coefficient estimates are, the more accurate the orbital propagation model will be.

The atmospheric density ρ strongly depends on the solar activity and the geomagnetic fluctuations caused by solar storms. Solar activity results in long-term 11-year solar cycles and other effects such as diurnal, seasonal, longitudinal and latitudinal fluctuations [2]. It can be quantified by means of the solar radiation indices $F_{10.7}$ (solar flux emitted at a wavelength of 10.7 cm), S_{10} (Extreme Ultra Violet radiation), M_{10} (Mid Ultra Violet radiation) and Y_{10} (a weighted index between X-ray emissions and Lyman-α ones). On the other hand, the geomagnetic fluctuations are included by means of the geomagnetic activity index A, which is calculated as the average value resulting from daily observations made every 3 hours from 12 locations around the world.

Then, under all these considerations, different models to evaluate ρ have been proposed in the literature [2, 3] depending on the altitude and the solar and geomagnetic activities. In this work, two dynamic models are implemented to estimate ρ, the Jacchia-Bowman 2008 (JB08) model [5] and the Harris-Priester (HP) model [2]. The first one considers most of the solar activity phenomena to give ρ in an empiric way, while the second one only considers diurnal variations and an average of the solar activity index and estimates ρ as an exponential interpolation from statistic tables.

Since the aforementioned solar indexes are only available from 1997 until 1 or 2 months before the present day, the JB08 atmospheric model can be used as an estimation of ρ until that date and then the HP model is used to estimate ρ in the final evolution time.

2.1 Ballistic coefficient estimation algorithm

The ballistic coefficient β involves the drag coefficient, C_D, the surface, A, and the mass, m, of the satellite. It is an important coefficient to be determined in order to get more realistic trajectories for the satellite and a better prediction of its lifetime. This coefficient is defined as

$$\beta = \frac{C_D A}{m} \tag{6}$$

Note that β is not constant since it depends on the drag coefficient which is related to the temperature and the orbit altitude. Throughout this work the surface A and the mass m are assumed to be constant along the orbit of the satellite.

Several models have been developed to estimate β using the Two-Line-Element (TLE) data [6, 7]. These TLE data provide β through the value B^* by the expression:

$$\beta = \frac{2B^*}{\rho_0} \tag{7}$$

where $\rho_0 = 0.1570$ kg·m^{-2}·R$_E^{-1}$ is the reference value of atmospheric density, and R$_E = 6371$ km is the mean Earth radius.

So, starting from a TLE data, the first estimation of the ballistic coefficient is given by the expression (7). Next, an iterative process based on the Secant method begins to get a better estimation of this parameter [7].

It is worthwhile to note that before the iterative algorithm begins, the code makes a filtering of the TLE data to remove those data that are not updated or have interferences with other satellites. Once this filtering has been done, the process begins as follows:

1) The first estimation of the ballistic coefficient, β_1, is obtained from the first TLE data using (7).
2) With this β_1 use the propagator to integrate the motion equations to the epoch of the following TLE data of the set.
3) Obtain the difference between the propagated semi-major axis, a_{prop1}, and the semi-major axes given by the second TLE data, a_{TLE2}:

$$\Delta a_{prop1} = a_{prop1} - a_{TLE2}$$

4) The second estimation of the ballistic coefficient will be

$$\beta_2 = \frac{\Delta a_{TLE}}{\Delta a_{prop1}} \beta_1$$

where $\Delta a_{TLE} = a_{TLE2} - a_{TLE1}$.

5) Now, use β_2 to propagate again from the first TLE data to the second TLE data epoch.
6) Calculate the difference between the new propagated semi-major axis, a_{prop2}, and the semi-major axes given by the second TLE data, a_{TLE2}:

$$\Delta a_{prop2} = a_{prop2} - a_{TLE2}$$

7) The third estimate will be:

$$\beta_3 = \beta_2 - \Delta a_{prop2} \cdot \frac{\beta_2 - \beta_1}{\Delta a_{prop2} - \Delta a_{prop1}}$$

8) Again, make a propagation to the second TLE data epoch using β_3.
9) Calculate the difference between the new propagated semi-major axis, a_{prop3}, and the semi-major axes given by the second TLE data, a_{TLE2}:

$$\Delta a_{prop3} = a_{prop3} - a_{TLE2}$$

10) If $\Delta a_{prop3} > 10^{-4}$km then repeat the process from 7), else continue in 11).
11) Repeat the whole process for the following pair of TLE data in the set.

This process end when all the pairs of TLE data have been analyzed and provides us an estimate of the ballistic coefficient that improves the initial value obtained from the TLE data.

3 Validation of the implemented code

In this section, a validation of the implemented integrator has been done. For that, two small spherical satellites, Starshine I and Starshine II [8], are used since their deorbit dates are known and their surface and area can be considered constant along their trajectory. In fact, both satellites have the same size and mass, that is, 0.178 m^2 and 39.46 kg but were launched in different epochs. The first one was launched on May 27th, 1999 and deorbits on February 18th, 2000; and the second one was launched on May 12th, 2001 and made its reentry on April 26th, 2002. It can be directly observed that their timelifes were different despite their characteristics were very similar.

Next, the implemented code is tested using these two satellites. Firstly, the code downloads from Celestrak [1] the satellite TLE data set in a given time before their reentry has taken place. The obtained TLE data set is filtered to remove those atypical data. Now, the program is ready to make an estimation of the ballistic coefficient using the algorithm given in Section 2. Once the ballistic coefficient has been estimated, the integrator uses it to obtain the orbit propagation solving the motion equations with the considered perturbations.

Figure 1 depicts the variations in the semi-major axis for both satellites, Starshine I and Starshine II, clearly illustrating the deorbit of each satellite. Starshine I's propagation began on June 11th, 1999. The propagator predicts its reentry to occur 258.017 days later, whereas the actual reentry took place 258.38 days after the starting date. As for Starshine II, the propagation started on December 21st, 2001. The code anticipated its reentry after 130.84 days, while the actual event happened 130.93 days later. These close estimations suggest that the code is functioning effectively and is capable of accurately predicting a satellite's deorbit timing.

In addition, it can also be seen in Figure 1 that the lifetime of both satellites is quite different although they have very similar characteristics (small spherical satellites with the same size and weight). These differences are due to the

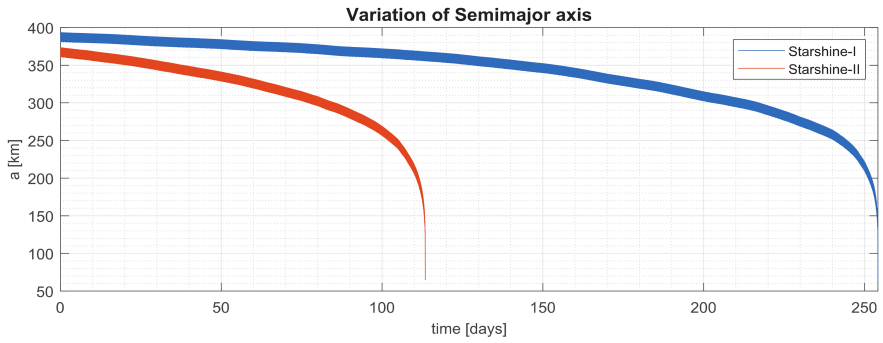


Fig. 1: Semi-major axis propagation for Starshine-I and Starshine-II

different atmospheric conditions the satellite suffered, since both satellite were launched at different epochs. As it can be seen in Figure 2, the solar activity was greater for the Starshine II epoch. This activity causes a greater drag coefficient for this satellite, as it can be seen in Figure 3.

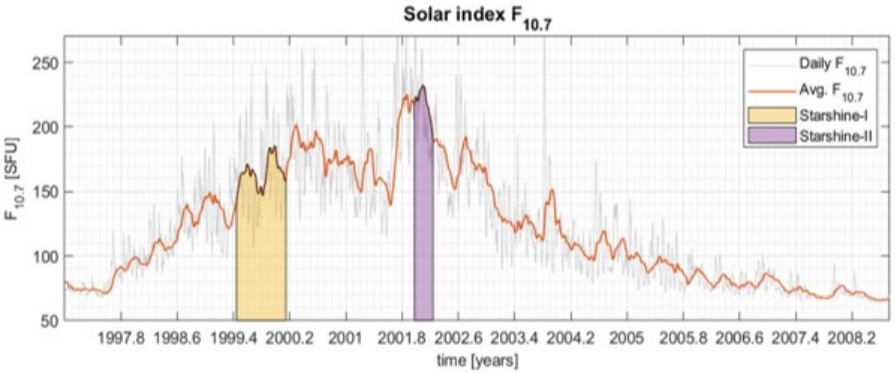


Fig. 2: Solar activity index $F_{10.7}$ for the orbit epoch of Starshine-I and Starshine-II

4 Conclusions

A propagation model has been developed to estimate the lifetime of LEO satellites and predict their orbital trajectories. This model includes the main perturbating effects acting on this kind of satellites, that is, Earth's oblateness, atmospheric drag, and solar and lunar gravity perturbations. The model has been validated to make propagation and prediction studies of LEO orbits.

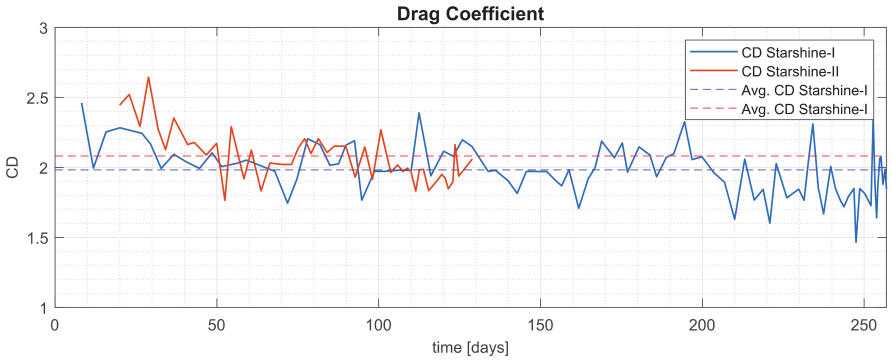


Fig. 3: Drag coefficient estimation C_D for Starshine-I and Starshine-II

Solar activity has been found to be the most influential factor in the estimation of atmospheric density, and consequently affects the estimation of the satellite's lifetime. So, the JB08 atmospheric model, which includes most of the solar activity indices, has been used until the epoch where these indices are available. Later the HP atmospheric model has been implemented for the final evolution epoch.

The estimation of the ballistic coefficient is also an important feature of the code. It has been done by means of an iterative process based on the Secant method. The validity of the iterative estimation method has been verified, since the orbital propagation obtained coincides with the observed data.

Acknowledgments

This work has been supported by the Spanish Ministerio de Ciencia, Innovación y Universidades, Project PID2019-109753GB-C22/AEI/10.13039/ 501100011033, and by the European Union NextGenerationEU/PRTR, Programa de Planes Complementarios I+D+I, ref. ASFAE/2022/014

References

1. North American Aerospace Defense Command (NORAD) website. https://celestrak.org/NORAD/elements.
2. Curtis, H.D.: Orbital Mechanics for Engineering Students. 3 rd Edition, Elsevier (2010).
3. Vallado, D.A., McClain, W.D.: Fundamentals of Astrodynamics and Applications, 4 th Edition. Space Technology Library, 103: 636?648 (2013).
4. Schaub, H., Junkins, J.L.: Analytical Mechanics of Aerospace Systems, 4th Edition. AIAA Education 335 Series (2002).
5. Bowman, B.R., Tobiska, W.K., Marcos, F.A.: A New empirical Thermospheric Density Model JB2008 Using New Solar and Geomagnetic Indices. AIAA 2008-6438 Astrodynamics Specialist Conference, 342 (2008).

6. Picone, J.M., Emmert, J.T., Lean, J.L.: Thermospheric densities derived from spacecraft orbits: Accurate procession of two-line elements sets. Jorunal of Geophysical Research, 110, A03302 (2005). doi:34810.1029/2004JA010585
7. Saunders, A., Swinerd, G.G., Lewis, H.G.: Deriving Accurate Satellite Ballistic Coefficient from 350 Two-Line Element Data. Journal of Spacecraft and Rockets (2012). doi:10.2514/1.A32023
8. ESA Earth Observation Portal. Starshine mission database. https://directory.eoportal.org/web/eoportal/satellite-missions/s/starshine

A new point of view about a biparametric family of anomalies in the elliptic motion

José Antonio López Ortí[1], Francisco José Marco Castillo[1], and María José Martínez Usó[2]

[1] University Jaume I of Castellón, Departamento de matemáticas, E-12071 Castellón, Spain,
lopez@mat.uji.es,
WWW home page: https://www.uji.es/departaments/mat/base/estructura/personal?p_departamento=92&p_profesor=65082
[2] Universitat Politècnica de València, Departamento de matemática aplicada, E-46022 Valencia, Spain

Abstract. One of the main problems in celestial mechanics is the well-known two-body problem. This problem can be studied by analytical and numerical methods. In the elliptical case, the performance of the numerical method is good if the eccentricity value is not high. The main problem, in this case, is the distribution of the points over the orbit. Given a fixed temporal step size, most points are located in the apoapsis region, while few are in the periapsis region. This paper tries to define a new temporal variable so that the point distribution is higher over the periapsis region than on the apoapsis region. It is also desirable to obtain a higher concentration of points with greater curvature. To this aim, we intend to study a modification of the biparametric family of anomalies.

Keywords: Celestial mechanics. Two-body problem. Orbital motion

1 Introduction

One of the main problems in celestial mechanics is studying the two-body problem.
The relative motion of the secondary with respect to the primary is defined by the second-order differential equations:

$$\frac{d^2\vec{r}}{dt^2} = -\mu\frac{\vec{r}}{r^3}$$

where $\vec{r}$ is the vector radius of the secondary and $\mu = G(m_1+m_2)$ where G is the gravitational constant and m_1, m_2 the masses of the primary and the secondary respectively.

This problem is also appropriate for testing numerical methods because we can compare the numerical and the analytical solution. The analytical solution is described by the set III of elements of Brouwer and Clemence $(a, e, i, \Omega, \omega, M)$ [1].

The performance of the numerical methods is good. However, in the case of very eccentric orbits, the point distribution when the natural time, or mean anomaly, is used is not according to the dynamics.

To solve this problem, there are several ways. In this paper, we use the analytical regularization of stepsize to reparametrize the orbit to get a more appropriate point distribution on the orbit. This method can be combined with variable stepsize integrators, symplectic integrators, and other techniques. The reparametrization technique is a very interesting method to solve several problems. An interesting review of these methods can be seen in [4].

In 1912, Sundman [16] introduced a change of temporal variable using the transformation $r\,d\tau = dt$, a temporal reparametrization of the motion known as the Sundman transformation. In this transformation, we can change the time by the mean anomaly multiplying by the mean motion n and including in τ a normalization factor $\frac{1}{a}$ the Sundman transformation can be rewritten as $dg = \frac{r}{a}dM$ where g is the eccentric anomaly.

This method has been used by several authors obtaining a set of time reparametrizations Ψ in the form $dM = Q(r)d\Psi$ where $Q(r)$ is called partition function and $\Psi = \Psi(M)$ is a 2π periodic function of M which satisfies $\Psi(0) = 0, \Psi(\pi) = \pi$. In this sense, Nacozy introduced a new parameter τ related by $r^{\frac{3}{2}}d\tau = dt$ [15], Janin and Bond extended this transformation to Ψ_α defined by $r^\alpha d\Psi_\alpha = dt$ [9], [10], Brumberg [2] introduced the regularized length of arc s^* by $vds^* = dt$, where v is the velocity of the secondary, Brumberg and Fukushima introduced the elliptic anomaly ω as $\omega = \frac{\pi u}{2K(e)} - \frac{\pi}{2}$, where $\operatorname{am} u = g + \frac{\pi}{2}$ [3].

All these variables can be reduced to anomalies, including a normalization scale factor so that they take values in $[0, 2\pi]$ along one revolution, The classic mean anomaly M, the eccentric anomaly g, and the true anomaly f can be considered temporal variables. López [14] defines the semifocalanomaly Ψ as the mean between f and f', where f' is the antifocal anomaly [8].

All of these anomalies are included in the biparametric family of anomalies $\Psi_{\alpha,\beta}$ defined by López [13] as

$$K_{\alpha,\beta}r^\alpha r'^\beta d\Psi_{\alpha,\beta} = dM \tag{1}$$

This family contains the anomalies defined by the previous time transformations, for $\alpha = \beta = 0$ we have the mean anomaly M, for $\alpha = 1$, $\beta = 0$, the eccentric anomaly g, for$\alpha = 1/2$, $\beta = -1/2$ the generalized length of arc introduced by Brumberg, for $\alpha = 3/2$, $\beta = 1/2$ the elliptic anomaly ω for $\alpha = 1$, $\beta = 1$ the antifocal anomaly, and for for $\alpha = 2$, $\beta = 1$ the semifocal anomaly.

Notice that the anomaly $\Psi_{\alpha,\beta}$ is symmetric with respect to the axis of the ellipse when $\alpha - \beta = 1$.

There are some interesting anomalies not included in this biparametric family, [7], [5], [11], [12]

This paper is focused on studying a new geometric point of view about the biparamétric family. This section presents the background and the primary goal of this paper. In Section 2, we study the relation between the curvature and the vector radii $r\ r'$, and from them, the biparametric family of anomalies is related

through the vector radius r and the curvature. In section 4, the main conclusions of this study will be discussed.

2 The biparametric family of anomalies as function of vector radius and the curvature.

In this section, we consider the curvature of an elliptical orbit as a function of the eccentric anomaly g. For this anomaly, it is well known that:

$$\xi = a(cosg - e), \quad \eta = a\sqrt{1-e^2}\sin g \tag{2}$$

where (ξ, η) are the orbital coordinates referred to the primary focus placed in the point F, $\overrightarrow{F\xi}$ running to the periapsis region and $\overrightarrow{F\eta}$ making a direct orthogonal system with $\overrightarrow{F\xi}$ the motion of the secondary whit respect to the primary, direct in the system (F, ξ, η). On the other hand, the vector radius r of the secondary with respect to primary and the vector radius r' with secondary focus of the ellipse F' are given by:

$$r = a(1 - e\cos g), \quad r' = a(1 + e\cos g). \tag{3}$$

The curvature $\kappa(t)$ of a planar parametric curve $\vec{r}(t) = (x(t), y(t))$ is given by [6]

$$\kappa(t) = \frac{\|\vec{r}'(t) \times \vec{r}''(t)\|}{\|\vec{r}'(t)\|^3} \tag{4}$$

Applying this formula to the equation of the ellipse parametrized by the eccentric anomaly $\vec{r} = (a(1 - e\cos g), a\sqrt{1-e^2}\sin g)$ we obtain:

$$\kappa(g) = \frac{a\sqrt{1-e^2}}{(\sqrt{1 - e^2\cos^2 g})^3}. \tag{5}$$

Taking into account that the minor semiaxis of the ellipse can be represented by $b = a\sqrt{1-e^2}$, we can represent the curvature in a more symmetric form as:

$$\kappa(g) = \frac{a\,b}{(\sqrt{a^2\sin^2 g + b^2\cos^2 g})^3} \tag{6}$$

. On the other hand, we have that the product of vector radii r and r' is given by:

$$r\,r' = a^2(1 - e^2\cos^2 g), \tag{7}$$

and comparing with (5) we obtain:

$$(r\,r')^2\kappa(g)^3 = a^{10}(1 - e^2), \tag{8}$$

and so (r r') can be represented using the curvature as

$$r\,r' = a\,a^{1/3}\sqrt[3]{1-e^2}\kappa(g)^{-2/3}. \tag{9}$$

For this reason the biparametric family of anomalies $\Psi_{\alpha,\beta}$ can be rewritten as $\Phi_{\gamma,\delta}$ where:

$$C_{\gamma,\delta}r^\gamma\kappa(g)^\delta d\Psi_{\gamma,\delta} = dM \tag{10}$$

where $\gamma = \alpha - \beta$, $\delta = -\frac{2}{3}\beta$, and $C_{\gamma,\delta} = K_{\gamma,\delta}a^{1+1/3}\sqrt[3]{1-e^2}$

3 Concluding remarks

This paper provides a new point of view on the biparametric family of anomalies.

This family can be rewritten int the form $\Phi_{\gamma,\delta}$ where $C_{\gamma,\delta} r^{\gamma} \kappa(g)^{\delta} d\Psi_{\gamma,\delta} = dM$. These anomalies depend on two factors r^{γ} and $\kappa(g)^{\delta}$, which allows a simple interpretation of the reparametrization. Taking n points on the ellipse with anomaly $\Phi_{\gamma,\delta} = k\,h$ $h = 2\pi/n$, $k = 1, \ldots, n$ the mean of the first factor is a displacement of the points from the region of apoapsis to the periapsis one. The second factor $\kappa(g)^{\delta}$ represents a symmetric displacement of the region from the semimajor to the minor axis, which provides an easy interpretation of the family that was not available until today.

Acknowledgments

This work has been partialy suported by grant UJI-B2021-15 of University Jaume I of Castellón.

Conflict of interest The authors declare that they have no conflict of interest with regard to this work.

References

1. D. Brouwer, G.M. Clemence, (1961). Methods of Celestial Mechanics. New York. Academic Press.
2. Brumberg, E.V. (1992). Length of arc as independent argument for highly eccentric orbits, *Celestial Mechanics.* **53**, 323–328.
3. E.V. Brumberg and T. Fufkushima. (1994). Expansions of elliptic motion based on elliptic functions theory, *Celestial Mechanics and Dynamical Astronomy* **60**, 69–89.
4. Cariñena J F, Martínez E and Muñoz-Lecanda M C. (2022). Infinitesimal time reparametrisation and its applications *Journal of Nonlinear Mathematical Physics* **29** 523–55
5. R. Cid, S. Ferrer, A. Elipe (1983). Regularization and linearization of the equations of motion in central force-fields, *Celestial Mechanics* **31** 73–80
6. P. do Carmo , (2017) Differential Geometry of Curves and Surfaces (second edition). Dover Publications Inc.
7. J.M. Ferrándiz, S. Ferrer and M.L. Sein-Echaluce (1987). Generalized elliptic anomalies, *Celestial Mechanics.* **40**, 315–328.
8. T. Fukushima. (2004). Efficient Orbit Integration by the Orbital Longitude Method Using Antifocal Anomaly.*The Astrononical Journal* **128**, 1455–1464.
9. Janin, G. (1974) Accurate Computation of Highly Eccentric Satellite Orbits, *Celestial Mechanics.* **10**, 451–467.
10. Janin, G., Bond, V. R. (1980) The elliptic anomaly, *NASA Technical Memorandum 58228.*
11. J.A. López, F.J. Marco, M. J. Martínez. (2014). A Study about the Integration of the Elliptical Orbital Motion Based on a Special One-Parametric Family of Anomalies. *Abstract and Applied Analysis. Article ID 162060*, 1–11.

12. J.A. López, F.J. Marco, M. J. Martínez. (2016). Geometrical definition of a continuous family of time transformations generalizing and including the classic anomalies of the elliptic two-body problem. *Journal of Computational and Applied Mathematics* **318**, pág 479–490.
13. J.A. López, V. Agost, M. Barreda. (2017) A new bi-parametric family of temporal transformations to improve the integration algorithms in the study of the orbital motion. *Journal of Computational and Applied Mathematics* **309**, 482–492.
14. J.A. López, V. Agost, M. Barreda. (2023). CMMSE: Study of a new symmetric anomaly in the elliptic, hyperbolic, and parabolic Keplerian motion *Mathematical Methods in the Applied Sciences* **46** 1429-1442
15. P.Nacozy. (1977). The intermediate anomaly.*Celestial Mechanics* **16**, 309–313.
16. K. Sundman, (1912) Memoire sur le probleme des trois corps, *Acta Mathematica* **36**, 105–179

Dynamical Systems and Entropy: state of research

Joan C. Micó[1]

Institut Universitari de Matemàtica Multidisciplinar,
Universitat Politècnica de Valéncia
Camí de Vera s/n, ciutat de Valéncia, Valéncia, Spain.
jmico@mat.upv.es

1 Introduction

The objective of the article is to introduce Entropy and Temperature in the context of abstract dynamical systems. The starting hypotheses to reach the objective are the similar patterns of: (1) the Internal Energy, provided by the second law of classical reversible Thermodynamics, and (2) Dirac's Hamiltonian provided by dynamical systems written as coupled systems of first order differential equations [1].

Note that this objective takes part of an old epistemological program of science: the unification of Dynamics and Thermodynamics, whose most decided research was due to I. Prigogine [2]. The modern attempts to reach that unification lie in the context of the Newtonian Hamiltonian systems (with second order Hamiltonian functions in moments) of Irreversible Thermodynamics, through Liouville, Boltzmann and Fökker-Planck equations, among others [3]. Other attempts are more related with the approach here presented, but with other starting points, such as Sinergetics by H. Haken [4].

The approach here followed was provided by first time in the last edition of this congress [5]. The objective was to include Entropy in the mathematical expression of Dirac's Hamiltonian. However, a better inspection of the followed method provides that it fails in [5]. Therefore, this failure must be here explained. This is the content of Section 2. In Section 3 the new hypotheses to include Entropy (and also Temperature) in Dirac's Hamiltonian are provided by stating three postulates and their consequences. Section 4 is devoted to demonstrate that the Internal Energy of the classical reversible Thermodynamics holds the formalism as a particular case. Section 5 presents the state of the research in the present. Section 6 is devoted to the paper conclusions.

2 Dirac's Hamiltonian and failure of the first approach

The objective of the article is to introduce Entropy and Temperature in the context of an abstract dynamical system and to present the state of research of this objective. Concretely, let q_k be, $k = 1, 2, ..., n$, the abstract variables of a dynamical system:

$$\dot{q_k} = f_k\left(t, \mathbf{q}\right) \tag{1}$$

Note that in (1) $\mathbf{q} = (q_1, q_2, ..., q_n)$. Then, Dirac's Hamiltonian corresponding to (1) is [1, 5]:

$$H\left(t, \mathbf{q}, \mathbf{p}\right) = \sum_{j=1}^{n} f_j\left(t, \mathbf{q}\right) \cdot p_j - \sum_{j=1}^{n} f_j\left(t, \mathbf{q}\right) \cdot g_j\left(t, \mathbf{q}\right) + h\left(t, \mathbf{q}\right) \tag{2}$$

Such that:

$$\sum_{k=1}^{n} F_{lk} \cdot f_k\left(t, \mathbf{q}\right) = -\frac{\partial g_l\left(t, \mathbf{q}\right)}{\partial t} - \frac{\partial h\left(t, \mathbf{q}\right)}{\partial q_l} \tag{3}$$

Where in (2):

$$F_{lk}\left(t, \mathbf{q}\right) = \frac{\partial g_l\left(t, \mathbf{q}\right)}{\partial q_k} - \frac{\partial g_k\left(t, \mathbf{q}\right)}{\partial q_l} \tag{4}$$

Note that the functions $g_j\left(t, \mathbf{q}\right)$ and $h\left(t, \mathbf{q}\right)$ in (2-4) come from the Lagrangian [5]:

$$L\left(t, \mathbf{q}, \dot{\mathbf{q}}\right) = \sum_{j=1}^{n} g_j\left(t, \mathbf{q}\right) \cdot \dot{q}_j - h\left(t, \mathbf{q}\right) \tag{5}$$

The corresponding canonical equations to Hamiltonian (2) are, for $k = 1, 2, ..., n$ [5]:

$$\dot{q_k} = f_k\left(t, \mathbf{q}\right) \tag{6}$$

$$\dot{p_k} = \dot{g_k} = -\frac{\partial h\left(t, \mathbf{q}\right)}{\partial q_k} + \sum_{j=1}^{n} f_j\left(t, \mathbf{q}\right) \frac{\partial g_j\left(t, \mathbf{q}\right)}{\partial q_k} \tag{7}$$

Two basic hypotheses taht were made in [5] to insert Entropy $S\left(t, \mathbf{q}\right)$ and Temperature $T\left(t, \mathbf{q}\right)$ are explained in the following.

Hypothesis 1 provided a way to find the functions $g_j\left(t, \mathbf{q}\right)$ and $h\left(t, \mathbf{q}\right)$:

$$g_k\left(t, \mathbf{q}\right) = \frac{\partial \chi\left(t, \mathbf{q}\right)}{\partial q_k} \tag{8}$$

$$h\left(t, \mathbf{q}\right) = -\frac{\partial \chi\left(t, \mathbf{q}\right)}{\partial t} \tag{9}$$

where $\chi\left(t, \mathbf{q}\right)$ is a function to be found.

Hypothesis 2 inserted Entropy and Temperature in Hamiltonian (2):

$$\sum_{j=1}^{n} f_j\left(t, \mathbf{q}\right) \cdot g_j\left(t, \mathbf{q}\right) - h\left(t, \mathbf{q}\right) = \sum_{j=1}^{n} f_j\left(t, \mathbf{q}\right) \frac{\partial \chi\left(t, \mathbf{q}\right)}{\partial q_k} - \frac{\partial \chi\left(t, \mathbf{q}\right)}{\partial t} = -T\left(t, \mathbf{q}\right) \cdot S\left(t, \mathbf{q}\right) \tag{10}$$

Then, Hamiltonian (2) becomes:

$$H(t,\mathbf{q},\mathbf{p}) = \sum_{j=1}^{n} f_j(t,\mathbf{q}) \cdot p_j + T(t,\mathbf{q}) \cdot S(t,\mathbf{q}) \tag{11}$$

See [5] for subsequent hypotheses. In fact, Hypothesis 1 already fails because when (8, 9) are substituted in Lagrangian (3):

$$L(t,\mathbf{q},\dot{\mathbf{q}}) = \sum_{j=1}^{n} \frac{\partial \chi(t,\mathbf{q})}{\partial q_j}\dot{q}_j + \frac{\partial \chi(t,\mathbf{q})}{\partial t} = \frac{\mathrm{d}\chi(t,\mathbf{q})}{\mathrm{d}t} \tag{12}$$

Then, from (12) the action $A(t) = \int_{t_1}^{t_2} L(t,\mathbf{q},\dot{\mathbf{q}})\,dt$ coincides with the function $\chi(t,\mathbf{q})$, and it is easy to demonstrate that the Hamilton-Jacobi equation becomes $0 = 0$.

Therefore, a new approach with new hypotheses must be stated in the following section.

3 Inclusion of Entropy in Dirac's Hamiltonian: the new approach

The new hypotheses are stated as 3 postulates. Again, if $S(t,\mathbf{q})$ is Entropy and $T(t,\mathbf{q})$ is Temperature, then Postulate 1 consists in adding an new entropic differential equation to the system (1):

$$\dot{S} = f_{n+1}(t,\mathbf{q},S) \tag{13}$$

Being defined the Temperature as:

$$T(t,\mathbf{q},S) = \frac{\partial H(t,\mathbf{q},S,\mathbf{p})}{\partial S} > 0 \tag{14}$$

See below why Temperature does not depend on $\mathbf{p}$. Note in addition in (13, 14) that a new moment p_{n+1} arises in the formalism, i.e, $\mathbf{p} = (p_1, ..., p_n, p_{n+1})$. Then, calling $\mathbf{x} = (\mathbf{q}, S)$, the new Hamiltonian, consequence of Postulate 1, is:

$$H(t,\mathbf{x},\mathbf{p}) = \sum_{j=1}^{n+1} f_j(t,\mathbf{x}) \cdot p_j - \sum_{j=1}^{n+1} f_j(t,\mathbf{x}) \cdot g_j(t,\mathbf{x}) + h(t,\mathbf{x}) \tag{15}$$

Postulate 2 introduces the restriction:

$$-\sum_{j=1}^{n+1} f_j(t,\mathbf{x}) \cdot g_j(t,\mathbf{x}) + h(t,\mathbf{x}) = 0 \tag{16}$$

in Hamiltonian (15). Therefore, from Postulates 1 and 2, the Hamiltonian becomes:

$$H\left(t,\mathbf{x},\mathbf{p}\right)=\sum_{j=1}^{n+1}f_j\left(t,\mathbf{x}\right)\cdot p_j=\sum_{j=1}^{n}f_j\left(t,\mathbf{q}\right)\cdot p_j+f_{n+1}\left(t,\mathbf{q},S\right)\cdot p_{n+1} \tag{17}$$

As a consequence of applying Postulates 1 and 2 to (4, 7), the corresponding canonical equations to the Hamiltonian (17) become:

$$\dot{x_k}=f_k\left(t,\mathbf{x}\right) \tag{18}$$

$$\dot{p_k}=\dot{g_k}=-\sum_{j=1}^{n+1}\frac{\partial f_j\left(t,\mathbf{x}\right)}{\partial x_k}g_j\left(t,\mathbf{x}\right) \tag{19}$$

Postulates 1 and 2 also provide the time derivative of the Hamiltonian (17) by using its result in [5] and the new canonical equations (18, 19):

$$\frac{\mathrm{d}H\left(t,\mathbf{x}\right)}{\mathrm{d}t}=\frac{\partial H\left(t,\mathbf{x}\right)}{\partial t}=\sum_{j=1}^{n+1}\frac{\partial f_j\left(t,\mathbf{x}\right)}{\partial t}g_j\left(t,\mathbf{x}\right) \tag{20}$$

And, from (14) and the $n+1$ canonical moment equation of (19), Temperature becomes:

$$T\left(t,\mathbf{q},S\right)=\frac{\partial H\left(t,\mathbf{x},\mathbf{p}\right)}{\partial S}=-\dot{p}_{n+1}=-\dot{g}_{n+1}=\frac{\partial f_{n+1}\left(t,\mathbf{q},S\right)}{\partial S}g_{n+1}\left(t,\mathbf{q},S\right)>0 \tag{21}$$

Note, as announced above, that Temperature does not depend on the moments. Moreover, also Postulates 1 and 2 provide the way new to find the functions $g_j\left(t,\mathbf{q}\right)$ and $h\left(t,\mathbf{q}\right)$, by substituting 2 in (2, 4), for $l=1,2,...,n+1$:

$$\frac{\partial g_l\left(t,\mathbf{x}\right)}{\partial t}+\sum_{j=1}^{n+1}\frac{\partial g_l\left(t,\mathbf{x}\right)}{\partial x_j}f_j\left(t,\mathbf{x}\right)=\dot{g}_l=-\sum_{j=1}^{n+1}\frac{\partial f_j\left(t,\mathbf{x}\right)}{\partial x_l}g_j\left(t,\mathbf{x}\right) \tag{22}$$

Postulate 3 provides a generalized Gibbs-Duhem equation:

$$\sum_{j=1}^{n+1}f_j\left(t,\mathbf{x}\right)\cdot dp_j=0 \tag{23}$$

Dividing (23) by dt, applying (19) for the moments and simplifying, (23) becomes:

$$\sum_{l=1}^{n+1}\sum_{j=1}^{n+1}\frac{\partial f_j\left(t,\mathbf{x}\right)}{\partial x_l}g_j\left(t,\mathbf{x}\right)\cdot f_l\left(t,\mathbf{x}\right)=0 \tag{24}$$

As a consequence of (23) in Hamiltonian (17):

$$dH\left(t,\mathbf{x},\mathbf{p}\right)=\sum_{j=1}^{n+1}p_j\cdot df_j\left(t,\mathbf{x}\right) \tag{25}$$

It is easy to demonstrate that dividing in (25) by dt, expanding $\dot{f}_j\left(t,\mathbf{x}\right)$, applying canonical equations (18) and simplifying, Eq.(24) arises again.

Therefore, Eqs.(22, 24) provide $n+2$ equations for the $n+2$ unknown variables, i.e, for $g_j\left(t,\mathbf{x}\right)$, $j=1,2,...,n+1$, and for $f_{n+1}\left(t,\mathbf{x}\right)$.

4 Reversible Thermodynamics as a particular case of the new approach

A crucial question arises: does the classical reversible Thermodynamics fit the formalism presented? The answer is affirmative by identifying the Internal Energy (U) of the reversible Thermodynamics with the Hamiltonian (H) (17).

Let $j=1,..,n$ be, being respectively $(q_j,\ S)$ and $(p_j,\ T)$ the extensive and intensive thermodynamic variables, then:

$$U=H=\sum_{j=1}^{n}\ p_j\cdot q_j+T\cdot S \tag{26}$$

Note in (26) that $f_j\left(t,\mathbf{q}\right)=\left(\pm\right)q_j$ (the sign depends on the physical intensive variable sense), $f_{n+1}\left(t,\mathbf{q},S\right)=S$ and $p_{n+1}=T$. Then, the corresponding canonical equations (18, 19) are, for $k=1,..,n$:

$$\dot{q_k}=\left(\pm\right)q_k;\dot{S}=S \tag{27}$$

$$\dot{p_k}=\dot{g_k}=\mp g_k=\mp p_k;\dot{T}=T \tag{28}$$

Eqs. (27, 28) can provide any dynamic relationship among intensive/extensive variables. In addition, taking into account the above identifications, $f_j\left(t,\mathbf{q}\right)=\left(\pm\right)q_j$, $f_{n+1}\left(t,\mathbf{q},S\right)=S$ and $p_{n+1}=T$, their substitution in (24) and subsequent simplification, the following equation arises:

$$\dot{T}\cdot S+\sum_{j=1}^{n}\dot{p}_j\cdot q_j=0 \tag{29}$$

which, after being multiplied by dt:

$$S\cdot dT+\sum_{j=1}^{n}q_j\cdot dp_j=0 \tag{30}$$

which is the classical Gibbs-Duhem equation of the reversible Thermodynamics. Applying (30) to the Internal Energy-Hamiltonian (26), its differential form arises:

$$dU = dH = T \cdot dS + \sum_{j=1}^{n} p_j \cdot dq_j = 0 \tag{31}$$

In other words, the formalism stated by Postulates 1, 2 and 3 embodies as a particular case the reversible Thermodynamics.

5 Present state of the research

Note that Postulate 3 has been imposed as the (23) restriction (called as the generalized Gibbs-Duhem equation) on the Hamiltonian (17). Jointly Postulates 1 and 2 the classical reversible Thermodynamics arises as a particular case. The question is that the (23) restriction will never hold in general for the Hamiltonian (17). However, the way to force that it holds is finding the unknown entropic equation, $f_{n+1}(t, \mathbf{q}, S)$, as the solution of Eq. (23).

If in (23) the function $f_{n+1}(t, \mathbf{q}, S)$ is written explicitly, the equation that must be held this function arises:

$$\sum_{l=1}^{n} \frac{\partial f_{n+1}}{\partial q_l} f_l + \sum_{l=1}^{n} \frac{\partial f_{n+1}}{\partial S} f_{n+1} = -\frac{1}{g_{n+1}} \sum_{l=1}^{n} \sum_{j=1}^{n} \frac{\partial f_j}{\partial q_l} f_l \cdot g_j \tag{32}$$

In addition, (32) cannot be solved independently. It is coupled with Eqs. (32), which are in the beginning presented again with the explicit writing of $f_{n+1}(t, \mathbf{q}, S)$. On the one hand, for $j = 1, 2, ..., n$:

$$\frac{\partial g_j}{\partial t} + \sum_{l=1}^{n} \frac{\partial g_j}{\partial q_l} f_l + \frac{\partial g_j}{\partial S} f_{n+1} = -\sum_{l=1}^{n} \frac{\partial f_l}{\partial q_l} g_l - \frac{\partial f_{n+1}}{\partial q_j} g_{n+1} \tag{33}$$

And on the other hand, for $j = n + 1$:

$$\frac{\partial g_{n+1}}{\partial t} + \sum_{l=1}^{n} \frac{\partial g_{n+1}}{\partial q_l} f_l + \frac{\partial g_{n+1}}{\partial S} f_{n+1} = -\frac{\partial f_{n+1}}{\partial S} g_{n+1} \tag{34}$$

Note in Eqs. (32-34) that $f_{n+1} = f_{n+1}(t, \mathbf{q}, S)$, $f_j = f_j(t, \mathbf{q})$ for $j = 1, ..., n$, and that $g_j = g_j(t, \mathbf{q}, S)$ for $j = 1, ..., n + 1$ due to in (22) Entropy is also involved in the $n + 1$ equations.

Therefore, the state of the present research is to solve the coupled system of equations (32-34). The solution of these $n + 2$ functions would provide that (24) holds, and then that both (23) and (25) also holds.

6 Conclusions

Note that the attempt to unify Dynamics and Thermodynamics has been a constant intellectual strength along the twentieth century and along the years of the present century, being the proposals by Prigogine the most emphasizing [2],

without trying to underestimate the research of Haken by Sinergetics [4] or those researches coming from the Irreversible Thermodynamics.

However, the approach here provided is radically different because it starts from the Dirac's Hamiltonian [1], which was developed by this physicist to derive his famous equation to find the relativistic version of Quantum Mechanics. Although this problem was applied for the electromagnetic field rather for functions, i.e., for the case of dynamical systems, the use of this formalism has never been used beyond a theoretical rarity topic [6, 7].

Therefore, keeping in mind the failure of Dirac's Hamiltonian fully exploit and, in addition, its nonlinear similarity with the Internal Energy of reversible Thermodynamics (26), this author's work has started to research the possibility to develop a nonlinear Thermodynamics of abstract systems. The fact that the reversible Thermodynamics holds the formalism is a first success that encourages to follow this way.

However, it is also obvious that following that way implies to solve Eqs. (32-34). Although they seem extremely complex to be solved simultaneously, even in particular systems, perhaps some simplifying hypotheses could help to solve them. Therefore, the intellectual strength must go on persevering on this objective.

References

1. Dirac, P. A. M., Lectures on Quantum Dynamics. New York, Belfer Graduate School of Science, Yeshiva University, 1964.
2. Prigogine, I., Non-linear Science and the Laws of Nature *J. Franklin Inst*, Volume(5/6):745–758, 1997.
3. Le Bellac, M., Mortessagne, F., and Batrouni, G. G., Equilibrium and Non-Equlibrium Statistical Thermodynamics. Cambridge, Cambridge University Press, 2004.
4. Haken, H., Synergetics. Berlin Heidelberg New York, Springer-Verag, 2004.
5. Micó, J. C., First Order Hamiltonian Systems. Modelling for Engineering and Human Behaviour 2022, 177-184, Universitat Politècnica de Valéncia, ciutat de Valéncia.
6. Havas, P., The conection between Conservation laws and Invariance Groups: Folklore, Fiction, and Fact, *Acta Physica Austriaca*, 38: 145-167, 1973.
7. Govaerts, J., Hamiltonian Reduction of First Order Actions, *International Journal of Modern Physics A*, 05: 3625-3640, 1990.

A revision of the concept of mass may calm the Hubble-tension

Miguel Portilla

Universitat de Valencia, Departament de Mecánica i Astronomia.

Abstract. The ΛCDM cosmological model cannot explain the discrepancy, currently referred as the Hubble tension, between the value obtained by studying the early CMB, Ho= 67.4 Km/s/Mpc, and the one obtained from the more recent supernovae, in the range 70 $\leq$Ho $\leq$ 76.
Using a recent revision of the concept of gravitational mass*, that introduces a time dependence for it, we can explain the Ho-discrepancy and interpret both, dark matter and dark energy, in terms of the revised gravitational mass:
Dark matter is the gravitational mass acquired by the particles ("first stars") that collapsed to form a galaxy, and dark energy as the gravitational mass acquired by the galaxies, tending to a constant at the present epoch that is interpreted as the cosmological constant. (*Journal of Modern Physics. M.P. 2021).

Keywords: Hubble tension, dark matter, dark energy, gravitational mass

We shall need, first of all, two basic equations, namely the total energy density and the Friedman equation:

$$\rho(a) = \frac{3H_0^2}{8\pi G}(\frac{\Omega_m}{a^3} + f(a)) \tag{1}$$

$$\frac{\dot{a}^2}{a^2} = H_0^2(\frac{\Omega_m}{a^3} + \frac{\Omega_k}{a^2} + f(a)), \tag{2}$$

though we shall assume $\Omega_k = 0$. In the current cosmological model the first and second summands in the energy density represent dark matter and dark energy respectively. The cosmological constant Λ is a case of dark energy, with $f(a) = \Lambda/3H_0^2$.

1 The Hubble tension

The Planck collaboration (Planck 2018. Results. VI. Cosmological parameters) studied the remote CMB radiation using the ΛCDM model that assumes $f(a) = \Lambda/3H_0^2 = \Omega_\Lambda = 1-\Omega_m$. They obtained precise estimations the basic cosmological parameters: $H_0 = 67.4, \Omega_b = 0.0493, \Omega_m = 0.315^7$. However, when the nearest objects were considered unexpected discrepancies ocurred: from a Cepheid-SNIa sample Adam G. Riess (2022) obtained: $H_0 = 73.04^{1.04}$, and from many different samples one obtain Hubble constants in the range (70...76).

2 The equation of the geodesics

It will be convenient to write the equation of the geodesics in the form given by Havas-Goldberg 1962, "Lorentz-invariant equations of motion of point masses in the general theory of relativity":

$$m\frac{du_\mu}{d\tau} = M(\tau)\frac{\partial U}{\partial x^\mu} \tag{3}$$

$$M(\tau) = \frac{m}{g(v,v)^{1/2}}\,, U = \frac{1}{2}g(v,v) = \frac{1}{2}g_{\mu\nu}v^\mu v^\nu \tag{4}$$

$$u^\mu = \frac{dz^\mu}{ds}\,, v^\mu = \frac{dz^\mu}{d\tau} \tag{5}$$

$$ds = \sqrt{g_{\mu\nu}dz^\mu dz^\nu}\,, d\tau = \sqrt{\eta_{\mu\nu}dz^\mu dz^\nu} \tag{6}$$

where ds stands for the true metric and $d\tau$ for an auxiliar Minkowski metric.

2.1 Revision of the concepts of inertial, and gravitational mass.

Recently we have considered the constant "m" as the inertial mass of a particle, and the time dependent function $M(\tau)$ as its gravitational mass (Journal of Modern Physics, 2021). Multiplying by the number density of particles $n = n_0/a^3$ we get the gravitational mass density $\rho(a) = n(a)M(a)$ as a function of the expansion factor.

We shall consider three epochs dominated by three kind of particles: atoms, first stars and galaxies, considered all of them as point particles. We shall denote the initial expansion factor of each epoch by: (a_i^{**}, a_i^*, a_i), and we shall use as independent variables the expansion factor a_i corresponding to the galaxy formation, and the quotients of expansion factors u and v defined as follows: $u = a_i^{**}/a^*{}_i$, $v = a_i^*/a_i$, from which we get $a_i^{**} = a_i uv$, and $a_i^* = a_i v$.

We have obtained $M, \rho, \Omega_b, \Omega*_m, \Omega_m$ and the gravitational mass fraction $f(a)$ as functions of the expansion factor "a " and the initial value a_i, and fractions of initial values u, v defined above:

$$g(a,a_i) = (1+\frac{a_i^3}{a^3})^3 - 6\frac{a_i^3}{a^3}(1+\frac{a_i^3}{a^3}) + 4\frac{a_i^(9/2)}{a^(9/2)}, g(a_i,a_i) = 0 \tag{7}$$

$$g(x) = (1+x^3)^3 - 6x^3(1+x^3) + 4x^{4.5} \tag{8}$$

$$M(u,v,a,a_i) = m(1+\alpha+\frac{\Omega_b^2(u,v,a_i)}{18\Omega_m(u,v,a_i)}\frac{a_i^3}{a^3}g(a,a_i) \tag{9}$$

$$f(u,v,a,a_i) = \frac{\Omega_b^2(u,v,a_i)}{18\Omega_m(u,v,a_i)}\frac{g(a,a_i}{a_i^3} \tag{10}$$

$$\rho(u,v,a,a_i) = \frac{3H_o^2}{8\pi G}(\frac{\Omega_m(u,v,a_i)}{a^3} + f(u,v,a,a_i)) \tag{11}$$

$$\frac{\Omega_m}{\Omega_b}(u,v) = 1 + \frac{1}{18}\frac{g(u)}{u^3} + \frac{g(v)/v^3}{18 + g(u)/u^3} \tag{12}$$

$$\Omega_b(u,v,a_i) = \frac{\frac{\Omega_m}{\Omega_b}(u,v)}{(\frac{\Omega_m}{\Omega_b}(u,v))^2 + \frac{g(1,a_i)}{18a_i^3}} \tag{13}$$

$$\Omega *_m (u,v,a_i) = \Omega_b(u,v,a_i) + \frac{1}{18}\Omega_b(u,v,a_i)\frac{g(u)}{u^3} \tag{14}$$

$$\Omega_m(u,v,a_i) = \Omega *_m (u,v,a_i) + \frac{g(v)}{g(u)}\frac{u^3}{v^3} \tag{15}$$

$$\alpha(u,v,a_i) = \frac{\Omega_{dm}(u,v,ai)}{\Omega_b(u,v,a_i} \quad , \Omega_{dm} = \alpha\Omega_b \tag{16}$$

$$\Omega_m(u,v,a_i) = \Omega_b(u,v,a_i)(1 + \alpha(u,v,a_i)) \tag{17}$$

$$H_0(u,v,a_i) = H_{0\Lambda}\sqrt{\frac{\Omega_\Lambda(1-a_i)}{B(u,v,a_i)}} \tag{18}$$

$$B(u,v,a_i) = \int_{a_i}^{1} f(u,v,a,a_i)da \tag{19}$$

$$B(u,v,a_i) = \frac{\Omega_b(u,v,a_i)^2}{18\Omega_m(u,v,a_i)a_i^3}(1 - \frac{513}{280}a_i + \frac{3}{2}a_i^3 - \frac{8}{7}a_i^{4.5} + \frac{3}{5}a_i^6 - \frac{1}{8}a_i^9) \tag{20}$$

The product $m\alpha(u,v,a_i)$ in equation (7) is the gravitational mass contained in a galaxy, generated during the pregalactic epoch, with $a < a_i$. This is our explanation of the galactic dark matter.

Taking into account the Planck collaboration results (*Astr.Aphys.*2019):

$$(\Omega_b, \Omega_m, \Omega_\Lambda, H_{0\Lambda}) = (0.049, 0.317, 0.683, 67.4) \tag{21}$$

obtained assuming the ΛCDM model, that introduced the cosmological constant (verifying $\Lambda/3H_{0\Lambda}^2 = \Omega_\Lambda$) we get the expansion factor a_i at the beginning of the galactic epoch, and using $a = 1/(1+z)$ we can obtain the corresponding redshift z_i. The observed values for Ω_b and Ω_m must verify $\Omega_m + f(u,v,1,a_i) = 1$, and substituting the gravitational mass fraction one gets the equation

$$\Omega_m + \frac{\Omega_b^2}{18\Omega_m}\frac{g(1,a_i)}{a_i^3} = 1 \tag{22}$$

from which we obtain the expansion factor $a_i = 0.085$ at the epoch of galaxy formation, and using the relation $a = 1/(1+z)$ one gets the corresponding redshift $z_i = 10.76$.

Finally , by using these values we shall give in the next tables the values of $z_i^*, z_i^{**}, \Omega_b, \Omega_m$ corresponding to values of the Hubble constant in the interval $70 - 76$ and for different values of $v = 0.60, 0.56$.

For $v = 0.60$ we have the first stars at redshift $z_i = 18.6$, and free atoms at $z_i = 18.60$

u	Ω_b	Ω_m	Ho	z_i^{**}
0.1854	0.0526146	0.52212	70.0005	104.76
0.2066	0.050606	0.388039	71.0004	93.9073
0.22	0.0478335	0.322069	72.0274	88.1266
0.23	0.0454214	0.282546	73.0876	84.2515
0.2369	0.0436911	0.259751	74.0009	81.7684
0.2433	0.0420811	0.259751	75.0012	79.5912
0.24885	0.0407014	0.227982	76.0007	77.7938

For $v = 0.56$ we have the first stars at redshift $z_i = 20.0084$, and free atoms at $z_i = 18.60$

u	Ω_b	Ω_m	Ho	z_i^{**}
0.18636	0.0526159	0.521887	70.0002	111.73
0.2034	0.0511497	0.414635	71.0176	102.286
0.2146	0.0490603	0.356985	72.0059	96.895
0.22325	0.0471032	0.319853	73.011	93.102
0.2302	0.0454144	0.294503	74.0051	90.261
0.23697	0.0437258	0.273403	75.0011	87.654
0.241301	0.0426415	0.261642	76.0007	86.063

Some references about the concept of mass:

References

1. Einstein A. Relativity. 1920. The special and the general theory. 15th edition 1962. Pg. 65: The ratio of the gravitational to the inertial mass must likewise be the same for all the bodies.
2. Einstein A. Geometry and experience. 1921. (In "Sidelights on Relativity" Dover 1983): The general theory of relativity teaches that the inertia of a given body is greater as there are more ponderable masses in proximity to it.
3. Einstein A. The meaning of relativity. 1950. (In "Concept of mass", chapter 5. Max Jammer 2000. Princeton University Press.) The inertial mass of a particle increases if other masses are piled up in its vicinity.
4. Brans C. H. Phys. Rev. 1962: Contrarily argued the equality of inertial and active gravitational mass.
5. Rosen N. Ann. Phys. 1965: Replied with a variable inertial mass, equal to the passive gravitational mass which increases as the universe expands, and with a constant active gravitational mass.
6. Portilla M. Journal of Modern Physics, 2021. One introduces in cosmology a constant inertial mass, and a time dependent gravitational mass.

Bancroft's GPS navigation solution: relativistic interpretation

Ramón Serrano Montesinos and Juan Antonio Morales-Lladosa

Departament d'Astronomia i Astrofísica, Universitat de València, Spain

Abstract. In the context of Global Navigation Satellite Systems (GNSS), a modern approach is that of Relativistic Positioning Systems (RPS). The purpose of this contribution is to bring the (non-relativistic) theoretical foundations of current GNSSs closer to the RPS approach, by recovering from the RPS coordinate transformation equation one of the classical solutions to the location problem which is still used in current GNSS receivers: Bancroft's closed-form solution (with four emitters).

Keywords: GPS navigation equations, Relativistic Positioning Systems, Bancroft's solution

1 Global Navigation Satellite Systems

There are currently four Global Navigation Satellite Systems (GNSS) in operation: GPS from the US, Galileo from the European Union, Glonass from Russia and BeiDou from China. Each of these systems consists of an earth segment and a constellation of 25 to 30 satellites with atomic clocks on board and which continuously transmit electromagnetic signals enconding information (ranging code, navigation message,...). The system time is referenced to the Coordinated Universal Time (UTC) and positions are expressed in an earth-centered, earth-fixed rotating reference frame (ECEF).

In a GNSS, the basic observable is the pseudorange, which is the apparent distance to a visible satellite as inferred from the travel time of the emitted signal. The emitter (satellite) sends a signal, the user receives it and correlates it with a replica of that signal generated in the receiver to compute the travel time of the signal. This time is multiplied by the speed of light in vacuum, c, to obtain the distance to that satellite at the moment of signal emission. This is, of course, an apparent distance because there are many effects that have not been considered: satellite and receiver clock synchronisation errors, relativistic effects, atmospheric effects and instrumental errors. So if the true geometric distance to a visible satellite is about 20.000 km, my pseudorange measurement includes other *distances* that have to be taken into account.

Apart from the pseudorange, the position of each satellite at signal emission is also computed from the signal received (navigation message).

1.1 GNSS navigation equations

The pseudorange[1] to satellite j (R^j) is the user's geometric distance to that satellite (ρ^j), plus the receiver's clock error (δt) and all other effects[2] including the satellite clock error, relativistic corrections, atmospheric effects, etc. (D^j):

$$R^j \approx \underbrace{\sqrt{(x^j-x)^2+(y^j-y)^2+(z^j-z)^2}}_{\rho^j} + c\delta t + D^j, \qquad j = 1, ..., n,$$

where $\{x, y, z\}$ and $\{x^j, y^j, z^j\}$ are the user's and satellite's spatial coordinates, respectively.

There are four unkowns, the user's spatial coordinates and the user's clock error, so we need at least 4 pseudorange measurements. In practice, measurements from all visible satellites are considered and the solution to such an overdetermined system is found by least-squares adjustment. The equations are linearised around an approximate solution ρ_0^j, which is found by solving the (simplified) non-linear equations directly, using Bancroft's method, for example. Once the approximate location is known, the following linearised system of equations is solved using least-squares:

$$R^j - D^j - \rho_0^{\,j} = \frac{x_0 - x^j}{\rho_0^{\,j}} dx + \frac{y_0 - y^j}{\rho_0^{\,j}} dy + + \frac{z_0 - z^j}{\rho_0^{\,j}} dz + c\delta t, \qquad j = 1, ..., n.$$

2 Relativistic Positioning Systems

GNSSs achieve very high precision correcting for relativistic and for all other effects. We have nothing against them as *technological* objects. But the principles on which they are based are not satisfactory from a *scientific* perspective. In contrast, Relativistic Positioning Systems (RPS) involve a purely relativistic approach to positioning systems in general. This theory was developed more than twenty years ago[3] not only for its application to GNSS, but as a relativistic theory for positioning systems in general that has additional advantages:

- any user can know its location in a 4D-coordinate system
- it aims to describe the user's space-time trajectory (proper acceleration) and the space-time itself (gravimetry, metric).

2.1 Positioning in Minkowski space-time

In Minkowski space-time[4], a relativistic positioning system can be thought of as a set of at least four emitters A ($A = 1, 2, 3, 4$) of world-lines $\gamma_A(\tau^A)$, broadcasting

[1] In this section we follow the notation used in [1].
[2] These effects are modelled and, in principle, can be computed from the navigation message.
[3] For the genesis and perspectives of RPS see Ref. [2].
[4] In this section we follow the notation used in [3].

their respective proper times τ^A by means of electromagnetic signals. In a RPS, the basic observable is the set of four proper times $\{\tau^A\}$ received at an event x by the user. These are the user's emission coordinates. Suppose the four world-lines $\gamma_A(\tau^A)$ are known in an inertial coordinate system $\{x^\alpha, \alpha = 0, 1, 2, 3\}$ with origin O. Now we can refer the different 4-vectors that describe this system to this reference frame (see Fig. 1 in [3]).

To solve the location problem, the 4 equations that we have to solve simply express the light-like character of the 4-vector m_a which is x (the user's position 4-vector) minus γ_a (the position 4-vector of the satellite's worldline):

$$(x - \gamma_A)^2 = 0, \ \forall A, \tag{1}$$

The solution to these equations, mapping the user's emission coordinates to its inertial coordinates, is what we may call the RPS coordinate transformation or the RPS solution. As detailed in [3], these equations can be conveniently solved by referring both the user's position x and three of the emitters, say $\{\gamma_1, \gamma_2, \gamma_3\}$, to the fourth one, γ_4,

$$\begin{aligned} y &= m_4 = x - \gamma_4, \\ e_a &= \gamma_a - \gamma_4, \qquad a = 1, 2, 3, \end{aligned} \tag{2}$$

and separating (1) into a system of three linear equations

$$e_a \cdot y = \Omega_a, \qquad a = 1, 2, 3, \tag{3}$$

where $\Omega_a = \frac{1}{2}(e_a)^2$ are the world-function scalars, and a quadratic equation

$$y^2 = 0. \tag{4}$$

The general solution to the underdetermined system (3) is of the form:

$$y = y_* + \lambda\chi, \tag{5}$$

where y_* is the particular solution to the system, λ is a real parameter and

$$\chi = *(e_1 \wedge e_2 \wedge e_3) \tag{6}$$

the configuration vector of the RPS. The particular solution is found by bringing in an arbitrary vector ξ such that $\xi \cdot \chi \neq 0$:

$$y_* = \frac{1}{\xi \cdot \chi} i(\xi) H, \tag{7}$$

with[5]

$$\begin{aligned} H &= \Omega_1 E^1 + \Omega_2 E^2 + \Omega_3 E^3, \\ E^1 &= *(e_2 \wedge e_3), \\ E^2 &= *(e_3 \wedge e_1), \\ E^3 &= *(e_1 \wedge e_2). \end{aligned} \tag{8}$$

[5] Where $*$ is the Hodge operator and $\wedge$ the wedge or external product.

The general solution to the location problem (1) is obtained by introducing (5) in the main quadratic equation (4) and solving for λ:

$$x = \gamma_4 + y_* + \lambda\chi, \quad \lambda = -\frac{y_*^2}{(y_* \cdot \chi) + \hat{\epsilon}\sqrt{\Delta}}, \tag{9}$$

where

$$\Delta = (y_* \cdot \chi)^2 - y_*^2\chi^2, \tag{10}$$

and $\hat{\epsilon} = \pm 1$ is the orientation of the positioning system, which can take two values depending on the configuration of the emitters as seen by the user. We are solving a system of quadratic equations, so in general we expect to obtain two solutions for a given set of emission coordinates. Identifying the correct one is known as the bifurcation problem.

Bifurcation problem Depending on the causal character of the configuration vector χ, we distinguish three situations[6] (see Figs. 3, 4 and 5 in [4]):

- If χ is time-like, there is only one emission solution, the other is a reception solution (the event P' would be on the opposite side of the configuration hyperplane Γ). In this case, the sign of $\hat{\epsilon}$ can be determined.
- If χ is light-like, there is only one valid solution (the other solution is degenerate). The sign of $\hat{\epsilon}$ can be determined.
- If χ is space-like, there are two valid emission solutions: in order to determine the sign of $\hat{\epsilon}$, additional observational information is necessary (relative positions of emitters on the user's celestial sphere).

In practical GPS applications, there is generally no bifurcation problem because the valid solution is always the one that is closest to the earth radius.

3 Bancroft's solution (4 satellites)

We will now talk about Bancroft's classical GPS solution [5], but using the concepts of RPS systems that we have just learned. Bancroft's solution is still used in current GNSS receivers as an initial approximate solution for the least-squares algorithms that are implemented.

Bancroft first defines the user's spatial coordinates, $\vec{x}$, and those of the n satellites, $\vec{s_i}$ (for our purposes, $1 \leq i \leq n = 4$). Then he introduces the pseudorange measurements t_i made by the user with respect to each of the satellites:

$$t_i = d(\vec{x}, \vec{s_i}) + b, \tag{11}$$

where $d(\vec{x}, \vec{s_i})$ is the geometric distance between the ith satellite and the user and b what he calls the user clock's *offset*.

[6] Please refer to [4] for more details.

Reinterpreting (11) as a past light-cone equation (with the user at its vertex), we identify t_i with the time coordinate, with respect to some coordinate system $\{x^\alpha\}$, of the word-line of the ith emitter:

$$t_i \leftrightarrow \gamma_i^t,$$

and the clock's *offset* b, with the inertial time coordinate of the user's position four-vector x:

$$b \leftrightarrow t.$$

Bancroft now defines the four-vectors a_i, which we identify as the emitters' world-lines:

$$a_i = (t_i,\ \vec{s_i}) \leftrightarrow \gamma_i = (\gamma_i^t,\ \vec{\gamma_i}).$$

He introduces a scalar product between four-vectors $\langle a, b\rangle$, which is equivalent to the scalar product in Minkowski space-time, $a \cdot b$, with metric signature $(-,+,+,+)$.

Bancroft's solution vector $y_{1,2} = (-b_{1,2},\ \vec{x}_{1,2})$ can be readily identified with the user's position four-vector x of the RPS. These correspondences are summarized in table 1.

Bancroft		RPS	
pseudorange	t_i	γ_i^t	coordinate time of emitter
data vector	a_i	γ_i	emitter worldline
clock offset	b	t	coordinate time of user
solution vector	$y_{1,2}$	x	user position four-vector

Table 1: Identifying Bancroft's notation and concepts with those of the RPS solution.

Now we are ready to write down the navigation equations solved by Bancroft, using the RPS notation. These are none other than equations (1):

$$(x - \gamma_i)^2 = 0,\ \forall i.$$

In order to solve this system of equations, Bancroft rewrites them with the help of the following scalar:

$$\rho = \frac{1}{2}x^2,$$

row vectors:

$$\mathbb{1} = \Big(1\ 1\ 1\ 1\Big),$$
$$r = \Big(r_1\ r_2\ r_3\ r_4\Big),$$

where $r_i = \frac{1}{2}\gamma_i^2$,

and matrix:

$$A = (\gamma_1\ \gamma_2\ \gamma_3\ \gamma_4),$$

where γ_i are column vectors.

The system (1) reads:

$$\frac{1}{2}x^2 - \gamma_i \cdot x + \frac{1}{2}\gamma_i^2 = 0 \quad \forall i,$$
$$\Leftrightarrow \rho\mathbb{1} - xA + r = 0,$$

or equivalently, provided that A is invertible:

$$x = \rho\, u + v, \tag{12}$$

where

$$u = \mathbb{1}A^{-1},$$
$$v = rA^{-1}.$$

Squaring (12) and substituting $x^2 = 2\rho$:

$$E\rho^2 + 2F\rho + G = 0,$$

where

$$E = u^2,$$
$$F = u \cdot v - 1,$$
$$G = v^2.$$

Equation (12) is Bancroft's solution to the problem, where u and v are known from the emitters' trajectories γ_i and ρ is obtained by solving equation (3). Here we can distinguish the case where $E = 0$ (which, as we will see, corresponds to a light-like configuration of the emitters):

$$E = 0 \Rightarrow \rho = -\frac{G}{2F},$$

and the cases where $E \neq 0$ (which correspond to time-like and space-like emitter configurations):

$$E \neq 0 \Rightarrow \rho = \frac{-F \pm \sqrt{F^2 - EG}}{E}. \tag{13}$$

Bancroft does not make this dinstinction and implicitly assumes $E \neq 0$, his solution being equation (12) with ρ as in (13).

4 Correspondence between Bancroft's and the RPS solution

We can now write both solutions and the correspondence between Bancroft's 4-vectors and scalars and those of the RPS solution (Table 2). At this point we can say that one of the keys to this correspondence is writing the inverse of the matrix A as,

$$A^{-1} = [det(A)]^{-1} \begin{pmatrix} -*(\gamma_2 \wedge \gamma_3 \wedge \gamma_4) \\ *(\gamma_1 \wedge \gamma_3 \wedge \gamma_4) \\ -*(\gamma_1 \wedge \gamma_2 \wedge \gamma_4) \\ *(\gamma_1 \wedge \gamma_2 \wedge \gamma_3) \end{pmatrix}, \quad det(A) = -*(\gamma_1 \wedge \gamma_2 \wedge \gamma_3 \wedge \gamma_4),$$

and choosing $\xi = \gamma_4$.
Then, Brancroft's and the RPS solution are expressed as:

Bancroft's solution:

$$x = v + \frac{-F \pm \sqrt{F^2 - EG}}{E}\, u\ .$$

RPS solution:

$$x = \gamma_4 + y_* - \frac{y_*^2}{(y_* \cdot \chi) + \hat{\epsilon}\sqrt{(y_* \cdot \chi)^2 - y_*^2 \chi^2}}\, \chi\ .$$

Bancroft	RPS ($\xi = \gamma_4$)
u	$[det(A)]^{-1}\ \chi$
v	$y_* - r_4[det(A)]^{-1}\chi + \gamma_4$
E	$[det(A)]^{-2}\ \chi^2$
F	$[det(A)]^{-1}\left(y_* \cdot \chi - r_4[det(A)]^{-1}\ \chi^2\right)$
G	$\left(y_* - r_4[det(A)]^{-1}\ \chi\right)^2$

Table 2: Correspondence between Bancroft's vectors and scalars and those of the RPS solution.

As we have already stated, Bancroft's solution is not valid when $E = 0$, that is, when the configuration vector χ is light-like. In contrast the RPS solution is valid in any configuration.

5 Conclusions

We think that current GNSS systems can benefit from the RPS approach to positioning. Their fundamental principles implicitly use relativistic concepts (such as 4-vectors, Minkowski scalar product, etc.) that can be correctly interpreted according to RPS theory. But RPS is much more, it's a theory for positioning systems in general that aims to describe the user's space-time trajectory dynamically and gravitationally and compute its space-time metric. Ultimately, it could transform relativity into an experimental branch of physics[7].

Acknowledgements

This work has been supported by the Spanish Ministerio de Ciencia, Innovación y Universidades, Projects PID2019-109753GB-C21/AEI/10.13039/501100011033 and PID2019-109753GB-C22/AEI/10.13039/501100011033.

References

1. J. Subirana, J. Zornoza, M. Hernández-Pajares, E. S. Agency, and K. Fletcher, GNSS Data Processing, vol. 1 of ESA TM. ESA Communications, 2013.
2. B. Coll, "Relativistic positioning systems: perspectives and prospects," Acta Futura, vol. 7, pp. 35–47, 2013.
3. B. Coll, J. J. Ferrando, and J. A. Morales-Lladosa, "Positioning systems in Minkowski spacetime: from emission to inertial coordinates," Classical and Quantum Gravity, vol. 27, no. 6, p. 065013, 2010.
4. B. Coll, J. J. Ferrando, and J. A. Morales-Lladosa, "Positioning systems in Minkowski space-time: Bifurcation problem and observational data," Phys. Rev. D, vol. 86, p. 084036, Oct 2012.
5. S. Bancroft, "An algebraic solution of the GPS equations," IEEE Transactions on Aerospace and Electronic Systems, vol. AES-21, no. 1, pp. 56–59, 1985.
6. B. Coll, Epistemic Relativity: An Experimental Approach to Physics, pp. 291–315. Cham: Springer International Publishing, 2019.

[7] Please refer to [6] for an interesting insight.

The Kasner Universe on the Plane

Michael M. Tung

Instituto de Matemática Multidisciplinar,
Universitat Politècnica de València,
Camino de Vera, s/n, 46022 Valencia, Spain,
mtung@imm.upv.es

Abstract. The Kasner spacetime has played a significant role in the initial development of modern cosmology, due to its straightforward yet clever incorporation of spatial anisotropy into modified models for our universe. Using the Cartan formalism, we derive fundamental analytical results for the generalized $(n+1)$D Kasner universe, which exhibit the most prominent features of this spacetime.
The particular $(2+1)$D variant of this spacetime model so far has remained largely unchartered. In this study, we aim to fill this gap by explicitly deriving and investigating several essential aspects of the $(2+1)$D Kasner spacetime geometry. Specifically, our efforts encompass the derivation of fully analytic solutions for both, the timelike geodesic equation, and the scalar wave equation within the $(2+1)$D Kasner background. Furthermore, we present some numerical simulations to highlight the central characteristic properties of the model.

Keywords: classical general relativity, exact solutions, Riemannian geometries, mathematical theory of wave propagation

1 Introduction

In 1921, a mere five years after the inception of the theory of general relativity, the mathematician Edward Kasner (1878–1955) discovered an exact solution to Einstein's vacuum field equations, representing a still homogeneous but anisotropic universe without matter [1–3]. The corresponding spacetime metric of this solution is simple yet highly effective for exploring the notable features of such a universe, providing an elementary model of a spacetime having extreme conditions and which may contract or expand.

Only recently the Kasner cosmological model has regained interest in the context of $f(T)$ gravity, a theory encompassing an arbitrary function of the torsion scalar, offering a novel explanation for the late-time universe acceleration [4].

In this study, our primary focus revolves around the Kasner model with two spatial dimensions, allowing us to investigate its characteristic properties for potential implementations in analogue models of gravity, see e.g. Ref. [5]. For this purpose, the Cartan formalism will help us to efficiently derive interesting analytical results and predictions for the generalized $(n+1)$D Kasner universe.

Until now the $(2+1)$D spacetime variant of the Kasner universe has been mainly overlooked. So a closer inspection is in order. Throughout this work, we explicitly derive and examine some of the most intriguing aspects of this spacetime geometry. This includes obtaining fully analytic solutions for the timelike geodesic equation—geodesics that are followed by massive particles—and solutions for the scalar wave equation—the mathematical framework for describing sound waves in acoustics—all within the $(2+1)$D Kasner background. To conclude the discussion, numerical simulations of the model will help to illustrate the principal properties of this spacetime.

2 Geometrical Framework

As a starting point for a geometrical discussion of the $(n+1)$D Kasner spacetime manifold M, we consider the following nonholonomic basis 1-forms $\theta^\mu \in T^*M$:

$$\theta^0 = c\,dt, \quad \theta^i = \left(\frac{t}{t_0}\right)^{p_i} dx^i, \quad i = 1, \ldots, n, \tag{1}$$

where the constants p_i, $i = 1, \ldots, n$, are the so-called Kasner exponents, entirely defining the anisotropy of the model. Each Kasner exponent, p_i, individually fixes the expansion or contraction rate for the corresponding direction with time evolving. However, we cannot arbitrarily choose these exponents due to physical restrictions as will be demonstrated later on. Note that Greek-letter indices cover the full range of spacetime components, i.e., $\mu = 0, 1, \ldots, n$, whereas the Latin indices only cover the spatial components, i.e., $i = 1, \ldots, n$. Further, we introduce $t_0 > 0$ as an arbitrary scale to render all factors in front of θ^i dimensionless. As usual, c denotes the speed of light.

By construction, this frame $(\theta^0, \theta^1, \ldots, \theta^n)$ will possess local flatness and have orthonormality so that the associated metric in this frame is Minkowskian: $\hat{\boldsymbol{g}} = -\theta^0 \otimes \theta^0 + \theta^1 \otimes \theta^1 + \cdots + \theta^n \otimes \theta^n$. As can be seen, we have chosen a metric with signature $(-, +, \ldots, +)$, so that the time component agrees with $\mu = 0$.

Now, for a torsionless theory, Cartan's structure equations serve to determine the associated curvature 1- and 2-forms in this frame:

$$\begin{aligned} \tau^\mu &= d\theta^\mu + \omega^\mu{}_\nu \wedge \theta^\nu = 0, \\ \Omega^\mu{}_\nu &= d\omega^\mu{}_\nu + \omega^\mu{}_\lambda \wedge \omega^\lambda{}_\nu = \tfrac{1}{2} \hat{R}^\mu{}_{\nu\rho\sigma}\, \theta^\rho \wedge \theta^\sigma, \end{aligned}$$

where d is the exterior derivative, $\omega^\mu{}_\nu$ denotes the rotation 1-forms, and $\hat{R}^\mu{}_{\nu\rho\sigma}$ are the local components of the Riemann curvature tensor.

After a straightforward computation of the Riemann tensor and performing the usual contraction, we obtain the Ricci tensor

$$\hat{\boldsymbol{R}} = \frac{1}{c^2 t^2} \sum_{i=1}^{n} p_i \left[-(p_i - 1)\, \theta^0 \otimes \theta^0 + \left(\sum_{j=1}^{n} p_j - 1 \right) \theta^i \otimes \theta^i \right]. \tag{2}$$

Then, Einstein's field equations in the absence of energy-matter require spacetime to be Ricci-flat, which implies $\hat{\boldsymbol{R}} = 0$. Thus, for the general Kasner model the following physical conditions emerge:

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{n} p_i \quad \text{and} \quad \sum_{i=1}^{n} p_i = 1, \tag{3}$$

corresponding to the time (θ^0) and space components (θ^i) in Eq. (2), respectively. These conditions, Eqs. (3), suggest that the parameter space of allowed Kasner exponents is restricted to the intersection of the unit $(n-1)$-sphere with a hyperplane. Obviously, for $n \geq 3$ the exponents therefore may vary continuously (including the possibility for the p_i's to have positive and negative values). For $n = 2$, however, the *only* possible solutions (given by the intersection of the unit circle $p_1^2 + p_2^2 = 1$ and the line $p_2 = 1 - p_1$) are the discrete parameter pairs $(0, 1)$ and $(1, 0)$, as shown in Fig. 1.

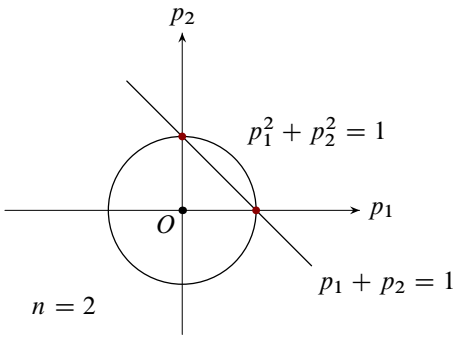


Fig. 1: Graphical representation of the solution space for the Kasner exponents in $n = 2$ spatial dimensions due to conditions Eq. (3).

In the following discussion, for the $(2+1)$D Kasner universe, we will henceforth take the exponents to have the values $p_1 = 0$ and $p_2 = 1$. Therefore, in local Cartesian coordinates, the spacetime metric has the form

$$\boldsymbol{g} = -c^2\, dt \otimes dt + dx \otimes dx + \left(\frac{t}{t_0}\right)^2 dy \otimes dy. \tag{4}$$

For this case, also all components of the Riemann tensor vanish, and $(2+1)$D Kasner spacetime is always flat, although representing a universe with rich features such as being anisotropic and non-static.

3 Particle Dynamics

In this section, we will briefly sketch the dynamics of a massive particle in the $(2+1)$D Kasner universe. All physically viable trajectories are following timelike geodesic paths, which as usual may be found by a variational principle, see Fig. 2.

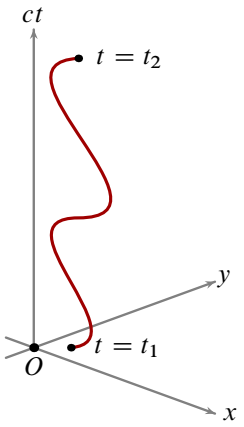


Fig. 2: In the $(2+1)$D Kasner universe a massive particle, connecting spacetime points (events) at $t = t_1$ and $t = t_2$, will travel along a timelike geodesic path. The corresponding Lagrangian, which governs the motion within this spacetime via a variational principle, is given by Eq. (5).

Accordingly, to derive the geodesic equation of a massive test body, one can directly use a free-particle action. So next we take the Kasner metric, Eq. (4), to obtain the following Lagrangian assuming a test mass of $m = 1$:

$$L(t, \dot{x}, \dot{y}) = -c^2 + \dot{x}^2 + \left(\frac{t}{t_0}\right)^2 \dot{y}^2, \tag{5}$$

and where the dot denotes the derivative with respect to the time coordinate t in the local frame. Then, the variational principle—also called the principle of maximal aging—is given by the variation of the action integral with respect to the coordinates x and y, respectively. For $i = 1, 2$ the expression

$$\frac{\delta}{\delta x^i} \int dt\, L = 0$$

yields explicit results for $x = x^1(t)$ and $y = x^2(t)$, given the initial conditions $x_0 = x^1(t_0), y_0 = x^2(t_0), v_x = x^1(t_0)$, and $v_y = x^2(t_0)$.

Eliminating the dependence on t for these solutions, one readily obtains the geodesic curves in the xy-plane

$$y(x) = y_0 + v_y t_0 \left[1 - \frac{1}{\frac{x-x_0}{v_x t_0} + 1}\right]. \tag{6}$$

Some observations are in order: (i) varying the initial position x_0, the shape of the curve does not change but translation invariance occurs in x-direction; (ii) there is damping in y-direction due to the change in scale; (iii) the velocity in y-direction will fall off as $1/t^2$ yielding a maximal possible height of $y_\infty = y_0 + v_y t_0$.

4 Wave Dynamics

For a discussion of scalar wave propagation in curved spacetime, again it will be convenient to depart from a variational principle. However, contrary to the case of a localized point particle, here it will be necessary to consider an action principle over a bounded spacetime domain, $\Omega \subseteq M$, over which the wave will spread, viz. Fig. 3. The general infinitesimal volume element is given by $d\mathrm{vol}_g = \theta^0 \wedge \theta^1 \wedge \ldots \wedge \theta^n$, which for the $(n+1)$D Kasner universe is $d\mathrm{vol}_g = \theta^0 \wedge \theta^1 \wedge \theta^2 = (t/t_0)\, c\, dt \wedge dx \wedge dy$, as produced by direct substitution of Eq. (1) with $p_1 = 0$ and $p_2 = 1$. This clearly indicates that a spacetime volume grows linearly with progressing time.

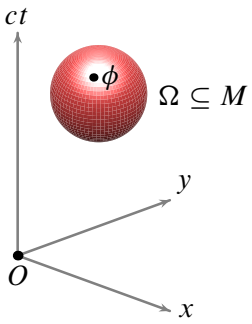


Fig. 3: The variational principle of wave dynamics describes the propagation of scalar waves within a compact and bound region of spacetime, $\Omega \subseteq M$. For the $(2+1)$D Kasner universe, a spacetime volume will expand linearly as time evolves. The scalar wave equation, governing all of its dynamics, derives from Eq. (7) by varying the wave potential ϕ. Typically, the boundary conditions are specified on $\partial\Omega$.

According to the variational principle governing wave dynamics, the following action is stationary with respect to variations of the scalar wave potential, $\phi : M \to \mathbb{R}$, such that [6]:

$$\frac{\delta}{\delta\phi} \int\limits_{\Omega \subseteq M} d\mathrm{vol}_g \, \boldsymbol{g}(\nabla\phi, \nabla\phi) = 0, \tag{7}$$

where $\boldsymbol{g}$ is the metric of the smooth manifold M, and ∇ denotes the associated covariant derivative. In particular, $\nabla\phi$ in Eq. (7) reduces to the partial derivative since ϕ is a scalar quantity. Therefore it is $\nabla\phi = g^{\mu\nu}\, \partial_\mu\phi\, \partial_\nu$, which in acoustics coincides with wave-particle velocity and sound pressure in the local coordinate base $\partial_\mu = \partial_{x^\mu} = \partial/\partial x^\mu$, see also Ref. [6]. The integrand of Eq. (7) is just the quadratic kinetic-energy term generalized to spacetime.

In geometrical language, the Euler-Lagrange equations solving Eq. (7) lead to $*d*d\phi = 0$, where $*$ is the Hodge dual, see e.g. Ref. [7, p. 21]. However,

in local coordinates, this operation just takes the form of the Laplace-Beltrami differential operator on the curved spacetime manifold $(M, \boldsymbol{g})$, denoted by Δ_M, and acting on the wave potential. Substitution of the specific metric, Eq. (4), yields after some simplification the governing wave equation:

$$\Delta_M \phi = -\frac{1}{c^2}\left(\frac{\partial^2\phi}{\partial t^2} + \frac{1}{t}\frac{\partial\phi}{\partial t}\right) + \frac{\partial^2\phi}{\partial x^2} + \left(\frac{t_0}{t}\right)^2 \frac{\partial^2\phi}{\partial y^2} = 0. \tag{8}$$

Eq. (8) can next be tackled by the separation-of-variables technique, making the ansatz $\phi(t, x, y) = \phi_0(t)\phi_1(x)\phi_2(y)$. A reasonable assumption for the component travelling in x-direction (where no scaling occurs) is $\phi_1(x) = e^{ik_x x}$. Then, it can be shown that the propagation in y-direction will necessarily also be harmonic, i.e., $\phi_2(y) = e^{ik_y y}$. Lastly, the final differential equation determining the remaining wave component, $\phi_0(t)$, will depend on the scale parameter t_0, and the wave numbers k_x and k_y. Note that we also adopt natural units where the speed of light is $c = 1$. Furthermore, with plausible initial conditions for prototype waves, the time component $\phi_0(t)$ may still be cast into analytical form in terms of logarithms, trigonometric and Bessel functions.

Figs. 4 and 5 illustrate the simplest scenarios, with prototype waves travelling in x- and y-direction, respectively. Here, the solutions in x-direction, analytically expressed in terms of Bessel functions, clearly display oscillatory behaviour with damping. As expected, in y-direction, where the change of scale occurs, a dramatically different behaviour emerges.

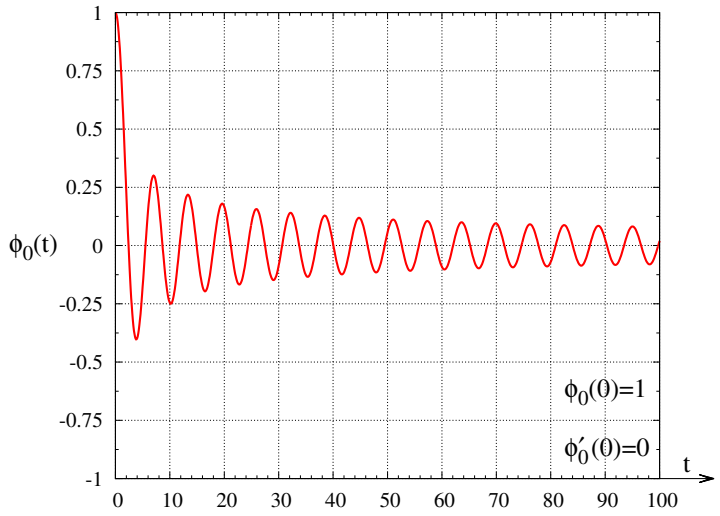


Fig. 4: Time-dependent potential ϕ_0 for the test wave travelling in x-direction $(k_x = 1, k_y = 0)$, with the scale parameter conveniently set $t_0 = 1$. The initial conditions are $\phi_0(0) = 1$ and $\phi_0'(0) = 0$.

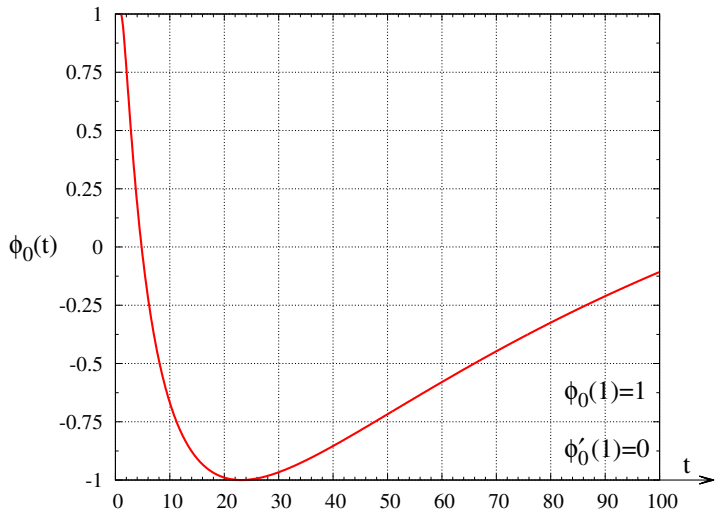


Fig. 5: Time-dependent potential ϕ_0 for the test wave travelling in y-direction ($k_x = 0, k_y = 1$), with the scale parameter conveniently set $t_0 = 1$. The initial conditions are $\phi_0(1) = 1$ and $\phi_0'(1) = 0$.

5 Summary and Outlook

In this presentation we have explored various geometric and physical aspects of (2+1)-dimensional Kasner spacetime. Initially, by employing Cartan's formalism, we first derived general expressions for the Ricci tensor in the suitable $(n+1)$-dimensional nonholonomic frame. Interestingly, the case for $n = 2$ does not exhibit singularities. However, the model of this universe becomes degenerate at $t = 0$ due to the spacetime volume linearly shrinking as it approaches earlier times, i.e., $t \to 0$. Additionally, we have identified only two possible combinations permitted for the Kasner exponents. As a consequence, one of the two spatial directions expands, causing distances parallel to this orientation to grow linearly with time, while the other direction remains unaffected, partially preserving its Euclidean nature.

In the next part we investigated the dynamical properties of the $(2+1)$D Kasner model by analyzing the timelike geodesic path of a massive test particle moving within this spacetime background. By applying a variational principle that maximizes the length between two events in $(2+1)$D Kasner spacetime, we find that fully analytical expressions for the geodesic trajectories may be derived. These curves exhibit characteristic asymptotic barriers in the direction of expansion, indicating that a test particle faces increasing difficulty in catching up with the expanding background. It is important to note that these barriers

are not genuine event horizons of the underlying spacetime since their location depends on the initial position and speed of each test particle.

Finally, we examined $(2+1)$D Kasner spacetime using prototype test waves featuring harmonic behavior propagating in the x-direction. Consequently, the y-dependent part of the potential must also exhibit harmonic behavior. This requirement, where both spatial parts of the potential demonstrate simple harmonic dependence simultaneously, reflects a symmetry property of the underlying spacetime. Accordingly, for the harmonic test case, all non-trivial behaviour lies in the time-dependent potential ϕ_0, which was separated from the total wave potential through the technique of separation of variables. As a result, we are able to derive fully analytical closed-form expressions for ϕ_0 and have provided some numerical examples simulating elementary waves.

In an extension of this study, we plan to consider more sophisticated phenomena of wave propagation, perhaps diagonal or otherwise oblique wave propagation with respect to the predetermined direction of expansion. These results might not be analytically feasible and a more thorough numerical simulation of the models will be necessary.

To conclude, one of the main motivation of this work has been to path the way for an implementation of an acoustic analogue model of gravity for the Kasner universe using metamaterials. Acoustic metamaterials are specially engineered materials that demonstrate unique properties in manipulating sound waves. They are designed to control and manipulate the propagation of acoustic waves—in a way usually not possible with naturally occurring materials, see e.g. Ref. [8]. In this context, the acoustic analogue implementation of the Kasner universe will be a challenging step forward in a series of published studies on the implementation of interesting spacetimes [6, 9–13], which may provide ideal laboratory testing grounds for probing exotic spacetime scenarios.

Acknowledgements

This research was supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21 and PAID-11-22).

References

1. Kasner, E.: Geometrical Theorems on Einstein's Cosmological Equations. Am. J. Math. 43(4), 217–221 (1921). https://doi.org/10.2307/2370192
2. Wainwright, J., Krasiński A.: Republication of: Geometrical theorems on Einstein's cosmological equations (by E. Kasner). Gen. Relat. Gravit. 40(4), 865–876 (2008). https://doi.org/10.1007/s10714-007-0574-4
3. Misner, C., Thorne, K., Wheeler, J.: Gravitation, W.H. Freeman and Company, San Francisco (1973)
4. Skugoreva, M., Toporensky, A.V.: On Kasner solution in Bianchi I $f(T)$ cosmology. Eur. Phys. J. C. 78/377, 1–7 (2018). https://doi.org/10.1140/epjc/s10052-018-5857-2

5. Visser, M., Barceló, C., Liberati, S.: Analogue models of and for gravity. Gen. Relativ. Gravit. 34(10), 1719–1734 (2002). https://doi.org/10.1023/A:1020180409214
6. Tung M.M.: A fundamental Lagrangian approach to transformation acoustics and spherical spacetime cloaking. Europhys. Lett. 98(3), 34002–34006 (2012). https://doi.org/10.1209/0295-5075/98/34002
7. Rosenberg S.: The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds, CUP, Cambridge (1997)
8. Craster, R.V., Guenneau, S.: Acoustic metamaterials: Negative refraction, imaging, lensing and cloaking, Springer, New York (2013)
9. Tung, M.M., Weinmüller, E.B.: Gravitational frequency shifts in transformation acoustics. Europhys. Lett. 101, 54006-–54011 (2013). https://doi.org/10.1209/0295-5075/101/54006
10. Tung, M.M: Modelling acoustics on the Poincaré half-plane. J. Comput. Appl. Math. 337, 336—372 (2018). https://doi.org/10.1016/j.cam.2017.10.037
11. Tung, M.M., Weinmüller, E.B.: Acoustic metamaterial models on the (2 + 1)D Schwarzschild plane. J. Comput. Appl. Math. 346, 162-–170 (2019). https://doi.org/10.1016/j.cam.2018.07.009
12. Tung, M.M: Metamaterial acoustics on the $(2+1)$D Einstein cylinder. Mathematics 9(17), 2079 (2021). https://doi.org/10.3390/math9172079
13. Tung, M.M: Metamaterial acoustics on the Poincaré disk. Math. Meth. Appl. Sci. 45(6), 3229—3238 (2022). https://doi.org/10.1002/mma.6912

Part IV

Mathematical Modelling in Public Health

Modelling the epidemic of oak decline in the Iberian forests

L. Acedo[1], E. Juárez[2] T. Corcobado[34], A. Daxer[3], and A. Solla[2]

[1] Department of Mathematics, Centro Universitario de Plasencia,
Avda. Virgen del Puerto, 2,
University of Extremadura, 10600, Plasencia, Spain.
acedo@unex.es,
WWW home page: https://matematicas.unex.es/ acedo/
[2] Faculty of Forestry, Institute for Dehesa Research (INDEHESA),
Avda. Virgen del Puerto, 2,
University of Extremadura, 10600, Plasencia, Spain.
[3] Austrian Research Centre for Forests BFW,
Seckendorff-Gudent-Weg 8, 1131 Vienna, Austria
[4] Phytophthora Research Centre, Faculty of Forestry and Wood Technology,
Mendel University in Brno, Zemědělská 3, 61300 Brno, Czech Republic

Abstract. Since the 1980s oak decline is severely affecting two main species of trees in Iberia: the holm oaks (Quercus ilex) and cork oaks (Quercus suber). In this work, we discuss a network mathematical model to analyze the propagation of P. cinnamomi in a "dehesa" system up to some thousands of hectares in extension. The model is fitted to epidemiological data obtained from surveys in Extremadura. By using a particle swarm optimization method we show that the preferred scenario corresponds to a very fast propagation of the disease (some months to one year in a hectare) followed by a slow death of the oaks (they survive an average of 56 years to the P. cinnamomi infection). Our model can have important consequences for the management and treatment of this disease as well as in reforestation plannings

Keywords: Forests epidemiology, Phytophthora cinnamomi, SIR model, PSO optimization

1 Introduction

The genus *Phytophthora* groups numerous species of pseudofungi (organisms similar to fungi but not related phylogenetically with them) causing pests in many species of wild plants and crops. The pathogen *Phytophthora infestans* was responsible for an epidemic of potato downy mildew in the 1840s, causing a severe famine. Consequently, there was a large movement of the population, since around one million people emigrated, mainly to the US and another million would perish during this period. In 1875 the founder of plant pathology, H. Anton de Bary, described the genus *Phytophthora* which means *plant destroyer* in Greek.

Specifically, the species *P. cinnamomi* (hereinafter Pc) is considered one of the most dangerous invasive species in the world [1]. Although its presence in the Iberian Peninsula has been known for decades [2], in recent years its activity has intensified in such a way that it threatens the very existence of the "dehesa" as an ecosystem [3,5]. The expansion of the epidemic is favoured by the combination of high temperatures and humidity [3]. To propose a model of propagation and control of Pc it is important to know, even in broad strokes, its transmission mechanism, as well as the effects on the host plant. Pc attacks the roots of its host causing its necrosis. This process can last for decades, although sometimes it is much faster and the trees show signs of decay in a short time, such as chlorosis in their leaves.

As in fungi, Pc develops a structure of very fine filaments that constitute the so-called mycelium. In these ramifications appear the sporangia that contain about 20 or 30 zoospores. These zoospores are biflagellate and can move through water and using chemotaxis they reach the roots of another tree which they colonize.

According to what has been indicated, the humidity in the soil favours the rapid spread of PC as well as the displacement of cattle in the pasture. Cattle transport Pc spores from one tree to another, and water further aids in the dispersal of the pathogen. For this reason, the trees that grow in areas of lower elevation, where puddles form in the rainy season, are the first to get sick and show signs of PC infection. But the dynamics of the "seca" (as the disease caused by Pc is commonly known) can be more complex, because trees already infected, and with rotten roots, suffer more from the effects of drought. unable to absorb nutrients from the soil, are affected by water stress and gradually succumb to the disease.

Several treatments can prevent and stop the disease. The most common is potassium phosphite, but this chemical compound has been forbidden as fertilizer. Various studies [4] propose the use of fosetyl aluminium as an alternative due to its proven preventive and therapeutic efficacy. A research group on agroforestry pathology from the University of Córdoba, Spain demonstrated the usefulness of this treatment for holm oaks and cork oaks [4]

With these basic notions of the phytopathology of Pc, we will propose in the next section (Sec. 2) an epidemiological model for the spread of this pathogen in the "dehesas" of Extremadura. This model can guide specialists in this field in decision-making that could help to combat this plague that is destroying the traditional ecosystem of the "dehesa". In Sec. 3 we will show how to fit the SIR model to real data on the incidence of the oak decline epidemic. This will be achieved by using the particle swarm optimization heuristic method. This work ends with some conclusions and guidelines for future work in Sec. 4.

2 Mathematical modelling of the Pc epidemic.

In this section we will propose the mathematical model to simulate the Pc epidemic in a "dehesa". The initial objective will be to fit this model with the real

data on tree mortality on a given site. We will use the data corresponding to the "dehesa" of Haza de la Concepción, located on the Malpartida de Plasencia road and belonging to the Diputación de Cáceres. This farm has an area of 4 hectares and has been studied for decades. It is known that the holm oaks and cork oaks located there have been sick with Pc and have been dying prematurely from this epidemic. The number of trees (trees still alive) depending on the year is shown in the table 1. To fit these data we will start using a SIR model with

Year	Feet
1956	230
1973	177
1984	140
1997	121
2002	105
2006	102
2010	95
2012	91
2016	79
2019	73
2022	68

Table 1: Census of live trees (feet) in the experimental site of Haza de la Concepción.

three states, as as shown in figure 1. These states would be:

- **Susceptible:** A healthy tree that can contract the "seca" disease if its roots are reached by zoospores.
- **Infected:** Tree whose roots have been colonized by the pathogen Pc and has developed the disease.
- **State R:** In this model, the state R does not correspond to those recovered but to the trees that have died from "seca". We assume that the trees have not received any treatment and that there is no possibility of recovery.

In the epidemic evolution algorithm, we consider that any tree on the farm, that is infected, can infect any other susceptible tree with probability p per month.

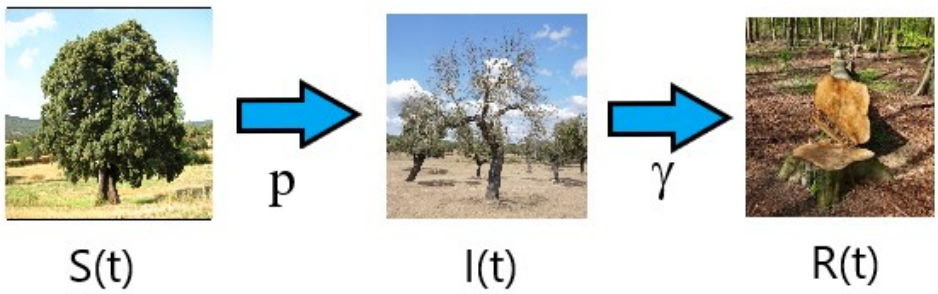


Fig. 1: Diagram of the SIR model for the Pc epidemic in a "dehesa" (here p is the probability that the pathogen is transmitted from a tree infected to a susceptible one while γ is the mortality rate of infected trees).

The infection is spread by water currents at the time of rains or by the movement of cattle and other animals.

In the case of not very extensive farms, such as Haza de la Concepción, it can be assumed that the probability of infection does not depend on the distance between the susceptible and the infected. On farms and extensive areas of several thousand hectares, it is more realistic to consider a function $p(r)$ of the form:

$$p(r) = p_0 \left(\frac{\sigma}{r}\right)^d , \tag{1}$$

where σ has distance dimensions and p_0, d are dimensionless [5]. This would mean carrying out a fit with four parameters, that is, p_0, d, σ and γ. But on the 4-hectare farm that we will study as an example only we will use the model with p and γ as parameters. In any case, to find optimal adjustments for these parameters it is necessary to resort to a metaheuristic algorithm such as the one described below.

2.1 Particle swarm optimization

One of the most widely used algorithms for fitting non-differentiable models is called PSO, particle swarm optimization. This algorithm is attributed Kennedy, Eberhart and Shi [6] and simulates the motion of a swarm of bees or a flock of birds through the space of parameters to find the best solution. In this algorithm, a large number N of possible solutions is constructed. characterized by their coordinates in the parameter space, $\mathbf{x}_i$, $i = 1, \ldots, N$. In our problem, for example, each of these vectors, called particles, would have dimension two. As objective function in model fitting, the root mean square distance of the model predictions to the data is used. In our case, $f : \mathbb{R}^2 \to \mathbb{R}$.

The PSO algorithm is built, then, with the following steps:

- The coordinates of the particles are initialized to values inside a hypercube H with vertices defined by the values of the reasonably possible minimums of the parameters and by the maximum possible values of these parameters. In our case, we would have a square in the parameter plane with vertices $(p_{\min}, \gamma_{\min})$ and $(p_{\max}, \gamma_{\max})$.
- The best known position of the ith-particle is initialized to the initial position, i. e., $\mathbf{b}_i = \mathbf{x}_i$ while if $f(\mathbf{x}_i) < f(\mathbf{g})$, the best global position of the swarm would also become $\mathbf{x}_i$, i.e., $\mathbf{g} = \mathbf{x}_i$.
- The velocity of the particles is initialized to random values within the hypercube defined by the vertices: $\mathbf{x}_{\max} - \mathbf{x}_{\min}$ and $\mathbf{x}_{\min} - \mathbf{x}_{\max}$.
- After initializing the positions and velocities of the particles, the evolution algorithm is carried out until a satisfactory solution is found. In this algorithm, for each particle and in each iteration, the positions and velocities are updated as follows:

$$\mathbf{v}_i \longleftarrow \omega\, \mathbf{v}_i + \phi_b\, r_b\, (\mathbf{b}_i - \mathbf{x}_i) + \phi_g\, r_g\, (\mathbf{g} - \mathbf{x}_i)\ , \tag{2}$$

$$\mathbf{x}_i \longleftarrow \mathbf{x}_i + \mathbf{v}_i\ . \tag{3}$$

  Here ω is the coefficient of inertia that modulates the trend of the particle to continue in the same direction and sense of passage of the current iteration. The cognitive coefficient, ϕ_b is a measure of the particle's tendency to go to its best position, among those found so far. Similarly, the social coefficient, ϕ_g, is the parameter that regulates the tendency of the particles to go towards the best position in the swarm. On the other hand, r_b and r_g are pseudorandom numbers chosen for each particle and each iteration and uniformly distributed between 0 and 1. The second equation simply describes the displacement of the particles in the parameter space according to their velocities. The parameters ω, ϕ_b and ϕ_g are chosen at the beginning of the fitting program. These are typically less than 1, but otherwise, your choice is based on arguments, merely, heuristical.
- Each time the position of a particle is updated, we check if it is a better solution than its best position so far. moment, $\mathbf{b_i}$. This means that the value of the target function has been reduced, i.e.,

$$f(\mathbf{x_i}) < f(\mathbf{b_i})$$

  . In that case, the new position supersedes the best position of that particle:

$$\mathbf{b}_i \longleftarrow \mathbf{x}_i\ ;. \tag{4}$$

- Similarly, we would check if the position $\mathbf{x}_i$ is better than the best global swarm solution found so far, i. e., we check if $f(\mathbf{x}_i) < f(\mathbf{g})$. In that case, we would replace the best global position with the new position of the ith particle:

$$\mathbf{g} \longleftarrow \mathbf{x}_i\ ;. \tag{5}$$

The key to the success of this algorithm is in eq. (2) that simulates the social behavior of the swarm. The new velocity vector of the particle is a composition of three contributions: (i) a fraction ω of the previous velocity vector (ii) a fraction $\phi_g r_g$ of the relative position vector from the current position of the particle to the best position of the swarm (iii) a fraction $\phi_b r_b$ of the relative position vector from the current position to the best position found by that particle.

In the swarm intelligence interpretation, it can be said that ω represents inertia, ϕ_g, cooperation, and ϕ_b, the memory of the best position found by the particle. The balance between the exploration carried out by the particles in the parameter space and the approach to the best solutions is what, intuitively, allows the algorithm to work. In any case, we must bear in mind that in the algorithms metaheuristics, convergence to the best solution is not assured and exploration may stall at a local minimum. To avoid these pathologies, it is advisable to carry out tests with several triads of coefficients ω, ϕ_b and ϕ_g.

The swarm of particles is initially dispersed and uniformly distributed in the solution space. When convergence occurs, some particles stay at or very close to the optimal solution. while others continue to explore.

We have used Fortran code to implement the PSO algorithm applied to the search of the best parameters that fit the SIR model to the data from the table 1. In the next section we will discuss the results of this simulation.

3 Results

To fit the SIR model to the data in the 1 table, we have used an agent-based model to simulate the transmission of Pc from infected trees to healthy trees. In this model, the trees occupy the vertices of a complete graph and starts from an initial state in which all trees except one are susceptible. That is, we start from a configuration with only an infected tree and the rest are susceptible. The model evolution algorithm checks for each link that connects an infected with a susceptible if the disease is transmitted during that month (one month is our time unit). For the infected, it is also verified if the tree dies that month, which in the SIR model occurs with probability γ. In this process, the feet still surviving the epidemic are $S + I$. Evolution takes place a certain number of iterations and is averaged over them to reduce the effect of statistical fluctuations. We have performed 20 iterations for each pair of parameters p, γ. A subroutine is called in the main program to return a prediction of the evolution of the epidemic for a set of given parameters (particle in the language of the PSO algorithm). The parameters of the PSO algorithm that we will use are:

- Number of particles: 100,
- Number of iterations of the PSO algorithm: 50,
- Coefficient of inertia, $\omega = 0.8$,
- Cognitive coefficient, $\phi_b = 0.1$,
- Social coefficient, $\phi_g = 0.1$.

The best fit to the data in the 1 table is shown in figure 2.

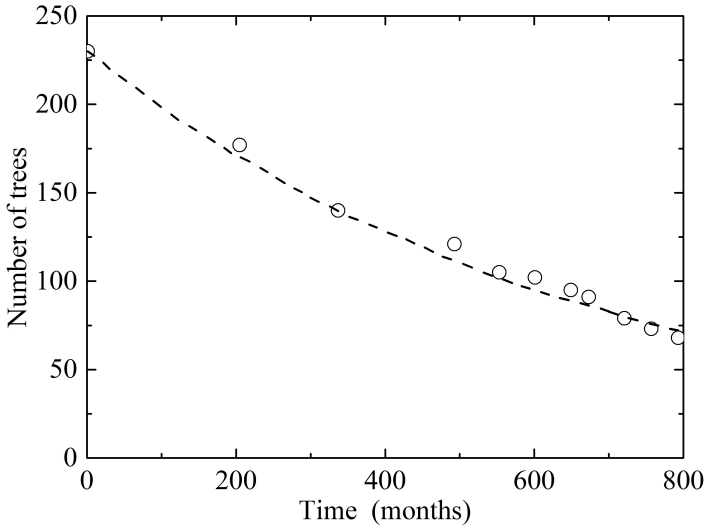


Fig. 2: Data for the number of feet (live trees) and fit to a SIR model. Here time is measured in months since January 1956, the circles are the data and the dashed line is the model prediction.

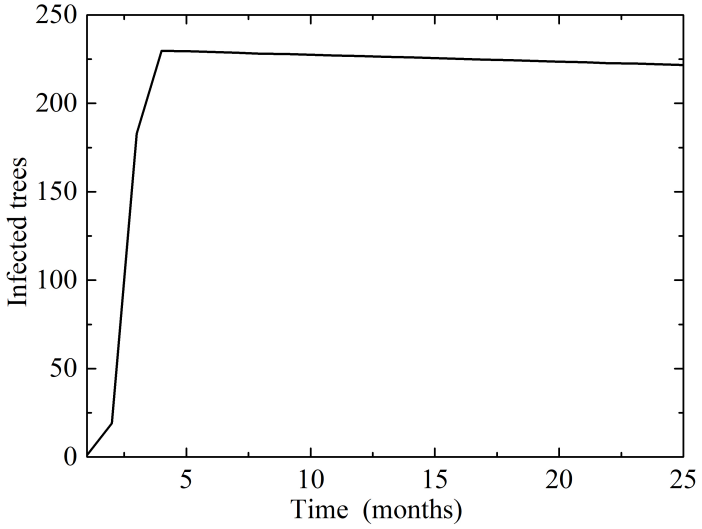


Fig. 3: Prediction for the number of infected trees as a function of time in months.

We note that the fit is very good. The coefficients of the best fit according to the PSO algorithm are:

$$\begin{aligned} p &= 0.078 \text{ per month} , \\ \gamma &= 0.00147 \text{ per month} . \end{aligned} \tag{6}$$

That means that *Phytophthora* spreads very quickly through the pasture, on the scale of the average life of a holm oak or cork oak, but that infected trees have a life expectancy:

$$T = \frac{1}{\gamma} = 56.7 \text{ years} . \tag{7}$$

This life expectancy is much less than that of an oak or cork oak healthy and in top condition.

A prediction of the mathematical model, for which no data exists with which to compare, is the number of infected trees (based on a single infected tree in January 1956 as a hypothesis). In figure 3 we show this prediction. It is observed that the PSO algorithm favours a scenario in which practically all of the trees on the farm end up infected with Pc in only five months from the start of the epidemic to subsequently slowly succumbing to the effects of the pathogen. These results are preliminary and should be validated with data from other pastures, which which will allow us to exclude the possibility of a local minimum. Likewise, it is necessary to consider the initial number of infected trees as an additional adjustment parameter.

4 Discussion and conclusions.

In this work we have proposed an epidemiological mathematical model to analyze the propagation of the pseudofungi *Phytophthora cinnamomi* in a "dehesa". The species *P. cinnamomi* is considered one of the most dangerous invasive species in the world [1] and it, in particular, attacks holm oaks and cork oaks, which are the most common trees in the "dehesas". The corresponding epidemic of oak decline (or "seca" as it is popularly known in Spain) is causing an important mortality of oaks and this is so serious that the "dehesa" ecosystem could disappear in a few decades. This would have, and it is already having, many consequences from the economic point of view, as well as for the ecology of the region of Extremadura, agriculture and cattle raising.

To study the progression of the epidemic we have retrieved data on the number of dead and surviving trees at the 'Haza de la Concepción' site in Extremadura. This site is known for being attacked by the *P. cinnamomi* epidemic for decades and this has caused the loss of a 70.4% of the initial trees back in 1956. To analyze this data we have performed a fit to a SIR model by using PSO optimization. Our results depict a scenario in which the pathogen propagates very fast to all the trees in the "dehesa" in a period of 5 months. Anyway, the oak trees can survive an average of 56.7 years to the "seca" disease caused by this pseudofungi.

To improve this initial mathematical model and apply it to larger "dehesas" we could consider a distance-dependent infection probability that takes into account the displacement of cattle, wild animals and other factors as in Eq. 1. The elevation of the terrain is also a key factor in this epidemic because lower areas are more likely flooded. Therefore, in these areas the Pc zoospores propagate more easily than in areas of higher elevation. Oak trees should be labelled with an elevation parameter and the probability of infection could be made linearly dependent on this parameter. Work along this line is in progress and will be published elsewhere.

References

1. Burgess, T. I. et al.: Current and projected global distribution of Phytophthora cinnamomi, one of the world's worst plant pathogens. In: Global Change Biology vol. 23 (2017).
2. Brasier, C. M., Robredo, F. G., Ferraz, J. F. P.: Evidence for Phytophthora cinnamomi involvement in Iberian oak decline. In: Plant Pathology vol. 42 (1993), pages 140-145.
3. Duque-Lazo, J. et al.: Forecasting oak decline caused by Phytophthora cinnamomi in Andalusia : identification of priority areas for intervention. In: Forest Ecology and Management vol. 417 (2018), pages 122-136.
4. Rodríguez-Navarro, D. N. et al.: Estudios in vitro de la capacidad de biocontrol de hongos y oomicetos fitopatógenos por cepas de Bacillus y Pseudomonas. In: Phytoma, (2019). https://www.phytoma.com/la-revista/phytohemeroteca/311-agosto-septiembre-2019.
5. Cardillo, E., Abad, E., Meyer, S.: Spatio-temporal analysis at landscape scale of the Iberian oak decline epidemic caused by Phytophthora cinnamomi. In: bioRxiv (2020).
6. Yang, X. S.: Nature-Inspired Metaheuristic Algorithms. Luniver Press, (2008).

A probabilistic description of the effect of vaccination in a Bayesian model of COVID-19 transmission dynamics

Javier Blecua[1,2], Juan Fernández-Recio[1], and José Manuel Gutiérrez[2]

[1] ICVV, CSIC-UR-Gobierno de La Rioja, Ctra. Burgos Km 6, 26007 Logroño
[2] Universidad de La Rioja, C. Madre de Dios 53, 26004 Logroño
javier.blecua@icvv.es

Abstract. Bayesian probabilistic models for COVID-19 transmission dynamics have been very efficient to interpret early data from the beginning of the pandemics. Using this approach, we showed that the impact of the non-pharmacological measures was clearly different in each country. We further extended this initial model for the analysis of multiple periods of different transmission rates, enabling the inclusion of an arbitrary number of non-pharmacological measures, and considering the impact of the vaccination and of the different virus variants on the transmission dynamics of the disease (paper currently under preparation).
A key element of the model are the probabilistic curves used to describe the immunity evolution of the population after the vaccination or the infection. The purpose of this paper is the description of the parameters of these curves that describe the degree and duration of the vaccination immunity, and the combinations of values with which the model obtains predictions for the number of daily deaths that best fit the reported data.

Keywords: COVID-19 transmission dynamics, vaccination, immunity.

1 Introduction

As of 23 July 2023, over 768 million confirmed COVID-19 cases and over 6.9 million deaths have been reported worldwide [1]. In Spain, 121,852 persons have already died from the disease [2]. Although the COVID-19 pandemic is officially over, the virus is still active, and we need to have efficient mathematical models of the transmission dynamics to be ready for periodic recurrence, appearance of new virus variants, or any other unwanted future scenarios. From the beginning of the pandemic, different approaches were reported to model the transmission dynamics of the virus in the population. Among them, Bayesian probabilistic models were very efficient to interpret early incomplete data [3].

We applied this approach to analyse the impact of non-pharmacological measures in different countries [4], as these were the only strategies to control the spread of the disease. The initial model was extended throughout the pandemic in order to include key features such as the possibility of defining an unlimited number of non-pharmacological measures, modelling the impact of detection rate

in the transmission, or predicting cases and deaths in hospitals (manuscript in preparation). But towards the end of 2020 two key aspects had a dramatic effect on COVID-19 transmission dynamics: massive vaccination of the population, and the appearance of different virus variants. The challenge was to include in the model the positive impact of the vaccination on the evolution of the disease, taking into account the immunity of the recovered population and considering specific transmission parameters for the different virus variants.

Different studies have estimated the evolution in time of the protection provided by vaccination based on clinical data as well as on the concentration of antibodies and other molecules of the immune system [5–8]. Such studies show important aspects that need to be considered to model the impact of vaccination on transmission dynamics. Basically, after vaccination, the level of protection against infection (immunity) increases until reaching a maximum value, which remains at a certain level during a period of time, and at some point, if no additional doses are applied, starts to decay. In addition, these studies show that immunity provided by vaccination is less effective against the variant omicron.

Here we propose the use of probabilistic curves to estimate the degree and duration of the vaccination impact on the transmission, considering the different vaccine doses. We show the effect of using different vaccination parameters on the general transmission model and the results of applying this model on data from different European countries.

2 Methodology

2.1 Bayesian model for COVID-19 transmission dynamics

We have used a Bayesian probabilistic SIR model in which the transitions between the different states (susceptible, infected, recovered, dead) are described by specific parameters and probability distributions over time [3, 4]. The expected number of new infections I_i occurring in each day i is defined as a function of the number of infected individuals I_j in the previous days, their probability of infecting other individuals after i-j days according to a serial interval (SI) distribution, and the reproduction number (R_t) in the day i (Equation 1).

$$I_i = (R_t)_i \cdot \sum_{j=1}^{i-1} (I_j \cdot SI_{i-j}) \tag{1}$$

The reproduction number (R_t) describes the total number of persons that can be infected by each newly infected individual in average. The value of (R_t) at a given day i is determined by an initial value R_0 and a set of factors α_k that quantify the effect of all non-pharmacological measures that are active on that day i as well as other events/conditions that can impact the virus transmission (Equation 2).

$$(R_t)_i = R_0 \cdot e^{-\sum_k \alpha_k} \tag{2}$$

The Serial Interval (SI) distribution (Figure 1a) describes the probability of when a newly infected individual is more likely to transmit the disease to another person. This distribution was estimated as a Gamma distribution $g \sim Gamma(6.5, 0.62)$ based on data from early epidemics [3].

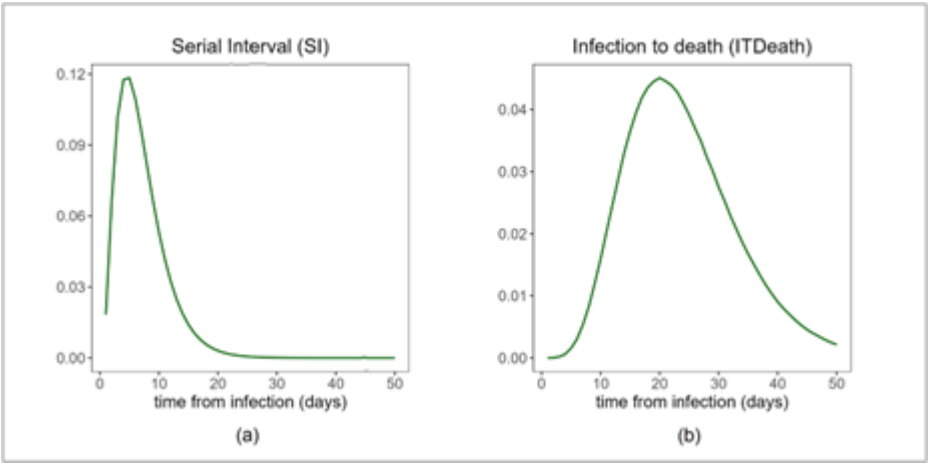


Fig. 1: Serial Interval (a) and Infection to Death (b) probability distributions.

Finally, the expected number of deaths D_i in each day i is a function of the number of infections I_j occurring in the previous days, the estimated infection fatality ratio (IFR) for each country, and the probability of occurrence of death in day i-j after infection according to a previously calculated infection-to-death (ITD) probability distribution (Equation 3).

$$D_i = \sum_{j=1}^{i-1} (I_j \cdot IFR \cdot ITD_{i-j}) \tag{3}$$

The infection fatality ratio (IFR) is the probability of death for an infected case, and it has been calculated for each country from clinical data as previously described [4].

The Infection to Death (ITD) distribution (Figure 1b) describes the probability of when a fatality is more likely to occur after infection. This distribution was estimated as the combination of two gamma distributions, representing the incubation period (infection to onset) and the time between onset of symptoms and death (onset to death), and is given by $\pi \sim Gamma(5.1, 0.86) + Gamma(17.8, 0.45)$ according to data from early epidemics [3].

The above-described model (Equations 1-3) estimates the new infected individuals and deaths for each day as a function of a set of parameters that will be optimized so that the model provides the best possible description of real data. These parameters are: i) the initial infected individuals (I_0) during the

first 6 days of the studied period, which usually starts 30 days before the first 10 deaths; ii) the initial value of the reproduction number (R_0), and iii) the factors α_k describing the impact of the measures and conditions that are applied in each period defined by the user. The values of the parameters are sampled from Bayesian prior distributions by using a Markov Chain Monte Carlo (MCMC) method in order to get the best possible fit of the model to the observed number of daily deaths, which are assumed to follow a negative binomial distribution, as previously described [3, 4].

2.2 Impact of vaccination and virus variants on transmission dynamics

The impact of vaccination in the transmission can be described by a reduction of the reproduction number by the relative amount of the initial population N_0 that is not susceptible of being infected due to the protective effect of the vaccination. The same is true also for the immunity acquired after infection. Both types of immunity contribute to the estimation of the reproduction number with a reduction factor (Equation 4), in which Imm_{i-1} represents the total amount of immune population (calculated on the previous day for practical purposes).

$$(R_t)_i = \ R_0 \cdot e^{-\sum_k \alpha_k} \cdot (1 - \ (Imm_{i-1}/N_0)) \tag{4}$$

The calculation of the amount of immune population (Imm_{i-1}) should take into account important aspects: i) vaccination does not provide immediate immunity; ii) vaccinated and infected individuals are immune against infection during an undetermined period of time; and iii) immunity provided by vaccination depends on the efficiency of vaccine and the virus variant. Thus, the amount of immune population (Imm_i) is calculated here as a function of the reported number of individuals vaccinated on previous days at each dose type ($Vacc1_j$, $Vacc2_j$, $Vacc3_j$), the probability of having acquired immunity in day i-j after vaccination according to a precalculated distribution for each dose type ($V1TImm$, $V2Timm$, $V3TImm$), and the efficiency of the vaccine against each virus variant, according to the reported proportion of variants each day ($IRvar_i$) (Equation 5). The value Imm_i also includes the number of infected people in previous days and the probability of retaining immunity since the time of infection. To avoid double counting in the case of infected individuals that are vaccinated, here we have included only detected cases when this number is larger than the number of vaccinated people.

$$\begin{aligned} Imm_i = \sum_{j=1}^{i-1} (IRvar_j \cdot (Vacc1_j \cdot V1TImm_{i-j} + \ Vacc2_j \cdot V2TImm_{i-j} + \\ Vacc3_j \cdot V3TImm_{i-j}) + \ I_j \cdot ITImm_{i-j}) \end{aligned} \tag{5}$$

The $V1TImm$, $V2ITmm$, $V3TImm$ distributions describe the probability of being immune at a given time after each vaccination dose. These were defined here based on approaches from other studies and on empirical data related to the

effectiveness of the vaccines (more curves could be defined for additional vaccination doses). Several factors define the parameters of the immunity probability distribution curves related to the vaccination:

- Time to reach the highest immunity after each vaccine dose.
- Highest immunity against infection after each vaccine dose.
- Period of highest immunity after each vaccine dose.
- Period of immunity decay from highest to lowest immunity.
- Lowest immunity against infection after immunity decay.

Figure 2 shows some examples of immunity curves for the individuals that have received 2 or 3 doses ($V2TImm$, $V3TImm$) when using specific values for the different parameters in three different scenarios: (a) All vaccine doses are applied on time, before the immunity of the previous doses start to decay; (b) only two vaccine doses (no booster) are applied; and (c) the booster is delayed and is applied when the immunity from the second vaccine dose has decayed to its minimum value. Figure 2 also includes the curve for immunity decay in infected people (d). The periods of increase and decrease of immunity are defined as truncated gaussians, and the periods of maximum and minimum immunity are just constant values.

Additionally, the infection fatality ratio (IFR), initially assumed to be a constant value for each country during the entire studied period, is actually variable in time given that vaccinated individuals that are infected have lower probability of death. In addition, IFR can be different for each virus variant. Thus, IFR will depend on the percentage of vaccinated population and the proportion of infections for the different virus variants on a given day (Equation 6).

$$D_i = \sum_{j=1}^{i-1} (I_j \cdot IFR_j \cdot ITD_{i-j}) \tag{6}$$

3 Results

3.1 Defining the optimal parameters for the immunity probability curves

The model has been applied to available data from a total of 30 European countries using different combinations of the parameters that define the vaccination immunity curves. Initial values for these parameters have been obtained from empirical data documented in different sources and from available information about the effectiveness of the vaccines used in the European countries [9, 10]. From these initial values, the optimal combination of parameters has been found by adjusting some of them to get the best fit results of the model for different countries.

The values of the parameters that provide the best fit for the majority of the European countries are shown in Table 1.

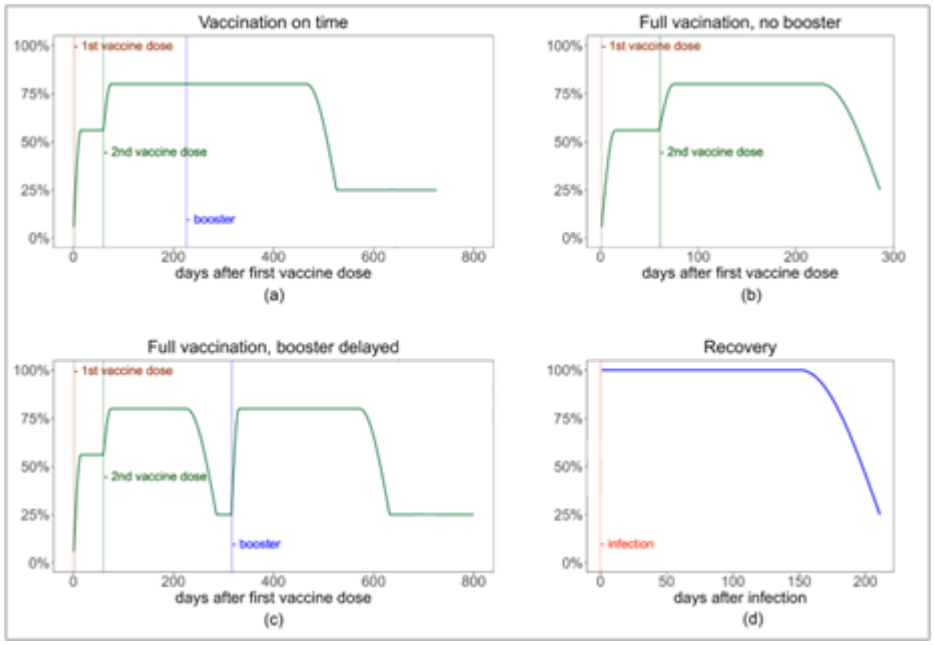


Fig. 2: Evolution of the immunity against infection in three different situations: when all vaccine doses are applied on time (a), when only two vaccine doses (no booster) are applied (b) and when the booster is delayed (c). Curve for immunity decay in infected people ITImm (d).

3.2 Application of the model to COVID-19 transmission dynamics in Spain

The above-described model, including the effect of vaccination and the new variants, has been applied to the available data of COVID-19 pandemics in Spain in the period from January 1st, 2020 to October 31st, 2022. The vaccination immunity model has been adjusted using different combinations of values for the parameters from Table 1. Here we show fit results for two scenarios, defined by variations in three of the parameters, as documented in Table 2.

- Scenario #1: Best fit (parameter values from Table 1).
- Scenario #2: Pessimistic scenario: reduced value for the highest immunity after the full vaccination, reduced period of immunity.

Figure 3 shows the predicted cumulative number of deaths (median with 95% credible interval) in comparison with the real data for the two scenarios. With the values from Table 1 (scenario #1), the predictions from the model replicate well the evolution of the number of reported deaths in the whole period of time. In the other scenario, the predictions deviate from the real data in the first months of 2021, when the massive vaccination of the population started to show its positive impact on reducing the spread of the disease.

Table 1: Parameters of the vaccination immunity curves.

Parameter	Value
Time to reach the highest immunity after each vaccine dose	15 days
Highest immunity against infection after first dose	56%
Highest immunity against infection after second dose	80%
Highest immunity against infection after booster	80%
Period of highest immunity after second dose	150 days
Period of highest immunity after booster	240 days
Period of immunity decay from highest to lowest immunity	60 days
Lowest immunity against infection after immunity decay	25%
Reduction factor of the previous immunity values for infections with the variant omicron	0.5

Table 2: Parameters of the vaccination immunity curves.

Parameter	#1	#2
Highest immunity against infection after second dose	80%	70%
Highest immunity against infection after booster	80%	70%
Period of highest immunity after second dose	150 days	120 days

4 Discussion

There is evidence that there are differences in the effect of the vaccination between seropositive (i.e., recovered from a COVID-19 infection) and seronegative individuals [8], and that this effect is also different depending on the age [11]. These factors are not considered in our model, where we assume common vaccination parameters for the whole population. This is consistent with other simplifications of the model, in which, for example, age weighted values for the infection fatality ratio (IFR) are used.

On the other side, while the initial value of the reproduction number (R_0) and the factors that quantify the impact of the non-pharmacological interventions are optimised during the fit process, the parameters that define the vaccination immunity curves are predefined as constant values in each fit run. Future versions of the model will include these parameters to be optimized during the fit process. The challenge is the limited availability of epidemiological data for the majority of countries, which is increasingly scarce since the COVID-19 pandemics is officially over.

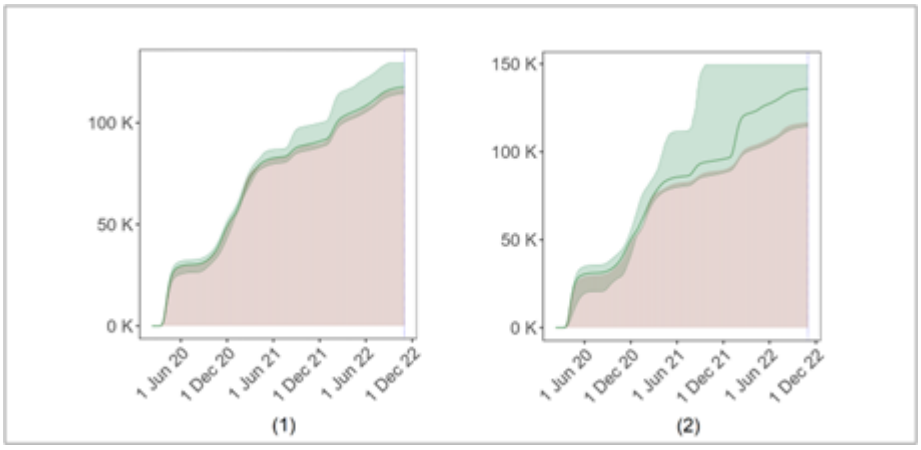


Fig. 3: Data fit for the cumulated number of deaths in Spain in the period from 01/01/2020 to 31/10/2022 with different vaccination parameters: best fit (1) compared to the result of assuming a reduced effect of the vaccination (2).

5 Conclusions

Bayesian inference models are very efficient to model transmission dynamics of infectious agents in complex scenarios, like in COVID-19. These models can be used to estimate the impact of non-pharmacological measures or other events that can affect the virus transmission. The versatility of these models makes it feasible to include different scenarios like vaccination or new variants. We have used a realistic distribution probability curve to describe the acquisition of immunity after vaccination, which provides very good fitting of the model to the available epidemiological data. This tool can be easily adapted to future situations and be ready for the appearance of new variants in which vaccination may have different effect.

References

1. World Health Organization. Coronavirus Disease 2019 (COVID-19) Situation Report - 106 (WHO, 2023). Available online: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20230727_weekly_epi_update_153.pdf?sfvrsn=de3d47d_3(accessedon29July2023)
2. World Health Organization. WHO Coronavirus (COVID-19) Dashboard (WHO, 2023). Available online: https://covid19.who.int/WHO-COVID-19-global-data.csv (accessed on 29 July 2023)
3. Flaxman, S., Mishra, S., Gandy, A. et al. (2020) Estimating the effects of nonpharmaceutical interventions on COVID-19 in Europe. 2020 Nature 584, pp. 257-261. https://doi.org/10.1038/s41586-020-2405-7
4. Fernández-Recio, J. (2020) Modelling the Evolution of COVID-19 in High-Incidence European Countries and Regions: Estimated Number of Infections and Impact of

Past and Future Intervention Measures. 2020 J. Clin. Med. 9, no. 6, 1825. https://doi.org/10.3390/jcm9061825
5. Pérez-Alós, L., Armenteros, J.J.A., Madsen, J.R. et al. (2022) Modeling of waning immunity after SARS-CoV-2 vaccination and influencing factors. Nat Commun 13, 1614 (2022). https://doi.org/10.1038/s41467-022-29225-4
6. Kristen W. Cohen. et al. (2021) Longitudinal analysis shows durable and broad immune memory after SARS-CoV-2 infection with persisting antibody responses and memory B and T cells. Cell Reports Medicine, Volume 2, Issue 7, 2021, https://doi.org/10.1016/j.xcrm.2021.100354
7. Rishi R. Goel et al. (2021) mRNA vaccines induce durable immune memory to SARS-CoV-2 and variants of concern. Science 374, abm0829 (2021). DOI: https://doi.org/10.1126/science.abm0829
8. Bayart J-L et al. (2021). Waning of IgG, Total and Neutralizing Antibodies 6 Months Post-Vaccination with BNT162b2 in Healthcare Workers. Vaccines. 2021; 9(10): 1092, DOI: https://doi.org/10.3390/vaccines9101092
9. Our World in Data. COVID-19 vaccine doses administered by manufacturer, European Union (2023). Available online: https://ourworldindata.org/grapher/covid-vaccine-doses-by-manufacturer (accessed on 29 July 2023)
10. Stein C. et al. (2023). Past SARS-CoV-2 infection protection against re-infection: a systematic review and meta-analysis. The Lancet, Volume 401, ISSUE 10379, P833-842, DOI: https://doi.org/10.1016/S0140-6736(22)02465-5
11. J. M. Ferdinands et al. (2022) Waning of vaccine effectiveness against moderate and severe covid-19 among adults in the US from the VISION network: test negative, case-control study. BMJ 2022; 379 DOI: https://doi.org/10.1136/bmj-2022-072141

Modeling of Wound Healing: The Proliferation and Maturation Stage

Amanda Patrick[1] and Benito Chen-Charpentier[12]

[1] University Of Texas at Arlington, Atlington, TX 76019, USA,
[2] bmchen@uta.edu,
WWW home page: https://mentis.uta.edu/explore/profile/benito-chen

Abstract. A wound is caused when the of the integrity of living tissue in the body is compromised. The body repairs a wound in overlapping stages, namely, homeostasis, inflammation, proliferation, and remodeling. In this paper we deal with the formation of the provisional matrix in the proliferation and maturation stages. The model is based on a system of six ordinary differential equations. It is constructed using assumptions and parameter values from the medical literature. A global sensitivity analysis is performed to determine which parameters cause the largest variations in the solutions. Also, since the reported data has large variations a system of stochastic differential equations is introduced and solved numerically. The models can help test hypothesis about the different species active and their interaction and importance.

Keywords: wound healing, mathematical model, differential equations

1 Introduction

Wounds, whether caused by cuts or impacts, happen very frequently. Most of them are minor and require little care, but more serious wounds can take a long time to heal or even fail to heal. Mathematical models are very important to understand the processes involved and to find how to improve the healing mechanisms, with the consequent benefits to humans and animals suffering wounds. The mechanisms in which the body repairs a wound consists of four overlapping stages: Homeostasis, during which, within minutes, blood clots form and the bleeding stops; inflammation, in which a first provisional matrix forms but the main process is to remove pathogens and debris; proliferation, during which the turnover of the provisional matrix proceeds and new tissues are built; and remodeling, in which the new tissue slowly gains strength and flexibility as collagen fibers reorganize, the tissue remodels, matures and strengthens. Mathematical modeling is a mean to help understand the processes and interactions involved and check theories and hypotheses that are difficult or even impossible to test under experimentation. Ordinary differential equation models are a useful tool for studying dynamics over time. Previous studies including mathematical models of wound healing are: Reynolds et al. [17] who focused on inflammation and anti-inflammation with their state variables being activated phagocytes, tissue

damage, and anti-inflammatory mediators. Cooper et al. [3] built a model expanding on the Reynolds model with the inflammation state variable replaced by the more specialized inflammatory variables, neutrophils and macrophages. Torres et al. [23] utilize these cell dynamics and compare with experimental results. Some studies that modeled the proliferation and remodeling stages are Jin et al. [10] and Segal et al. [20]. Jin et al. incorporate macrophages, MMP-9, TGF-β, fibroblasts, and collagen to model the healing process after a myocardial infarction. Segal et al. [20] construct a model incorporating inflammation, pathogens, fibroblasts, and collagen. In the next sections we present a model for the proliferative stage.

2 Proliferative Stage

In the proliferative stage the focus is shifted from removing debris and pathogens to rebuilding and improving the provisional matrix. An important immunoregulatory cytokine, TGF-β, causes fibroblasts to migrate to the wound by chemotaxis. TGF-β is produced by macrophages,neutrophils, fibroblasts,and myofibroblasts. The production of TGF-β is also enhanced by the process of efferocytosis. These fibroblasts produce the major protein component of the ECM which is collagen. Through the influence of TGF-β, fibroblasts can differentiate into a more specialized cell known as a myofibroblast which also produces collagen but also α-smooth muscle actin which causes the edges of the wound to contract.

Proteases play another important part of the proliferative phase. More specifically, matrix metalloproteinases (MMPs) break down collagen. This contributes to the turn over of collagen as the wound heals. There are different types of MMPs such as MMP-1, MMP-3, and MMP-9. MMPs can be produced by macrophages, neutrophils, fibroblasts, and myofibroblasts. TGF-β induces the expression of tissue inhibitors of matrix metalloproteinases (TIMPS) which inhibit MMPs ability to break down collagen.

TGF-β is a cytokine that contributes to the anti-inflammatory processes. We incorporate the production of TGF-β by M2 macrophages with the term, $k_{\beta M2}M_2$. The production from efferocytosis is represented by $k_{\beta ap}A_N(M_1+M_2)$. The other terms $k_{\beta N}N$, $k_{\beta F}Fk_{\beta My}M_y$, and $\mu_\beta T_\beta$ will represent the production of $TGF-\beta$ from neutrophils, production via fibroblasts, and exit rate of TGF-β, respectively.

Incorporating these mechanisms we get the following differential equation

$$\frac{dT_\beta}{dt} = k_{\beta N}N + k_{\beta ap}A_N(M_1 + M_2) + k_{\beta M2}M_2 + k_{\beta F}F + k_{\beta My}M_y - \mu_\beta T_\beta.$$

Estrogen mediation effects will be also taken into consideration for the proliferative portion. According to Zhou et al. [25], presence of estrogen is associated with an increase production of TGF-β. Adding the effect of estrogen we get the updated equation:

$$\frac{dT_\beta}{dt} = [k_{\beta N}N + k_{\beta ap}A_N(M_1+M_2) + k_{\beta M2}M_2 + k_{\beta F}F + k_{\beta My}M_y](1+k_{\beta e}E) - \mu_\beta T_\beta.$$

MMPs are produced by M1 and M2 macrophages [6, 9], neutrophils [7, 12], fibroblasts [8, 21], and myofibroblasts [14, 22]. In addition to TGF-β's ability to influence migration of fibroblasts to produce collagen, Leivonen et al. [13] notes that TGF-β also plays a role in down regulating the expression of MMPs. They can do this by inducing the expression of tissue inhibitors of MMPs (TIMPs).This inhibition is incorporated into the M_{MP} equation by including the inhibition term $\frac{1}{1+(\frac{T_\beta}{T_{\beta inh}})^2}$. The production via M1, M2, N, F, and M_y is represented by $k_{MMPM1}M1$, $k_{MMPM2}M_2$, $k_{MMPN}N$, $k_{MMPF}F$, and $k_{MMPMy}M_y$, respectively. Finally the exit term for MMPs are represented by $\mu_{MMP}M_{MP}$.

These give the following equation:

$$\frac{dM_{MP}}{dt} = \frac{k_{MMPM1}M1 + k_{MMPM2}M_2 + k_{MMPN}N + k_{MMPF}F + k_{MMPMy}M_y}{1+(\frac{T_\beta}{T_{\beta inh}})^2} - \mu_{MMP}M_{MP}.$$

2.1 Fibroblast and Myofibroblast Equations

For the fibroblast equation the migration to the wound via chemotaxis from TGF-β is represented by the term $c_{F\beta}T_\beta$, and then after these cells migrate, they can proliferate [16, 19]. This is represented by the term $p_F F$. This proliferation may be enhanced by the presence of TGF-β. This is represented by the term $k_{F\beta}T_\beta F$. Fibroblasts can then differentiate into myofibroblasts ($d_F F$) and this is a process that can also be enhanced by the presence of TGF-β [18, 24] ($k_{MyF\beta}FT_\beta$). Fibroblasts that do not differentiate either leave the wound or commit apoptosis. The exit term is represented by $\mu_F F$.

Taking these mechanisms into account, the differential equation for fibroblasts is constructed as the following:

$$\frac{dF}{dt} = c_{F\beta}T_\beta + p_F F + k_{F\beta}T_\beta F - d_F F - k_{MyF\beta}FT_\beta - \mu_F F.$$

For myofibroblasts, we have the differentiation of fibroblasts ($d_f F + k_{MyF\beta}FT_\beta$) and the exit term $\mu_{my}M_y$ giving the following equation:

$$\frac{dM_y}{dt} = d_F F + k_{MyF\beta}FT_\beta - \mu_{my}M_y.$$

Fibroblasts and myofibroblasts secrete different types of collagen including collagen type I and type III [1]. Type III collagen is a weaker form of collagen than type I. At the beginning stages of extracellular matrix reformation, type III. collagen is produced, but is later replaced by stronger type I collagen [4]. The presence of TGF-β enhances the process of collagen secretion, so

the secretion of type I collagen with and without the enhancement of TGF-β is represented by $k_{cwf}(F + M_y)(1 + k_{ctb}T_\beta) + k_{cmy}M_y$. Collagen is defined as a percentage with 0 indicating no collagen in the wound and 1 indicating collagen has filled the wound. The state of existing collagen will affect the rate at which collagen is formed and broken down. To account for this, a collagen deposition multiplier was implemented and defined as $I(C_I + C_{III}) = \frac{1}{1+e^{a(C_I+C_{III}-b)}}$. A similar inhibition is also implemented for degradation by MMPs. Here the term $1 - \frac{1}{1+e^{a(C_{III}+C_I-b)}}$ is used. After collagen type III collagen comes into contact with MMPs and is broken down, it is assumed that type III collagen will be deposited in its place by fibroblasts and myofibroblasts. Finally, the negative effect on the ECM from by-products of neutrophils are taken into account for collagen type III by the term $d_{cn}C_{III}N$.

Taking into account these mechanisms, the following is the resulting equations for collagen type I and collagen type III:

$$\begin{aligned}
\frac{dC_{III}}{dt} &= \frac{k_{cwf}(F + M_y)(1 + k_{ctb}T_\beta) + k_{cmy}M_y}{1 + e^{a(C_{III}+C_I-b)}} \\
&\quad - d_{cn}C_{III}N - d_{cMmp}M_{MP}C_{III}(1 - \frac{1}{1 + e^{a(C_{III}+C_I-b)}}) \\
\frac{dC_I}{dt} &= k_{csf}\frac{M_{MP}C_{III}(F + M_y)}{1 + e^{a(C_{III}+C_I-b)}}
\end{aligned}$$

where $C_{III} + C_I \leq 1$.

3 Final Equations

The final system for the proliferation and remodeling stage is the following:

$$\begin{aligned}
\frac{dT_\beta}{dt} &= k_{\beta N}N + k_{\beta ap}A_N(M_1 + M_2) + k_{\beta M2}M_2 + k_{\beta F}F + k_{\beta My}M_y \\
&\quad - \mu_\beta T_\beta \\
\frac{dM_{MP}}{dt} &= \frac{k_{MMPM1}M1 + k_{MMPM2}M_2 + k_{MMPN}N + k_{MMPF}F + k_{MMPMy}M_y}{1 + (\frac{T_\beta}{T_{\beta inh}})^2} \\
&\quad - \mu_{MMP}M_{MP} \\
\frac{dF}{dt} &= c_{F\beta}T_\beta + p_F F + k_{F\beta}T_\beta F - d_F F - k_{MyF\beta}FT_\beta - \mu_F F \\
\frac{dM_y}{dt} &= d_F F + k_{MyF\beta}FT_\beta - \mu_{my}M_y \\
\frac{dC_{III}}{dt} &= \frac{k_{cwf}(F + M_y)(1 + k_{ctb}T_\beta) + k_{cmy}M_y}{1 + e^{a(C_{III}+C_I-b)}} \\
&\quad - d_{cn}C_{III}N - d_{cMmp}M_{MP}C_{III}(1 - \frac{1}{1 + e^{a(C_{III}+C_I-b)}}) \\
\frac{dC_I}{dt} &= k_{csf}\frac{M_{MP}C_{III}(F + M_y)}{1 + e^{a(C_{III}+C_I-b)}}
\end{aligned}$$

4 Proliferation and Remodeling stage parameters

In order to estimate parameter values, the following assumptions were used in addition to the data from an immunohistochemistry experiment in Kajikawa et al. [11]:

1. Fibroblasts peak between day 7 and day 14 [2, 20]
2. Myofibroblasts peak after fibroblasts peak
3. Collagen finishes being deposited by day 56 [20]
4. MMPs peak around day 5 [10]

In order to use the data, the data for collagen type I and collagen type III was scaled. In a normal state a pre-wounded area has a certain amount of different type of collagen. Collagen type I encompasses a larger amount. For some type of tissue this is around 80 percent, and collagen type III encompasses 20 percent or less [5, 15]. Using the assumption that collagen finishes being deposited by day 56, the signal on the last day of the data in Kajikawa et al. [11] is scaled so that these values are 0.8 and 0.2 for collagen I and collagen III, respectively. The the rest of the values are scaled using the same factor giving the proportion value in the wound.Using the assumptions and the new data a parameter set is found.

5 Global sensitivity analysis for proliferation and remodeling stage

Global sensitivity analysis for the whole model was conducted with respect to total collagen, that is, $C_I + C_{III}$. In order to help reduce the sampling space, the parameters associated with the smallest sensitivity indices for total inflammation variable average where left out. For the parameters associated with higher indices, the same range that was used is used for the inflammation sensitivity analysis was used again here.

Parameters that resulted in a higher sensitivity index were kmmpn (production of MMPs by neutrophils), kpg (growth rate of pathogens), ummp (decay rate of MMPs), dcn (destruction of type III collagen by byproducts of neutrophils) , sb (source of background immune response), kem (estrogen increase in the phagocytic abilities of macrophages), and kmmpm2 (production of MMPs by M2 macrophages).

5.1 Stochastic Differential Equation System

Next a random process was implemented for each state variable equation as before, giving a stochastic differential equation system. Let the random processes $W_{T\beta}(t), W_{mmp}(t), W_F(t), W_{My}(t), W_{C3}(t)$, and $W_{C1}(t)$ be independent standard Brownian motions affecting the densities of $T_\beta, M_{MP}, F, M_y, C_{III}$, and C_I, respectively. Similarly to the inflammation system, the white noise terms proportional to the state variable are implemented for each equation. The stochastic

differential equation system for the proliferation and remodeling variables is as follows:

$$
\begin{aligned}
dT_\beta &= [k_{\beta N}N + k_{\beta ap}A_N(M_1+M_2) + k_{\beta M2}M_2 + k_{\beta F}F + k_{\beta My}M_y - \mu_\beta T_\beta]\,dt \\
&\quad + \sigma T_\beta dW_{T\beta}(t) \\
dM_{MP} &= [\frac{k_{MMPM1}M1 + k_{MMPM2}M_2 + k_{MMPN}N + k_{MMPF}F + k_{MMPMy}M_y}{1+(\frac{T_\beta}{T_{\beta inh}})^2} \\
&\quad - \mu_{MMP}M_{MP}]\,dt + \sigma M_{MP}dW_{mmp}(t) \\
dF &= [c_{F\beta}T_\beta + p_F F + k_{F\beta}T_\beta F - d_F F - k_{MyF\beta}FT_\beta - \mu_F F]\,dt + \sigma F d_F W_F(t) \\
dM_y &= [d_F F + k_{MyF\beta}FT_\beta - \mu_{my}M_y]\,dt + \sigma M_y dW_{My}(t) \\
dC_{III} &= [\frac{k_{cwf}(F+M_y)(1+k_{ctb}T_\beta) + k_{cmy}M_y}{1+e^{a(C_{III}+C_I-b)}} - d_{cn}C_{III}N \\
&\quad - d_{cMmp}M_{MP}C_{III}(1-\frac{1}{1+e^{a(C_{III}+C_I-b)}})]\,dt + \sigma C_{III}dW_{C3}(t) \\
dC_I &= [k_{csf}\frac{M_{MP}C_{III}(F+M_y)}{1+e^{a(C_{III}+C_I-b)}}]\,dt + \sigma C_I dW_{C1}(t)
\end{aligned}
$$

Realizations were simulated for $\sigma = 0.1$ using Milstein method. Some realizations were simulated showing fluctuations in the peak of M1 macrophages and resulting low and high values for collagen type I. The mean of 50, 1000, and 5000 simulations of each variable was analyzed. These means are identical to the deterministic solution.Then the result of taking the mean over 60 days for each iteration is analyzed, the mean for each variable is bounded between a certain range.

6 Conclusions

We have presented a mathematical model based on ordinary differential equations of the proliferation phase of wound healing. The model takes into account the most important processes, immune system cells and factors as reported in the literature. The dynamics of the wound healing process is compared to experimental data. The FAST global sensitivity indices show which parameters have to be measured more accurately, especially when investigating non-typical healing responses. Random biological fluctuations are always present, and if large enough, may cause a delay in the remodeling stage. Stochastic differential equations are introduced to deal with the variability in the variables, errors and unknown factors involved.

Wound healing is a very complicated process and simplifications are necessary, but still results compare well with data.

References

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. Fibroblasts and their transformations: The connective-tissue cell. *Molecular biology of the cell (4th ed.). New York, NY: Garland Science.[Google Scholar]* (2002).
2. Ather, S., Harding, K., and Tate, S. 1 - Wound management and dressings. In *Advanced Textiles for Wound Care (Second Edition)*, S. Rajendran, Ed. Woodhead Publishing, Jan. 2019, pp. 1–22.
3. Cooper, R. L., Segal, R. A., Diegelmann, R. F., and Reynolds, A. M. Modeling the effects of systemic mediators on the inflammatory phase of wound healing. *Journal of theoretical biology 367* (Feb. 2015). Place: London : Publisher: Academic Press.
4. Czubryt, M. Common threads in cardiac fibrosis, infarct scar formation, and wound healing. *Fibrogenesis & tissue repair 5* (Nov. 2012), 19.
5. Davison-Kotler, E., Marshall, W. S., and García-Gareta, E. Sources of Collagen for Biomaterials in Skin Wound Healing. *Bioengineering 6*, 3 (2019).
6. Elkington, P., Green, J., and Friedland, J. Analysis of Matrix Metalloproteinase Secretion by Macrophages. *Methods in molecular biology (Clifton, N.J.) 531* (Feb. 2009), 253–65.
7. Heissig, B., Nishida, C., Tashiro, Y., Sato, Y., Ishihara, M., Ohki, M., Gritli, I., Rosenkvist, J., and Hattori, K. Role of neutrophil-derived matrix metalloproteinase-9 in tissue regeneration. *Histology and histopathology 25 6* (2010), 765–70.
8. Howard, E., Crider, B., Updike, D., Bullen, E., Parks, E., Haaksma, C., Sherry, D., and Tomasek, J. MMP-2 expression by fibroblasts is suppressed by the myofibroblast phenotype. *Experimental cell research 318* (Mar. 2012), 1542–53.
9. James, B. P., Sicheng, W., Qiang, P.-H., and Marianne, Q.-J. Secretion of matrix metalloproteinase-9 by macrophages, in vitro, in response to Helicobacter pylori. *FEMS Immunology & Medical Microbiology 45*, 2 (Aug. 2005), 159–169.
10. Jin, Y., Han, H.-C., Berger, J., Dai, Q., and Lindsey, M. Combining experimental and mathematical modeling to reveal mechanisms of macrophage-dependent left ventricular remodeling. *BMC Systems Biology 5* (2010), 60 – 60.
11. Kajikawa, Y., Morihara, T., Sakamoto, H., Matsuda, K.-i., Oshima, Y., Yoshida, A., Nagae, M., Arai, Y., Kawata, M., and Kubo, T. Platelet-rich plasma enhances the initial mobilization of circulation-derived cells for tendon healing. *Journal of Cellular Physiology 215*, 3 (June 2008), 837–845. Publisher: John Wiley & Sons, Ltd.
12. Kurihara, T., Shimizu-Hirota, R., Shimoda, M., Adachi, T., Shimizu, H., Weiss, S. J., Itoh, H., Hori, S., Aikawa, N., and Okada, Y. Neutrophil-Derived Matrix Metalloproteinase 9 Triggers Acute Aortic Dissection. *Circulation 126*, 25 (Dec. 2012), 3070–3080. Publisher: American Heart Association.
13. Leivonen, S.-K., Lazaridis, K., Decock, J., Chantry, A., Edwards, D., and Kähäri, V.-M. TGF-β-Elicited Induction of Tissue Inhibitor of Metalloproteinases (TIMP)-3 Expression in Fibroblasts Involves Complex Interplay between Smad3, p38α, and ERK1/2. *PloS one 8* (Feb. 2013), e57474.
14. Ma, C., and Chegini, N. Regulation of matrix metalloproteinases (MMPs) and their tissue inhibitors in human myometrial smooth muscle cells by TGF-β1. *Molecular Human Reproduction 5*, 10 (Oct. 1999), 950–954.
15. Pauschinger, M., Doerner, A., Remppis, A., Tannhäuser, R., Kühl, U., and Schultheiss, H.-P. Differential myocardial abundance of collagen type I

and type III mRNA in dilated cardiomyopathy: effects of myocardial inflammation. *Cardiovascular Research 37*, 1 (Jan. 1998), 123–129.

16. Postlethwaite, A. E., Keski-Oja, J., Moses, H. L., and Kang, A. H. Stimulation of the chemotactic migration of human fibroblasts by transforming growth factor beta. *Journal of Experimental Medicine 165*, 1 (Jan. 1987), 251–256.
17. Reynolds, A., Rubin, J., Clermont, G., Day, J., Vodovotz, Y., and Bard Ermentrout, G. A reduced mathematical model of the acute inflammatory response: I. Derivation of model and analysis of anti-inflammation. *Journal of theoretical biology 242*, 1 (Sept. 2006), 220–236.
18. Scharenberg, M. A., Pippenger, B. E., Sack, R., Zingg, D., Ferralli, J., Schenk, S., Martin, I., and Chiquet-Ehrismann, R. TGF-β-induced differentiation into myofibroblasts involves specific regulation of two MKL1 isoforms. *Journal of Cell Science 127*, 5 (Mar. 2014), 1079–1091.
19. Schreier, T., Degen, E., and Baschong, W. Fibroblast migration and proliferation during in vitro wound healing: A quantitative comparison between various growth factors and a low molecular weight blood dialyzate used in the clinic to normalize impaired wound healing. *Research in experimental medicine 193* (1993), 195–205.
20. Segal, R. A., Diegelmann, R. F., Ward, K. R., and Reynolds, A. A differential equation model of collagen accumulation in a healing wound. *Bulletin of mathematical biology 74* (2012), 2165–2182.
21. Shin, J.-W., Kwon, S.-H., Choi, J.-Y., Na, J.-I., Huh, C.-H., Choi, H.-R., and Park, K.-C. Molecular Mechanisms of Dermal Aging and Antiaging Approaches. *International Journal of Molecular Sciences 20*, 9 (2019).
22. Singh, S., and Torzewski, M. Fibroblasts and Their Pathological Functions in the Fibrosis of Aortic Valve Sclerosis and Atherosclerosis. *Biomolecules 9*, 9 (2019).
23. Torres, M., Wang, J., Yannie, P. J., Ghosh, S., Segal, R. A., and Reynolds, A. M. Identifying important parameters in the inflammatory process with a mathematical model of immune cell influx and macrophage polarization. *PLoS Computational Biology 15*, 7 (July 2019). Place: San Francisco Publisher: Public Library of Science.
24. Vallée, A., and Lecarpentier, Y. TGF-[beta] in fibrosis by acting as a conductor for contractile properties of myofibroblasts. *Cell & Bioscience 9*, 1 (Dec. 2019), NA. NA.
25. Zhou, T., Yang, Z., Chen, Y., Chen, Y., Huang, Z., You, B., Peng, Y., and Chen, J. Estrogen Accelerates Cutaneous Wound Healing by Promoting Proliferation of Epidermal Keratinocytes via Erk/Akt Signaling Pathway. *Cellular Physiology and Biochemistry 38*, 3 (2016), 959–968.

Sparse multivariate methods to assess immune response in actively treated oncology patients after COVID-19 vaccination

Conchado, Andrea[1], Fernández-Murga, Leonor[2], Garde-Noguera, Javier[2], Serrano, Lucía[2], Portero, María[2], Llombart-Cussac, Antonio[2], Domínguez-Marques, Victoria[3] and Martín, Nerea[1].

[1] Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València, Spain.
[2] Molecular and Clinical Oncology Department, Hospital Universitari Arnau de Vilanova, Valencia, Spain
[3] Microbiology Department, Hospital Universitari Arnau de Vilanova, Valencia
anconpei@eio.upv.es

Abstract. Patients with cancer are at higher risk of manifesting severe disease and high mortality with COVID-19 virus infection than the non-cancer population. SARS-CoV-2 vaccines have been tested in healthy adult populations. However, specific data on their ability to generate antibodies and cellular immune response in cancer patients receiving anti-tumor treatments are still lacking. In this article, we analyze the usefulness of sparse multivariate methods for the evaluation of immune response in oncology patients receiving anti-tumor treatment and who have received the anti-SARS-CoV2 vaccine during the COVID-19 pandemic. To do so, we will analyze how each set of molecules can be reduced to a smaller set to obtain the underlying patterns among the data. In addition, the humoral and cellular immune response will be related to the level of specific antibodies in the subsample of patients who received the second dose of anti-COVID 19 vaccines. Numerical and graphical results will illustrate the performance and advantages of sparse multivariate methods, and specific sets of molecules will be identified as potential markers of the level of antibodies against COVID 19 in oncology patients.

Keywords: Sparse multivariate methods

1 Introduction

The emergence in December 2019 of the new virus known as Coronavirus (SARS-CoV-2) had devastating consequences worldwide. Although control measures such as the use of masks, physical distancing and isolation were implemented, these actions were not sufficient to stop the spread of the virus. In response to this situation, vaccine development efforts were intensified with the aim of reducing the spread and mortality associated with the virus. As progress has been

made in understanding the efficacy of mRNA-based vaccines, the importance of evaluating their immune response in special patient groups, such as those diagnosed with neoplastic diseases and under active treatment, has also been recognized.

Neoplastic diseases refer to any presence of tumors or neoplasms in the body, which can be benign or malignant, and are characterized by abnormal and uncontrolled cell growth. Patients with neoplastic disease represent a particularly vulnerable population, with an increased risk of acquiring SARS-CoV-2 virus infections and suffering a more severe course of disease. However, until now, the efficacy of vaccines in these patients has been a matter of debate due to their exclusion from the initial clinical studies that led to the approval of mRNA-based vaccines.

Previous research has suggested that the immune response generated by vaccines in patients with solid cancer may be diminished compared to healthy individuals. In addition, it has been observed that those undergoing treatments such as chemotherapy and immunotherapy may have lower levels of antibodies compared to those receiving other types of therapies due to a different reaction of their immune system [6].

The immune response is a reaction that occurs within an organism for the purpose of defense against foreign invaders (viruses, bacteria, tumor cells, etc.). The human immune system consists of a complex network of cells and molecules, including different types of T-lymphocytes (cytokine producers) and B-lymphocytes (antibody producers), which interact to fight foreign agents. Likewise, antibodies, also known as immunoglobulins, are considered key proteins in the functioning of the adaptive immune response. For the production and creation of these antibodies is the vaccine, designed to stimulate a specific immune response against a particular pathogen or its components [4]. The immune response triggered by vaccination is not only limited to the elimination of the pathogen present in the vaccine, but also has a long-lasting effect on the individual's immune system. This is the basis for the development of immunological memory, which allows the immune system to remember the exposure to the pathogen and respond more quickly and effectively in future encounters with the same infectious agent (Montoya, 2021).

Consequently, the analysis of cellular and humoral immunity (antibody generation) generates large data sets due to the number of variables or molecules identified. However, these data sets cannot be treated by classical multivariate dimension reduction methods, given the high number of variables over the number of patients analysed.

In light of the above, this work aims to examine whether cancer patients in active treatment can develop immunity against the SARS-CoV-2 virus through vaccination.

This generic objective has been divided into three specific objectives:

1. To examine differences in Immune Response between different cancer patients' profiles

2. To study the relationship between SARS-CoV-2 antibodies and patients' cellular response

2 Methods

2.1 Participants

The database used in this study contains relevant information from oncology patients under active treatment. These data, obtained from actual patients, include detailed information on their immune response using various cellular indicators. In addition, it is recorded whether the information was collected before the patient received any vaccine or after administering the second dose against COVID-19.

First, the data include patient characteristics, such as age, sex, cancer location, type of treatment and dose (unvaccinated or second dose). Also, the measurement of specific IgG antibody levels against the SARS-CoV-2 virus (Anti_SARS_CoV_2) has been recorded. Finally, there is the cellular response block, the most extensive data set in the database. It includes variables such as CD3, CD4, CD8, CD3_CD56, CD3 central memory, etc. However, for this study, neither the percentage of Apoptosis nor the Total value are considered since they are not the objective of interest of the study. In addition, variables with zero variability have been eliminated. Therefore, although the database contains 94 observations and 219 variables, the number of variables was reduced to 139 after excluding the variables above.

2.2 Sparse Partial Least Squares Regression (sPLS)

The methods for sparse PLS are based on reducing the dimensionality of two data sets measured on the same observations while proposing a selection of variables. The procedure for introducing this sparse approach to identify the most relevant variables distinguishes the various methods, which are structured in one or two phases. The guidelines for sparse PLS are not intended to replace the classical PLS method proposed by [5] and the subsequent improvements in this line of work but rather to provide tools for the selection of variables useful for the analysis of large data sets.

Recently, a new approach to sPLS analysis has been proposed based on introducing Lasso penalties in the factor loadings vectors obtained by PLS when performing the SVD decomposition (singular value decomposition or diagonalisation). These l_1 (lasso) penalties are extracted on each pair of factor loadings vectors in each of the dimensions of the analysis [1]. This approach is integrated into the package mixOmics, which integrates a two-step procedure described into a single step for variable selection and dimensionality reduction based on a modification of the PLS method [2].

Next, it is shown how this method is based on the SVD decomposition associated with the PLS method. Any matrix M_{pxq} of rank r can be decomposed

into three matrices U, Δ and V^T, as exposed: $M = U\Delta V^T$, where U and V^T are orthonormal matrices and Δ is a diagonal matrix whose r elements on the diagonal correspond to the eigenvalues of the $M^T M$ and MM^T matrices. Thus, the present sPLS method is based on the definition of a matrix M such that $M = X^T Y$ such that the column vectors of matrices U and V are the factor loadings of X and Y, respectively.

Based on this approach, in this method for sPLS, it is interesting to penalize both factor loading vectors, u sub k and v sub k, so that variable selection is performed in both matrices. This property is of particular interest since it is intended to facilitate the interpretation of the factor loading vectors.

Thus, the optimisation problem is based on the minimisation of the Frobenius norm between the product of matrices $M = X^T Y$ and the factor loadings vectors u and v, for each dimension h,

$$\min ||M_h - u_h v_h'||_F^2 + g_{\lambda_1}(u_h) + g_{\lambda_2}(v_h)$$

These two penalty parameters $g_{\lambda_1}^h$ and $g_{\lambda_2}^h$ can be selected simultaneously for each dimension h by calculating the root mean squared error prediction criterion (RMSEP) by cross-validation $k - fold$ or leave-one-out cross-validation *leave-one-out*. However, in practical terms it is simpler to select the number of non-zero variables in each dimension h and for each vector of factor loadings u_h, v_h, or both, as proposed in the method for sparse PCA included in the package elastic net [7].

Usually, the number of extracted dimensions is small to facilitate the interpretation of the results. A quantitative criterion to evaluate the number of dimensions to extract as a result of the sPLS method is the assessment of the marginal contribution of each latent variable to the predictive ability of the model (Tenenhaus, 1998), known as Q_h^2.

To select the most relevant variables in each component, a grid vector of values equispaced between 5 and 50 variables is created, in intervals of 5 by 5 variables. This vector contains different numbers of variables to be retained in each of the components. During PLS component extraction, the model is tested with each of these numbers of variables selected using the criterion of minimizing the Mean Absolute Error (MAE).

3 Results and Discussion

The results of the analyses selected to address each of the application objectives are shown below. Figure 1 shows that the production of antibodies against the virus depends mainly on vaccination, but not on the tumour's location. Nor does it depend on other variables related to the patient's characteristics and the treatment, although it is not shown in the work due to lack of space.

The graph shows how the variable representing SARS-CoV-2 antibodies (Y) has a small weight in both the first and second components. This shows the poor relationship of the COVID antibodies with the cellular response, as indicated by the Q2 graph, which shows the model's predictive capacity and the MAE error.

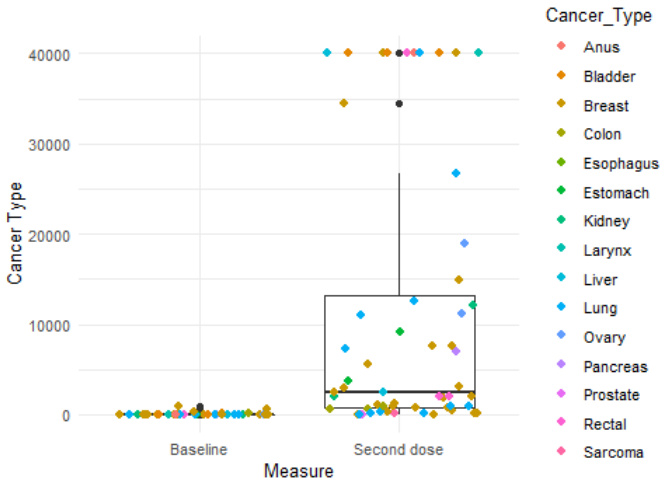


Fig. 1: SARS-CoV-2 antibodies were measured before vaccination and after the second dose of vaccination.

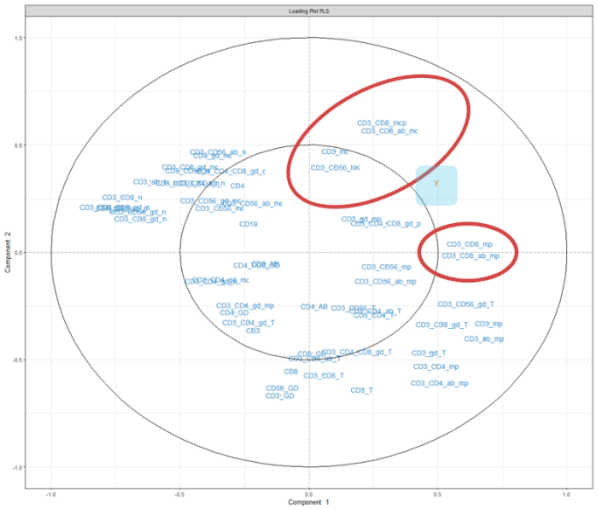


Fig. 2: sPLS Weightings Graph (1st and 2nd component).

On the other hand, the factor scores of both the cellular response and the Y antibodies are shown coloured according to the type of cancer. In both cases, there does not appear to be a significant weight of cancer type on cellular response and Covid antibodies.

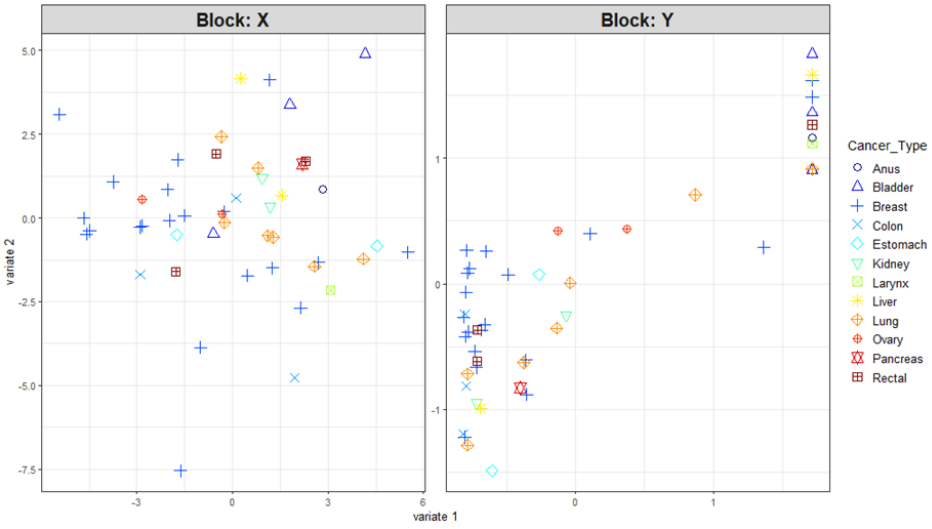


Fig. 3: sPLS Scores Graph (1st and 2nd component), according to cancer location.

Therefore, when analysing the cellular response at different antibody levels, the application of sparse techniques such as sPLS has made it possible to reduce the data from 72 variables to two components of 45 variables each. Although these do not explain much, CD3 or combinations of CD3 stand out as the most important. It can also be seen that the levels of central memory cells and TEMRA are more significant regarding the difference in the ranks of antibodies against the virus.

4 Conclusions

Throughout this work, the fundamentals of regularised regression methods have been reviewed, and the latest methodological proposals for specific multivariate methods have been synthetically presented. Including penalties in the objective functions associated with the optimisation problems generated by dimensionality reduction techniques has opened a very promising field of study. However, the theoretical formulation of these methods does not guarantee their applicability in other fields of study. Each method studied in this work shows how

the implementation of each methodology is conditioned by the algorithm capable of solving the optimisation problem associated with each technique, and its consequent practical characteristics, such as computational requirements or determination of the model fitting parameters.

In the biomedical sciences, the scientific community relies on data sets generated by technological data collection systems. Adapting these data to the functionalities offered by the statistical packages that finally implement the sparse multivariate methods is a key element in selecting one method or another. In this sense, increasing and improving the existing documentation on sparse multivariate methods is advisable to promote their correct application and interpretation. On the other hand, it is advisable to provide the scientific community with standardised and contrasted procedures to assess the suitability of these sparse multivariate methods for each research question and the required characteristics of the available data sets.

Acknowledgements

"Project co-financed by the Consellería de Sanitat Universal i Salut Pública de la Generalitat Valenciana and the European Union through the European Regional Development Fund (ERDF) Operational Program of the Comunitat Valenciana 2014-2020, in the framework of the REACT-EU action lines, as the Union's response to the COVID-19 pandemic."

References

1. Lê Cao, Kim-Anh, Rossouw, Debra, Robert-Granié, Christèle and Besse, Philippe. A Sparse PLS for Variable Selection when Integrating Omics Data. Statistical Applications in Genetics and Molecular Biology, vol. 7, no. 1, 2008. https://doi.org/10.2202/1544-6115.1390
2. Lê Cao, K. A., & Welham, Z. M. (2021). Multivariate data integration using R: methods and applications with the mixOmics package. CRC Press.
3. Montoya, A. F. (2021). Memoria inmunológica, stress y emociones. Le corps et l'analyse. Revue des societés francophones d'analyse bioénergétique.
4. Poland, G., Ovsyannikova, I., Jacobson, R., & Smith, D. (2007). Heterogeneity in Vaccine Immune Response: The Role of Immunogenetics and the Emerging Field of Vaccinomics. Nature publishing group. https://doi:10.1038/sj.clpt.6100415
5. Sjöström, M., Wold, S., Lindberg, W., Persson, J. Å., & Martens, H. (1983). A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. Analytica Chimica Acta, 150, 61-70.
6. Su, E., Fischer, S., Demmer-Steingruber, R., Nigg, S., Güsewell, S., Albrich, W. C., Kahlert, C. R. (2022). Humoral and cellular responses to mRNA-based COVID19 booster Humoral and cellular responses to mRNA-based COVID19 booster. ESMO, 7(5). https://doi:10.1016/j.esmoop.2022.100587.
7. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301-320.

A simulation approach for an extended 2D Quarantine Model

Jesús M. Gandía[1,2] [0000−0001−8997−997X] and R. Dale[1,3] [0000−0002−7262−474X]

[1] Departament d'Estadística, Matemàtiques i Informàtica. Universitat Miguel Hernández (UMH). Avinguda de la Universitat, s/n. 03202 Elx (Alacant)
[2] Dpto. de Matemáticas, Física y Ciencias Tecnológicas. Universidad CEU Cardenal Herrera. Pl. Reyes Católicos, 19, 03204 Elche (Alicante).
[3] Centre of Operations Research (CIO), University Miguel Hernández of Elche (UMH), Elche, 03202 Alicante, Spain
jesus.martinezg@umh.es
rdale@umh.es

Abstract. During the pandemic, the implemented measures led to create predictive mathematical models considering the impact of isolation on infection rates in various regions. The reference [1] focused on the "Quarantine Model" (QM) which considers a sub-population of latently infected individuals, protected from infection due to isolation. However, these models did not account for the spatial spread of the disease. To address this, a spatiotemporal diffusion model is proposed, extending the one-dimensional model to a two-dimensional surface and introducing additional parameters. The analysis performs simulations using real data and different numerical resolution methods for partial differential equations (PDEs). Our study emphasises the importance of qualitative analysis in situations with minimal mobility and compares the results in areas under stricter measures. The findings shed light on the effectiveness of various measures in controlling the spread of infections.

Keywords: Quarantine Model, Pandemic diffusion, PDEs, Covid-19

1 Pandemic Evolution Models

1.1 SIR Model

Recent pandemic conditions have triggered the study of different mathematical models that try to analyse the development of the pandemic under various scenarios. First of all, we must establish the theoretical framework on which the so-called Quarantine Model (QM) is built. The QM represents a modified and extended version of the SIR model, a fundamental model frequently employed in epidemiology for assessing the critical progression of diseases within a population. Among the various versions of the model, we shall focus on the subsequent system of Ordinary Differential Equations (ODEs):

$$\begin{aligned} \frac{dS}{dt} &= -kI(t)S(t), \\ \frac{dI}{dt} &= kI(t)S(t) - \beta I(t) - \sigma I(t), \\ \frac{dR}{dt} &= \beta I(t). \end{aligned} \tag{1}$$

The term kIS describes the disease transmission rate due to contact with infected individuals, βI characterises the rate of recovery for the infected, and σI represents the mortality rate due to the infection. The analysis of this model is well-known and will not be discussed further here. Let us focus solely on determining the conditions of disease progression in cases where the number of infected, recovered, and deceased individuals is significantly lower than the number of susceptible individuals. In this scenario, the susceptible population S in the model can be approximated by a constant value ($S = S_0$). By considering this simplification, we can derive an ordinary differential equation with constant coefficients, making the solution readily obtainable.

1.2 Quarantine Model

In the context of the present study, QM introduces a novel approach aimed at modifying the existing model, considering the extraordinary measures implemented to address the coronavirus infection. The standard SIR (Susceptible-Infectious-Recovered) model, while widely employed in epidemiological analyses, neglects the consideration of the disease's incubation period, which has been demonstrated to play a pivotal role in the spread of the coronavirus.

Notably, individuals afflicted with the coronavirus can become infective even before exhibiting any discernible symptoms. To account for this aspect, we must incorporate a sub-population of latently infected individuals into the model as we see in Eq. (2). These individuals have already contracted the infection during the incubation period but do not exhibit any outward symptoms. At the end of the incubation period, the disease manifests itself, accompanied by the characteristic symptoms, and the affected individual is subsequently isolated in quarantine to prevent further transmission of the infection to others.

$$\begin{aligned} \frac{dS}{dt} &= -kI(t)S(t) \\ \frac{dI}{dt} &= kI(t)S(t) - kI(t-\tau)S(t-\tau) \end{aligned} \tag{2}$$

The modification that QM proposes entails to consider I as the sub-population of latently infected individuals, and τ which represents the duration of the incubation period. The second term in the Eq. (2) corresponds to individuals who were infected at time $t - \tau$, and whose incubation period concludes at time

τ, leading to their placement in quarantine, effectively curbing their ability to transmit the infection further.

To simplify the analytical treatment, we proceed by approximating the susceptible population as a constant ($S(t) = S(t-\tau) = S_0$), and accordingly, we replace it in the equation, resulting in the formulation of the following ODE:

$$\frac{dI}{dt} = kI(t)S_0 - kI(t-\tau)S_0.$$

Limitations of QM. In spite of this improvement, the current model exhibits several evident limitations. Initially, it solely accounts for the initial phase of disease development, wherein the number of susceptible individuals can be reasonably approximated as a constant. While this approximation behaves well under circumstances where the disease propagation is effectively curtailed or regulated, it may prove inadequate when the number of infected cases remains relatively small in comparison to the total population.

Moreover, a notable disadvantage of the model lies in its neglect of the spatial distribution of infected individuals and their subsequent displacement. Addressing this concern, we propose several extensions to the model aimed at ameliorating the impact of these factors on the outcomes. In a previously referenced article [1] introducing the quarantine model, an extension encompassing a term describing one-dimensional spatial diffusion has been put forth. This term characterises diffusion akin to heat propagation as we can see in Eq. (3), albeit involving certain approximations such as representing the population as a fluid, in detail,

$$\frac{\partial I(x,t)}{\partial t} = \delta \frac{\partial^2 I(x,t)}{\partial x^2} + kI(x,t)S_0 - kI(x,t-\tau)S_0. \tag{3}$$

Furthermore, a significant drawback of the current model is the absence of mortality considerations. As the infected individuals are placed in quarantine post-incubation, and they become immune to reinfection, the model precludes the exploration of the dynamics concerning groups susceptible to such outcomes.

To advance the comprehensiveness and accuracy of the model, we must undertake further investigations to encompass these critical aspects and enhance its applicability in real-world scenarios.

2 A Quarantine Model Extension

As we have already mentioned (see ref. [1]), we are faced with a model that incorporates an explicit spatial dimension, where the variables I and S_0 now represent corresponding densities instead of sizes. The addition of a diffusion term characterises the motion of individuals within the system, with an intensity parameter denoted as δ. The new proposed equation, improving the model, is

$$\frac{\partial I(x,y,t)}{\partial t} = \delta \left(\frac{\partial^2 I(x,y,t)}{\partial x^2} + \frac{\partial^2 I(x,y,t)}{\partial y^2} \right) + kS_0(I(x,y,t) - I(x,y,t-\tau)). \tag{4}$$

Our proposal applies the Laplacian operator to the spatial distribution component, thereby extending the model's scope to a specific region of interest (Eq. (4)). In this context, we make an isotropic approximation, as the intensity of movement does not exhibit directional dependence. However, it is important to highlight that we still consider S_0 as an independent constant, unaffected by both time and space. Nevertheless, as the pandemic progresses and the number of infected cases increases, this assumption may no longer hold true, necessitating further refinement of the model to be aware of such circumstances. Then, we can complete the model by introducing a new unknown field that turns the problem back into a system of partial differential equations replacing S_0 by $S(x, y, t)$. In summary, we present two significant modifications to the previously proposed quarantine model, incorporating spatial and temporal diffusion. One approach assumes constant population density (S0) at all times and locations, while the other one considers susceptible as part of the unknown factors in the system of partial differential equations.

The problem at hand can be analogously understood as thermal diffusion with a time-delay system of differential equations resulting from the incubation period, wherein infected individuals cannot be reinfected or transmit the disease to others, leading to stability intervals after the quarantine period. However, the spatially averaged model may not accurately capture dynamics when certain parameters vary across space. In situations where disease development exhibits distinct motion intensities in different locations, eradication may occur in restricted areas, but new outbreaks may arise in non-adopted or inadequately restricted regions. Furthermore, human movement can follow more complex patterns than the Fickian Diffusion assumed, such as a network formed by long-distance connections between major airports. Such considerations emphasise the need for comprehensive and adaptable models to account for real-world complexities in disease spread.

2.1 Simulation Approach

Let us now pay attention to the discretisation of the equations to obtain numerical solutions of the problem. For this purpose, we opted for the FTCS (Forward Time-Centred Space) method, renowned for its simplicity and efficacy, well-suited to the problem's specific values. Henceforth we will employ upper indices to denote temporal coordinates and lower indices for spatial coordinates. Therefore we write

$$I_{ij}^{n+1} = I_{ij}^{n} + \delta\Delta t \left(\frac{I_{i+1,j}^{n} - 2I_{ij}^{n} + I_{i-1,j}^{n}}{\Delta x^2} + \frac{I_{i,j+1}^{n} - 2I_{ij}^{n} + I_{i,j-1}^{n}}{\Delta y^2} \right) + \Delta t k S_0 I_{ij}^{n} - \Delta t k S_0 I_{ij}^{n-\tau}, \quad (5)$$

$$I_{ij}^{n+1} = I_{ij}^{n} + \delta\Delta t \left(\frac{I_{i+1,j}^{n} - 2I_{i,j}^{n} + I_{i-1,j}^{n}}{\Delta x^2} + \frac{I_{i,j+1}^{n} - 2I_{i,j}^{n} + I_{i,j-1}^{n}}{\Delta y^2} \right) + \Delta t k S_{ij}^{n} I_{ij}^{n} - \Delta t k S_{ij}^{n} I_{ij}^{n-\tau}, \quad (6)$$

$$S_{ij}^{n+1} = S_{ij}^{n} + \delta\Delta t \left(\frac{I_{i+1,j}^{n} - 2I_{i,j}^{n} + I_{i-1,j}^{n}}{\Delta x^2} + \frac{I_{i,j+1}^{n} - 2I_{i,j}^{n} + I_{i,j-1}^{n}}{\Delta y^2} \right) + \Delta t k S_{ij}^{n} I_{ij}^{n}. \quad (7)$$

As we pointed out, we segregated the solutions of the model into two versions and proceeded to implement a Python-based routine to obtain results as a function of time. Equations (6) and (7) describe the model in which we consider $S(x, y, t)$ as an unknown vector field of the system of differential equations, while equation (5) refers to the model in which we approximate S as a constant value S_0. The subsequent deliberation shall encompass considerations regarding simulation-dependent factors, such as temporal and spatial stepping, in further detail.

Boundary Conditions. The establishment of appropriate boundary conditions holds paramount significance in ensuring the accurate and reliable development of the simulation. As well-documented, the Von Neumann conditions are widely employed for simulating heat diffusion phenomena. Ensuring a null flow at specific borders, such as the coast of the province of Valencia, becomes vital for the simulation's validity. These conditions are defined by the expressions

$$\frac{\partial I(t, C_x, y)}{\partial x} = 0 \Rightarrow \frac{I_{N+1,j}^{k} - I_{N-1,j}^{k}}{2\Delta x} = 0 \Rightarrow I_{N+1,j}^{k} = I_{N-1,j}^{k} \quad \text{and}$$
$$\frac{\partial I(t, x, C_y)}{\partial y} = 0 \Rightarrow \frac{I_{i,N+1}^{k} - I_{i,N-1}^{k}}{2\Delta y} = 0 \Rightarrow I_{i,N+1}^{k} = I_{i,N-1}^{k}. \quad (8)$$

Through the discretization process conducted earlier, we can ascertain that the flow in both directions is null, which allows us to implement boundary conditions that effectively cancell the flow as wee see in Eqs. (8). This strategic approach serves to prevent errors and discrepancies that might otherwise arise during the simulation, reinforcing the overall robustness and fidelity of the model.

Our focus of investigation centres on the province of Valencia, renowned for its geographical boundaries encompassing a coastal region to the east (Fig. 1). However, it is crucial to acknowledge that for the northern, southern, and western borders, we have refrained from implementing such restrictive conditions. This decision stems from our understanding that despite inter-regional restrictions, the movement of infected individuals might still extend beyond these boundaries, which requires a more comprehensive study in the future to augment the accuracy and reliability of the results.

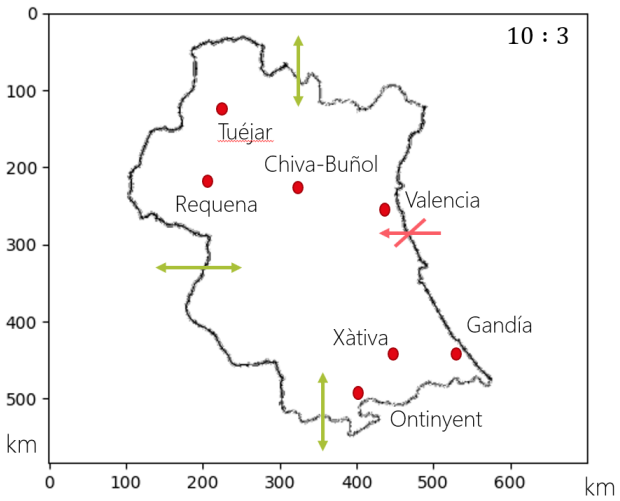


Fig. 1: Establishment of several initial infected zones within prominent cities or regions situated within the province. While considering the spread of infection within this province, we have imposed the necessary boundary conditions along the eastern coast, accounting for the constrained flow of infected individuals due to the maritime barrier.

It is imperative to acknowledge that our approach entails an approximation of reality, considering the complexities of real-world scenarios. This method allows us to explore and analyse the potential transmission dynamics originating from multiple focal points, contributing to a more comprehensive understanding of the disease's spread within the region of interest (Fig. 1). By accounting for these diverse initial conditions, we aim to capture a more representative and nuanced depiction of the epidemiological situation within the simulated area. In this particular simulation, it is crucial to note that each spatial step corresponds to a discrete pixel, with a population density that can either remain constant or vary based on the chosen model. This density directly influences the value of susceptibility attributed to each pixel.

Considering the scale of our simulation, which operates at a 10:3 ratio, we can deduce that each pixel encompasses an area of 0.09 km^2. For the model where S is assumed constant, we have employed the average population density of the entire province. By considering an average population density of 31.3 inhabitants/km^2 in each pixel, we can gauge the epidemiological dynamics at this scale. Moreover, we have assumed a mean incubation time of 6 days for our simulation. These considerations contribute to the accuracy and relevance of our findings as we investigate the disease spread within the province.

FCTS Stability. Given the intricate numerical nature of the simulation and the multitude of constants upon which it relies, undertaking a comprehensive stability analysis becomes imperative. For this purpose, we employ the well-established Von Neumann stability analysis, which entails decomposing numerical errors of the approximations into Fourier series (see [3] and [2]). In a particular case of a function $\psi(x, y, t)$ we have

$$\phi(x, y, t) = \sum_{k,l} c_{k,l}(t) e^{i(kx+ly)},$$

$$c_k(t + \Delta t) = \left[1 - \frac{4\delta\Delta t}{(\Delta x)^2} \sin^2\left(\frac{k\Delta x}{2}\right) - \frac{4\delta\Delta t}{(\Delta y)^2} \sin^2\left(\frac{k\Delta y}{2}\right) + kS_0\Delta t\right] c_k(t).$$

In our specific problem, we can simplify the method by considering the lag term as that of the $n - 1$ iteration and subsequently grouping to obtain an iteration constant. After our calculations, we derive two crucial conditions that determine the maximum permissible value of the time step Δt:

$$\Delta t \leq \min\left\{\frac{3}{4\delta} \frac{(\Delta x \Delta y)^2}{(\Delta x)^2 + (\Delta y)^2}, \frac{1}{kS_0}\right\}. \tag{9}$$

This condition plays a pivotal role in ensuring the stability of the simulation. By incorporating these parameters into the simulation, we can confidently uphold the stability and reliability of our numerical model as we explore the intricate dynamics of the disease spread within a particular area.

2.2 Simulation Results

To acquire the pertinent constants governing the alignment of the infected population with public health data, we have adopted one-dimensional optimisation models concerning the density of infected individuals in cities designated as initial infection zones. Afterwards, several iterations of the simulation have been conducted, wherein the relationship between the constants k and δ has been incrementally fitted.

These iterations enable us to explore the parameter space and gauge the impact of different k and δ values on the simulation's outcomes. By iterative fine-tuning the relationship between these constants, we endeavour to achieve a more accurate representation of the real-world epidemiological dynamics and establish a robust and validated numerical model for our study.

In cities with too high population density, the simulation demonstrates a favourable fit with the data from the initial wave of infections. This success can be attributed to the accurate approximation of S_0 as a constant, which holds true due to the relatively low number of infected cases during that period (Fig. 2). However, a phenomenon emerges in cities heavily impacted by the largest wave of infections, generated by the influence of Valencia. This results in a peak arising from the simultaneous presence of two incubation periods originating from distinct sources (Fig. 3). The overlapping effect contributes to the complexity of the infection dynamics within these cities.

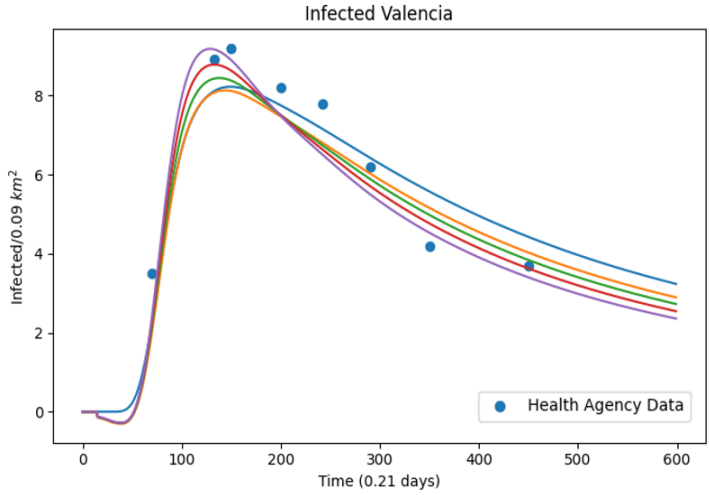


Fig. 2: Evolution of the density of infected population in Valencia as an initial area of infection.

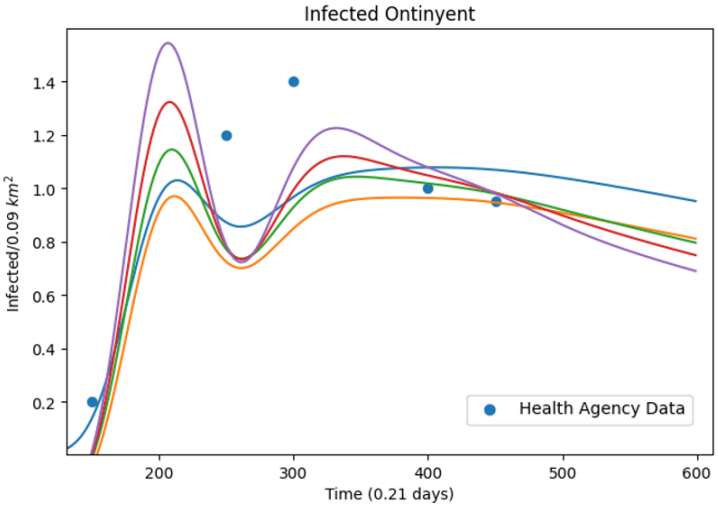


Fig. 3: Evolution of the density of infected population in Ontinyent as an initial area of infection.

In contrast to the observed situation in the city of Ontinyent, we find an opposing trend in cities located farther from the initially selected infection points. These distant cities do not strictly align with the data, mainly because of the absence of nearby sources that could bolster the flow of infections. The lack of nearby infection sources results in a decreased influx of infected individuals to

these cities, causing deviations in the pattern of disease progression compared to the initial infection zones (Fig. 4).

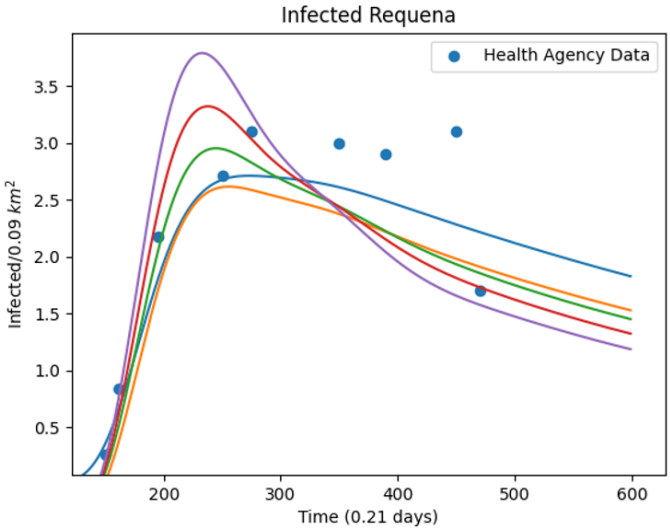


Fig. 4: Evolution of the density of infected population in Requena as an initial area of infection.

An intriguing observation arises from the model's adaptation to the variation in the trend of infected cases attributed to the incubation time. Over the course of a comprehensive 121 days simulation, we noted that the change in the trend of infections did not manifest until approximately 45 days from the onset of the initial wave outbreak, aligning closely with the trends observed in health agency data.

Remarkably, our simulation pinpointed a distinct trend change occurring in approximately 42 days for the city of Valencia. This observation highlights the model's ability to describe and anticipate changes in infection dynamics with remarkable accuracy, reinforcing the validity and reliability of our numerical approach in representing real-world epidemiological patterns. The temporal alignment between simulation outcomes and actual data underscores the model's potential as a valuable tool for predicting and understanding disease spread dynamics, thereby facilitating more effective public health actions and mitigation strategies.

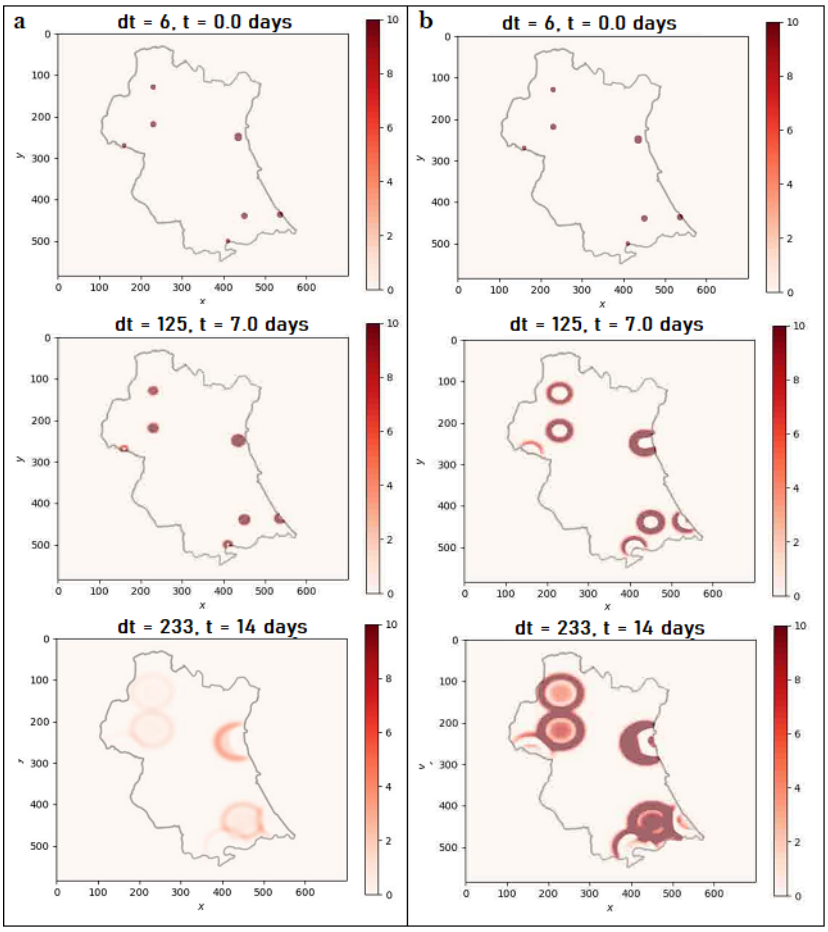


Fig. 5: Images obtained from the dynamic simulation carried out in the province of Valencia. In scenario"a" the simulation parameters facilitate a controlled pandemic evolution, where the number of infected individuals diminishes after reaching its peak. Conversely, scenario "b" depicts an uncontrolled evolution characterised by an unrestrained diffusion of infections with a high value of δ.

References

1. Volpert, V., Banerjee, M., Petrovskii, S. On a quarantine model of coronavirus infection and data analysis. *Mathematical Modelling of Natural Phenomena*, 15, 24 (2020). doi:10.1051/mmnp/2020006
2. G. Recktenwald. FCTS Solution to the Heat Equation. ME 448/548 Notes. Portland State University.
3. W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (2007b). Numerical Recipes 3rd Edition: The Art of Scientific Computing: Chapter 20. Cambridge University Press.

A novel molecular clock model based on anomalous diffusion

Lucas Goiriz[1,2], Raúl Ruiz[1] Òscar Garibo-i-Orts[2], J. Alberto Conejero[2], and Guillermo Rodrigo[1,†]

[1] Institute for Integrative Systems Biology (I2SysBio), CSIC – Universitat de València, 46980 Paterna, Spain
[†] guillermo.rodrigo@csic.es,
WWW home page: https://biosysdesign.csic.es
[2] Institute for Pure and Applied Mathematics (IUMPA), Universitat Politècnica de València, 46022 Valencia, Spain

Abstract. **Keywords:** anomalous diffusion, dynamic systems, virus evolution rate, stochastic process

1 The Molecular Clock Hypothesis

The molecular clock hypothesis, introduced by Emile Zuckerkandl and Linus Pauling in 1965, had a profound impact on evolutionary biology, becoming a vital tool for understanding and dating evolutionary events across various life forms, including archaea, eukarya, bacteria, and even viruses. This hypothesis proposes that the rate of genetic mutations in DNA sequences remains relatively steady over time, resembling a continuous "ticking clock", enabling researchers to estimate species divergence and evolutionary histories [1]. Motoo Kimura further developed this theory into a comprehensive framework in 1968, known as the neutral theory of molecular evolution. According to Kimura's proposal, the rate of molecular evolution is determined by the fixation rate of neutral mutations, which are not influenced by natural selection [2]. This theory also predicts that molecular clocks behave like Poissonian point processes [3, 4].

The application of the molecular clock hypothesis and its framework has been extensive in a wide range of biological problems, including estimating divergence times between species, reconstructing evolutionary relationships, calibrating phylogenetic trees, and tracking disease transmission and epidemics [5–7].

Despite its widespread use, the molecular clock hypothesis has been a subject of intense debate within the scientific community as the process of evolution is complex and affected by environmental changes, transmission bottlenecks, recombination, and speciation events, making it a highly volatile and stochastic phenomenon. Some studies have even shown that the molecular clock model is not valid in numerous cases [8], and others argue that it may not be applicable to all species and populations [9, 10].

Given these discrepancies, several modified molecular clock models have been proposed to address specific complexities and challenges in evolutionary studies.

These models include relaxed molecular clocks, Bayesian molecular clocks, birth-death molecular clocks, and relaxed clocks with covariates [11,12], among others. These adaptations build upon the foundational ideas of the original molecular clock hypothesis while accommodating real-world complexities and expanding the scope of evolutionary research.

To motivate the development of a novel molecular clock model, we will briefly explore how a Poisson point process can be employed to model DNA sequence evolution. We will then extend the Poisson point process into a continuous stochastic process.

1.1 Evolution as a Poisson point process

The Poisson distribution is commonly used to model the occurrence of infrequent events within a fixed time or space interval. In the context of genetic mutations during DNA replication, each generation (defined as a replicative cycle) can introduce changes or substitutions in the DNA sequence due to various factors like errors induced by the DNA polymerase, radiation, or chemicals. Since the likelihood of a mutation at a specific position in the DNA sequence is assumed to be small, constant, and independent between generations (as supported by experimental evidence), the number of mutations in a lineage over n generations can be accurately described using the Poisson distribution.

Let u be the rate of mutations per generation, and n the number of generations. In this scenario, the number of mutations that occur in a lineage during these n generations follows a Poisson distribution with a mean value of un. In addition, if each generation takes the same amount of time, the number of mutations in the lineage during a specific time period t can be described as a homogeneous Poisson point process, denoted as $\{N(t), t \geq 0\}$, where $N(t)$ represents the total number of mutations that have taken place up to (and including) time t. Consequently, the probability of observing exactly n mutations, denoted as $N(t) = n$ at time t, is given by

$$\Pr(N(t) = n) = \frac{e^{-\kappa t} (\kappa t)^n}{n!} \tag{1}$$

where κ the rate of substitutions for a given unit of time. Importantly, **Equation 1** implies that the number of mutations in a lineage at time $t = 0$ is 0 and that the increments of the process are independent.

For further developments, it is convenient to compute the moment generating function, $M_{N(t)}(s)$, of the Poisson process:

$$
\begin{aligned}
M_{N(t)}(s) &= \mathbb{E}\left[e^{tN(t)}\right] && (2)\\
&= \sum_{v=0}^{\infty} e^{sn} \frac{e^{-\kappa t}\left(\kappa t\right)^n}{n!}\\
&= e^{-\kappa t} \sum_{n=0}^{\infty} \frac{\left(\kappa t e^s\right)^n}{n!}\\
&= e^{-\kappa t} e^{\kappa t e^s}\\
&= e^{\kappa t (e^s - 1)} && (3)
\end{aligned}
$$

This way it is trivial to demonstrate that the mean and variance of the process are both given by κt:

$$
\begin{aligned}
\mathbb{E}\left[N(t)\right] &= \frac{\partial}{\partial s} M_{N(t)}(s)\bigg|_{s=0} && (4)\\
&= \left[\kappa t e^{\kappa t (e^s - 1) + s}\right]_{s=0}\\
&= \kappa t && (5)\\
\mathbb{V}\left[N(t)\right] &= \mathbb{E}\left[\left(N(t) - \mathbb{E}\left[N(t)\right]\right)^2\right] && (6)\\
&= \mathbb{E}\left[N^2(t) - 2N(t)\mathbb{E}\left[N(t)\right] + \mathbb{E}\left[N(t)\right]^2\right]\\
&= \mathbb{E}\left[N^2(t)\right] - \mathbb{E}\left[N(t)\right]^2\\
&= \frac{\partial^2}{\partial s^2} M_{N(t)}(s)\bigg|_{s=0} - \left(\kappa t\right)^2\\
&= \left[\kappa t \left(\kappa t e^s + 1\right) e^{\kappa t (e^s - 1) + s}\right]_{s=0} - \left(\kappa t\right)^2\\
&= \left(\kappa t\right)^2 + \kappa t - \left(\kappa t\right)^2\\
&= \kappa t && (7)
\end{aligned}
$$

As a corollary, it is trivial to assess that the process' dispersion index $\rho_{N(t)}$, defined as the ratio between mean and variance, is equal to 1.

1.2 Evolution approximated as a continuous stochastic process

Similar to how the Poisson distribution can be approximated by a Gaussian distribution through the central limit theorem, a Poisson point process can also be approximated by a Wiener process. The Wiener process, also known as Brownian motion, is a continuous-time stochastic process characterized by independent and stationary increments. It is usually represented as $\{W(t), t \geq 0\}$, where $W(t)$ is a random variable representing the displacement of a particle at time t, its increments follow a normal distribution with a mean $\mathbb{E}[W(t)] = 0$ and, if it's the

standard Wiener process, a variance $\mathbb{V}[W(t)] = 1$. The Wiener process is widely used as a model for random fluctuations in various physical systems.

Therefore, the number of mutations during DNA replication can be reformulated as the following Langevin stochastic differential equation

$$\frac{dm(t)}{dt} = \kappa + \sqrt{\kappa}\xi(t) \tag{8}$$

where $\xi(t)$ is a Gaussian white noise characterized by $\mathbb{E}[\xi(t)] = 0$ and covariance function $\mathrm{Cov}[\xi(t)\xi(s)] = \delta(t-s)$. Note that $\xi(t)$ is defined as the formal derivative of the standard Wiener process $W(t)$, an assertion which has to be handled with caution because the Wiener process is nowhere differentiable with probability 1. **Equation 8** can be solved analytically:

$$\begin{aligned}
\frac{dm(t)}{dt} &= \kappa + \sqrt{\kappa}\xi(t) \\
m(t) &= m(0) + \kappa t + \sqrt{\kappa}\int_0^t \xi(s)ds \\
&= \kappa t + \sqrt{\kappa}\int_0^t \xi(s)ds
\end{aligned} \tag{9}$$

Note that this reformulation maintains the Poisson process' mean and variance:

$$\begin{aligned}
\mathbb{E}[m(t)] &= \mathbb{E}\left[\kappa t + \sqrt{\kappa}\int_0^t \xi(s)ds\right] && (10)\\
&= \kappa t + \sqrt{\kappa}\int_0^t \mathbb{E}[\xi(s)]\,ds \\
&= \kappa t && (11)\\
\mathbb{V}[m(t)] &= \mathbb{E}\left[(m(t) - \mathbb{E}[m(t)])^2\right] && (12)\\
&= \mathbb{E}\left[\left(\kappa t + \sqrt{\kappa}\int_0^t \xi(s)ds - \kappa t\right)^2\right] \\
&= \kappa\int_0^t\int_0^t \mathbb{E}[\xi(s)\xi(u)]\,dsdu \\
&= \kappa\int_0^t\int_0^t \delta(s-u)dsdu \\
&= \kappa\int_0^t 1du \\
&= \kappa t && (13)
\end{aligned}$$

As a corollary, the corresponding dispersion index $\rho_{m(t)}$ remains equal to 1 as expected, since the Wiener process is a continuous-time approximation of

the Poisson process. One concern that arises from this reformulation is that the number of mutations $m(t)$ is no longer an integer. However, this issue can be easily solved by applying a rounding function to $m(t)$ whenever it is necessary to obtain an integer value.

2 Anomalous Diffusion

In the preceding section we demonstrated that, according to the molecular clock hypothesis, the number of mutations occurring in a lineage during a specific time period t can be described as a Brownian motion exhibiting a mean and variance equal to κt, where κ represents the rate of substitutions for a given unit of time, akin to a microscopic particle moving in a fluid as a consequence of thermal forces.

However, it is well known that the diffusion of microscopic particles in a fluid does not always conform to Brownian motion. In fact, the diffusion of particles in a fluid can be classified into three main categories depending on their mean squared displacement (MSD; also understood as the variance of the stochastic process governing the motion): normal diffusion, subdiffusion, and superdiffusion. Under normal diffusion, the MSD of the particle is proportional to t, while under subdiffusion and superdiffusion the MSD of the particle is proportional to t^α, where α is known as the diffusion exponent, with $\alpha < 1$ for the former case and $\alpha > 1$ for the latter [13].

Similarly to a microscopic particle moving in a fluid, the number of mutations in a lineage during a specific time period t may not conform to a Brownian motion, as described by several studies observing overdispersed and underdispersed populations. Therefore, it is reasonable to consider that the number of mutations in a lineage during a specific time period t may exhibit anomalous diffusion.

2.1 Evolution as a fractional Brownian motion

Multiple stochastic definitions of anomalous diffusion exist, and it is usually left to the researcher to use the one that best fits their problem. In this work, fractional Brownian motion (fBm) is used as a model for anomalous diffusion due to its simple yet powerful mathematical properties.

fBm is a continuous-time stochastic process, represented as $\{W_\alpha(t), t \geq 0\}$, where $W_\alpha(t)$ is a random variable representing the displacement of a particle at time t, characterized by stationary increments, mean $\mathbb{E}[W_\alpha(t)] = 0$ and a covariance function of the form $\mathrm{Cov}[W_\alpha(t)W_\alpha(s)] = \frac{1}{2}(t^\alpha + s^\alpha - |t-s|^\alpha)$, where α is the diffusion exponent, which determines the degree of long-term dependence of the process. Indeed, the fBm is a generalization of the Wiener process, which corresponds to the case $\alpha = 1$.

To reformulate the number of mutations in a lineage during a specific time period t as a fBm, we will modify the Langevin stochastic differential equation shown in **Equation 8**:

$$\frac{dm(t)}{dt} = \kappa + \sqrt{\kappa}\eta(t) \tag{14}$$

where $\eta(t)$ is an appropriate noise source characterized by $\mathbb{E}[\eta(t)] = 0$ and a covariance function such that $\mathrm{Cov}[W_\alpha(t)W_\alpha(s)] = \int_0^t \int_0^s \mathrm{Cov}[\eta(u)\eta(v)]\,dudv$. It is trivial to compute that $\mathrm{Cov}[\eta(t)\eta(s)] = \frac{\alpha}{2}(\alpha-1)\,|t-s|^{\alpha-2}$. This definition allows for the computation of the appropriate mean and variance of the process:

$$\begin{aligned}
\mathbb{E}[m(t)] &= \mathbb{E}\left[\kappa t + \sqrt{\kappa}\int_0^t \eta(s)ds\right] && (15)\\
&= \kappa t && (16)\\
\mathbb{V}[m(t)] &= \mathbb{E}\left[(m(t) - \mathbb{E}[m(t)])^2\right] && (17)\\
&= \kappa\mathbb{E}\left[\left(\int_0^t \eta(s)ds\right)^2\right]\\
&= \kappa\int_0^t\int_0^t \mathbb{E}[\eta(s)\eta(u)]\,dsdu\\
&= \frac{\alpha\kappa}{2}(\alpha-1)\int_0^t\int_0^t |s-u|^{\alpha-2}\,dsdu\\
&= \frac{\alpha\kappa}{2}(\alpha-1)\int_0^t\left[\int_0^u (u-s)^{\alpha-2}\,ds + \int_u^t (s-u)^{\alpha-2}\,ds\right]du\\
&= \frac{\alpha\kappa}{2}\int_0^t \left[s^{\alpha-1} + (t-s)^{\alpha-1}\right]du\\
&= \kappa t^\alpha && (18)
\end{aligned}$$

Therefore, by using fBm as a model for anomalous diffusion, the number of mutations in a lineage during a specific time period t can be described as a stochastic process with a mean and variance equal to κt (in line with the molecular clock hypothesis) and κt^α, respectively. These values provide valuable insights into the reasons behind overdispersed and underdispersed genetic populations and the extent of their long-term dependence. It is essential to note that when $\alpha = 1$, the fBm simplifies to the Wiener process, resulting in the number of mutations in a lineage during a specific time period t being described as a Brownian motion.

3 Data-driven model validation

Viruses have frequently served as a valuable model system for investigating evolution because of their high mutability and rapid evolutionary changes [14]. In this study, SARS-CoV-2 viral DNA sequences were employed to validate the proposed model due to the comprehensive coverage of the virus's evolution and the availability of high-quality data.

For each sequence in the dataset (all available viral sequences collected in the United Kingdom up to May 2022), the number of mutations was computed by comparing it to the reference SARS-CoV-2 genome sequence (NC_045512.2). Next, the number of mutations were binned in a weekly manner, and the mean and variance of the number of mutations were computed for each week and variant of concern (VoC) annotated. In particular, to perform computations for variants, only sequences annotated as variant v were considered. Certainly, the number of sequences in week k, dubbed N_k, obeys $N_k = \sum_{v \in V} N_{v,k} + N_{\emptyset k}$, where V is the set of variants and $N_{\emptyset k}$ denotes the number of sequences that are not linked to any variant of V in the kth week.

Thus, if there are $N_{v,k}$ sequences in the kth week that are linked to variant v, the mean and variance of the number of mutations are given by

$$\mathbb{E}\left[m_{v,k}\right] = \frac{1}{N_{v,k}} \sum_{i=1}^{N_{v,k}} m_{v,k,i} \tag{19}$$

$$\mathbb{V}\left[m_{v,k}\right] = \frac{1}{N_{v,k}} \sum_{i=1}^{N_{v,k}-1} \left(m_{v,k,i} - \mathbb{E}\left[m_{v,k}\right]\right)^2 \tag{20}$$

The variances were then fitted following the expression

$$\log\left[\left(\mathbb{V}\left[m_{v,k}\right] - \sigma_0^2\right)/\kappa\right] = \alpha \log k \tag{21}$$

Resulting in subdiffusion in the Primal, Alpha and Omicron BA.1 variants, while weak superdiffusion in the case of the Delta variant (Pearson's correlations in log scale, $P < 10^{-4}$ for Primal, Alpha, and Delta and $P = 0.020$ for Omicron BA.1), which resulted in a significant improvement with respect to the null model (Brownian motion). A more elaborated discussion regarding the biological significance of these results, including the implications of the diffusion exponent α and graphical representation of the fitted parameters, can be found in [15].

4 Closing remarks

Anomalous diffusion is gaining traction as a model for describing a great variety of naturally ocurring processes, starting with the diffusion of microscopic particles in a fluid. Models based on anomalous diffusion patterns may be suited to describe the evolution of living entities, including viruses, as they can account for the long-term dependence of the process, which is not possible with the Brownian motion model.

In this work, we have proposed a novel molecular clock model based on anomalous diffusion, which can be used to describe the number of mutations in a lineage during a specific time period t as a stochastic process with a mean and variance equal to κt and κt^{α}, respectively, where κ represents the rate of substitutions for a given unit of time and α is the diffusion exponent, which determines the degree of long-term dependence of the process. This model has

been validated using SARS-CoV-2 viral DNA sequences, resulting in a significant improvement with respect to the null model (Brownian motion).

Acknowledgements

We are indebted to numerous scientists and health professionals that contributed to generate a public database of SARS-CoV-2 sequences. We thank Roser Montagud-Martínez and Nicolás Firbas for their useful discussions regarding evolution and phylogenetic inference. This work was supported by the CSIC PTI Global Health (grant SGL2021-03-040 to GR) through the NextGenerationEU Fund (regulation 2020/2094) and the Valencia Regional Government (grant GVA-COVID19/2021/100 to JAC and grant GVA-COVID19/2021/036 to GR). LG was supported by a predoctoral fellowship from the Valencia Regional Government (ACIF/2021/183).

References

1. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. In: Bryson, V. & Vogel, H.J., Eds., *Evolving Genes and Proteins*. Academic Press, New York, pp. 97-166 (1965)
2. Kimura, M. Evolutionary rate at the molecular level. *Nature*. **217** pp. 624-626 (1968)
3. Kimura, M., & Ohta, T. On the rate of molecular evolution. *J. Mol. Evol.*. **1** pp. 1-17 (1971)
4. Kimura, M. Molecular evolutionary clock and the neutral theory. *J. Mol. Evol.*. **26** pp. 24-33 (1987)
5. Ho, S. The molecular clock and estimating species divergence. *Nature Education*. **1** pp. 168 (2008)
6. Ho, S. Y. & Duchêne, S. Molecular-clock methods for estimating evolutionary rates and Timescales. *Mol. Ecol.*. **23** pp. 5947-5965 (2014)
7. Park, S. Y., Love, T. M., Perelson, A. S., Mack, W. J. &Lee, H. Y. Molecular clock of HIV-1 envelope genes under early immune selection. *Retrovirology*. **13** pp. 38 (2016)
8. Jenkins, G., Rambaut, A., Pybus, O. & Holmes, E. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.*. **54** pp. 156-165 (2002)
9. Langley, C. H., & Fitch, W. M. An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.*. **3** pp. 161-177 (1974)
10. Bedford, T., Wapinski, I. & Hartl, D. Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. *Genetics*. **179** pp. 977-984 (2008)
11. Ayala, F. Molecular clock mirages. *BioEssays*. **21** pp. 71-75 (1999)
12. Kumar, S. Molecular clocks: four decades of evolution. *Nat. Rev. Genet.*. **6** pp. 654-662 (2005)
13. Muñoz-Gil, G., Volpe, G., Garcia-March, M.A. et al. Objective comparison of methods to decode anomalous diffusion. *Nat. Commun.*. **12** 6253 (2021).
14. Koonin, E. & Dolja, V. A virocentric perspective on the evolution of life. *Curr. Opin. Virol.*. **3** pp. 536-557 (2013)
15. Goiriz, L., Ruiz, R., Garibo-i-Orts, Ò., Conejero, J. A., & Rodrigo, G. A variant-dependent molecular clock with anomalous diffusion models SARS-COV-2 evolution in humans. *Proc. Natl. Acad. Sci. USA*. **120** pp. e2303578120 (2023)

Mathematical modeling of COVID-19 vaccine allocation

Gilberto González-Parra[1], Giulia Luebben[1], and Bhumika Bhakta[1]

Department of Mathematics, New Mexico Tech, New Mexico, 87801, USA.
Gilberto.GonzalezParra@nmt.edu

Abstract. In this work we investigate the topic of COVID-19 vaccine allocation. We present some previous studies that have considered the problem of allocating vaccines and in particular for COVID-19. We develop a mathematical model that is useful to investigate how the distribution of vaccines affects the dynamics and outcomes of the COVID-19 pandemic. The model is based on a non-autonomous system of nonlinear differential equations. The model considers age, gender, willingness to vaccinate and comorbidity status. The developed model is complex since it considers a total of 164 ordinary differential equations, which allow us to explore a variety of vaccination programs for the COVID-19 pandemic. We test these programs for the early COVID-19 pandemic when vaccine availability was limited and the number of deaths per week were very large. The evaluation of the different programs is done using the particular scenario of the USA. However, the constructed mathematical model can be applied to other countries or even regions. The findings of this work highlight the significance of developing an effective vaccination program in order to save human lives.

Keywords: Mathematical modeling; COVID-19; vaccination program; age structure; comorbidity; gender.

1 Introduction

During the COVID-19 pandemic there have been more than 6,900,000 deaths worldwide as of July 2023 [41]. This despite that a total of 13,490,832,730 vaccine doses have been administered worldwide (as of 23 July 2023) [41]. However, some low income countries have more than 70% of their populations unvaccinated [41]. For instance, Haiti has only administered a total of 5.83 dose per 100 population and Burundi an impressive 0.34 dose per 100 population [41]. The COVID-19 pandemic brought to light persistent obstacles to adult vaccination, such as lack of availability, lack of accessibility, and lack of trust. It has been recognized that vaccination reduces the burden of the COVID-19 pandemic and as a consequence has been an important factor for the public health worldwide [5].

In December 2020, two COVID-19 vaccines were authorized for emergency use in the United States [42]. Due to a limited vaccine supply the Advisory Committee on Immunization Practices (ACIP) gave priority vaccination to health

care workers and residents and staff members of long-term care facilities during the first phase of the U.S. COVID-19 vaccination program [42]. Ideally, everyone would be vaccinated immediately, since vaccination has been proven to be effective in reducing deaths and cases [9, 35, 49, 51]. However, the reality was that in December 2020 and over the first semester of 2021 there was a very limited supply of Covid vaccines in many countries around the world against the SARS-CoV-2 virus and as a consequence it was necessary to prioritize some subpopulations [22,32,34]. Studies that investigate the development of good vaccination programs is of paramount importance for the people and therefore for public health policymakers.

Allocating vaccines to subgroups under limited availability is a difficult task due to the complexity of the topic and also due to the societal pressure to vaccinate people to save lives. From a scientific viewpoint this distribution problem can be investigated using many different methods and mathematical tools. There are previous studies that have investigated the problem of finding the best protocols to allocate vaccines to the population [24,32,40,47]. There are many factors that can be taken into account in order to assign the vaccines. For instance, a vaccination program may aim to avert the maximum number of deaths, reduce the years of life lost (YLL), the number of cases, or even maintain essential social services [5, 40, 47]. All these aims are plausible and depending on different viewpoints policymakers can choose one over another. In summary, the selection of the main objective of a vaccination program is a very complex decision.

A variety of modeling studies to investigate vaccination programs have been developed [32,33,40]. Due to the intricacy of the COVID-19 pandemic situation and human behavior, each model or study has some limitations [1,10]. It is interesting to note that some works have determined that children should be given vaccine priority because of their crucial role in the spread of influenza [5, 52]. Obviously, this strategy was not possible in the case of the COVID-19 pandemic since the vaccinations were not approved for use in children at the time when vaccines became available. There are a variety of approaches to mathematical modeling studies, which prove helpful in many ways [13,24,38,54]. For instance, numerous simulations can be run, allowing the exploration of various aspects under numerous conditions where uncertainty is a significant component. Some studies have employed SIR or SEIR models without the age-group structure that is essential for finding the most effective vaccination program [28,32,40,46]. More sophisticated mathematical models where age-structure is taken into account have been constructed [9,17,22,24,54]. In [38] the authors found that to minimize deaths or quality adjusted life year losses in the UK, the best vaccination program is to prioritize older age groups first. In [23] the authors also found that for some countries the best vaccination program should not prioritize prioritize the oldest groups if the aim is to minimize expected number of life years saved. In [9] the authors used an age-structured SEIR type model to study the best vaccination strategies. Three age-groups were considered; 20–40 years, 40–60 years and 60+ years. They used contact matrices and different environments such as home, school and work. They assumed a fixed number of vaccine

doses available per day and found that the benefit of prioritizing vaccine allocation among older adults is higher when pace of vaccination is slow. Additional research on vaccination regimens that considered various doses has been carried out [15, 25, 37]. In summary, the significance of COVID-19 vaccine allocation programs under various priority schemes is highlighted by all of these previous results and others [32, 40].

In this work we construct a more detailed and complex mathematical model to investigate vaccine prioritization programs. The proposed model takes into account age, comorbidities, gender and vaccination hesitancy. We decided to consider the gender of people since men are at greater risk of more severe COVID-19 outcomes than women, with biological, socio-cultural and behavioral differences playing fundamental roles [3, 4, 11, 27]. We also chose to include comorbidities because it impacts CFR and social contacts. Taking data from December 2020 to June 2021, we evaluate the specific vaccination availability in the USA. This feature sets the current study apart from many others where the proportion of the population receiving vaccinations is often constant [28, 32, 40, 46]. In our work finding the best vaccination programs is challenging and computationally demanding since the mathematical model has a detailed structure and therefore a large number of equations. Furthermore, the number of potential vaccination programs is very large. In this work we consider the minimization of deaths as the main objective. Based on the previous discussion it can be seen that due to the vast number of variables, high dimensionality, and nonlinearities, related to the vaccine allocation problem this type of research becomes extremely complex.

2 Materials and methods

In this work, we present a mathematical framework to investigate which are the best vaccination programs to reduce the number of deaths during COVID-19 pandemic. First, we design and construct a mathematical model based on a nonautonomous system of nonlinear ordinary differential equations. The non-homogeneous term is related to a time-varying vaccination rate. In this model each state variable represents a different group with regard to COVID disease, vaccination, and comorbidity status. Furthermore, these groups also take into account five different age groups and gender. The model includes symptomatic and asymptomatic individuals. The designed model also includes groups of people who are hesitant to be vaccinated.

The population is split up into subpopulations that are mutually exclusive in the model. Based on disease status, age group, comorbidity status, gender and vaccination status, these subpopulations have been created. Regarding Covid disease status, the following subpopulations are taken into account: susceptible, infected (capable of infecting others), asymptomatic (capable of infecting others), and recovered (not infectious). We only take into account the two comorbidity statuses of zero comorbidities and one or more comorbidities. We take into account five age ranges: 0–39 years old, 40–59 years old, 60–69 years old,

70–79 years old, and 80 years and above. We selected these categories based on the case fatality rate (CFR).

We attempt to keep the mathematical model reasonably simple in order to have a tractable optimization problem. The model considers only two statuses associated to vaccine hesitancy: willing to be vaccinated and vaccine hesitant. People are able to move between subpopulations. During the brief simulation period, the model presupposes that recovered individuals have enduring protection against reinfection [34]. Given the low percentage of breakthrough cases occurring before one semester, this is a plausible assumption. The model assumes that, during the research time span, only susceptible individuals can be inoculated with the vaccine. For those who have received vaccinations, the model assumes that they can become infected. However, their the likelihood of infection is reduced due to the vaccine. This has often been the case in prior investigations [16, 24, 37, 43].

2.1 Mathematical model considering vaccination

As previously indicated, the epidemiological model incorporates a time-varying vaccination term that enables modeling the specific availability of vaccines per unit time in order to have a closer simulation to reality. Most of previous mathematical models are overly simplified compartmental models. The mathematical model constructed in this work considers the social contacts between people from different groups. This is achieved by using different transmission rates for each interaction between individuals from one group with another.

The mathematical framework proposed here employs the subsequent nonautonomous system This allows the model to be written as

$$
\begin{aligned}
&\dot{S}h_{ijk}(t) = -\lambda(t)\, Sh_{ijk}(t),\\
&\dot{S}w_{ijk}(t) = -\lambda(t)\, Sw_{ijk} - v(t),\\
&\dot{S}v_{ijk}(t) = -(1-\epsilon)\,\lambda(t)\, Sv_{ijk}(t) + v(t),\\
&\dot{I}_{ijk}(t) = (1-a)\,\lambda(t)\,(Sh_{ijk}(t) + Sw_{ijk}(t)) - \gamma\, I_{ijk}(t),\\
&\dot{I}v_{ijk}(t) = (1-\epsilon)\,(1-a)\,\lambda(t) Sv_{ijk}(t) - \gamma\, Iv_{ijk}(t),\\
&\dot{A}h_{ijk}(t) = a\,\lambda(t)\, Sh_{ijk}(t) - \gamma\, Ah_{ijk}(t),\\
&\dot{A}w_{ijk}(t) = a\,\lambda(t)\, Sw_{ijk}(t) - \gamma Aw_{ijk}(t),\\
&\dot{A}v_{ijk}(t) = (1-\epsilon)\, a\,\lambda(t)\, Sv_{ijk}(t) - \gamma\, Iv_{ijk}(t),\\
&\dot{R}(t) = \gamma\,[Ah_{ijk}(t) + Aw_{ijk}(t) + Av_{ijk}(t) + (1-\delta_{ijk})\, I_{ijk}(t) + (1-\delta_{ijk})\, Iv_{ijk}(t)],\\
&\dot{D}(t) = \gamma\,\delta_{ijk}\,[I_{ijk}(t) + Iv_{ijk}(t)].
\end{aligned}
\tag{1}
$$

where i, j, k represent the index of the age, sex and comorbidities groups. The variable D denotes the cumulative number of deaths. The force of infection $\lambda(t)$ represents the pace by which a susceptible person contracts the SARS-CoV-2 virus by contact with an infected or asymptomatic carrier. The variable, $Sh_{ijk}(t)$, is the susceptible subpopulation hesitant to vaccinate from age group i, sex j

and comorbidity status k. The variable, $Sw_{ijk}(t)$, represents the susceptible subpopulation willing to vaccinate from age group i, sex j and comorbidity status k. The variable, $Sv_{ijk}(t)$, represents the susceptible vaccinated subpopulation from age group i, sex j and comorbidity status k. Analogously, the mathematical model has the variables $Ah_{ijk}(t), Aw_{ijk}(t)$ and $Av_{ijk}(t)$ that represent the asymptomatic individuals who are unwilling to be vaccinated, willing (to be vaccinated), and vaccinated respectively. The variables $I_{ijk}(t)$ and $Iv_{ijk}(t)$, describe the non-vaccinated infected subpopulations (willing and unwilling) and vaccinated, respectively. The variable $R(t)$ is the recovered people.

2.2 Vaccination rate $\nu(t)$, transmission rates and social contacts

The model includes a parameter ϵ that describes the vaccine efficacy. For this work we use a specific time varying vaccination doses $v(t)$ taking this form data of the USA (December 2020 to June 2021). The inoculation of vaccines is only done on the individuals willing to be vaccinated.

As is typical in mathematical models used to study epidemiology, the force of infection is the the main drift of the model's dynamics. The force of infection is $\lambda(t) = \frac{1}{N}\sum_{i=1}^{5}\sum_{j=1}^{2}\sum_{k=0}^{1}\beta_{ijk}(Ah_{ijk} + I_{ijk} + Aw_{ijk} + Iv_{ijk} + Av_{ijk})$. It is important to note that the transmission rate, β_{ijk}, is included in the force of infection and has been considered to vary depending on the subpopulation. While it is true that each person has a unique chance of contracting an infection, adding this information in a model based on differential equations would make it too complex to use. However, agent-based models have been used to approximate individual behaviors during epidemics [19, 21, 45].

In this research, the people within a group have an average behavior. This is a more traditional mathematical method due to the use of ordinary differential equations. Therefore, each group has its own transmission rate β_{ijk}. This rate depends mainly on the social contact rate and the infectivity of the circulating SARS-CoV-2 variants [17, 24, 32, 40]. To estimate the averaged contacts of each subpopulation in this work, we also use a social contact matrix [26, 44]. For the transmission rate, we assumed a variety of values that have been used in other studies [6, 8, 36]. Younger adults, for instance, are less likely to stay at home or take precautions [47]. Thus, the transmission rate of younger groups is higher. For the model we assume that vaccine hesitant individuals are less likely to adhere to behavioral norms to prevent COVID infection. Therefore, it was anticipated that those who are vaccine hesitant have a transmission rate that is 1.5 times higher than that of those willing to be vaccinated [30]. Additionally, we predicted that individuals with one or more comorbidities would be more inclined to adhere to behavioral norms, which would lower their transmission rate [14].

2.3 Case fatality rates (CFRs) and initial conditions for the subpopulations

In numerous research pertaining to the dynamics of the COVID-19 pandemic, various death rates based on age, comorbidities, and vaccination status have been used (and estimated) extensively [12,24,31,48]. We compute the base case fatality ratio, δ_{base}, for each demographic group [6,33]. We used the fact that people with one or more comorbidities have a CFR increased by 1.97 times [20,50]. Although not all of the initial conditions for each of the subpopulations are available, we assessed initial conditions based on published data for the USA situation from December 2020 to June 2021. We made assumptions about proportions for the initial subpopulations that were unclear based on actual demographic data and information from scientific journals [6]. We also used CDC data to determine the number of people in each demographic group who had contracted COVID-19 prior to December 12, 2020, putting them in the recovered compartment

3 Preliminary Results

In this work we designed a more realistic and complex mathematical model that includes a structure that takes into account age, comorbidity, gender and willingness to be vaccinated. This large structure allows us to consider many different vaccination programs. As a matter of fact a total of 20 factorial potential possibilities just without taking into account the possibility of vaccinating people from different groups at the same time. Therefore, from a computational viewpoint the problem of finding the best vaccination programs with regard to deaths is very demanding. In this work we implement a randomized algorithm in order to find the best vaccination program with regard to the number of deaths. This decision is due to the computational intractability of the 20 factorial potential vaccination programs that can be implemented. In [33] there were 10 factorial feasible vaccination programs and randomized algorithms were not necessary. Randomized algorithms have been applied in many different fields to solve a variety of problems [2, 39]. Thus, we can address the computational challenges of finding the best vaccination programs.

We perform numerical simulations varying the vaccination programs. We use the number of deaths as the metric to evaluate the performance of the ordered priority vaccination programs. The numerical simulations are performed using the mathematical model (1) and varying the order of the priority groups to be vaccinated. The model (1) is fitted to actual data of the number of deaths in order to estimate the base transmission rate. The population structure of the constructed model differs from the one the CDC used to determine the USA vaccination program [24, 33].

Additionally, the CDC employed vaccination by stages, in which multiple subpopulations received vaccinations concurrently. Since the CDC vaccination program was the one that was really employed, we selected the vaccination strategy that was more similar to the vaccination roll out that was implemented in the USA in order to fit the model (1) [7]. With this method, we simply estimated

the base transmission rate because predicting additional parameters based on the data at hand would not guarantee the solution's uniqueness [18, 29, 53].

Fig. 1 displays the total number of fatalities for some random vaccination campaigns using the base transmission rate. As expected, the vaccination programs have a variety of results. From the perspective of public health, this factor is essential since it means that by selecting an ideal vaccination allocation, many lives can be saved. For a fair comparison, all results employ the same beginning subpopulations and base transmission rate. As it can be seen it is quite challenging to identify the best vaccination programs because there were so many different ones tested. However, at a first glance it can be observed that the potential best vaccination programs seem to require to start vaccinating the group one first or group eleven first. The group 1 corresponds to women with zero comorbidities and age group 0-39 years old. The group 11 corresponds to men without comorbidities and age group 0-39 years old. These groups are prioritized due to their high transmission rates, despite the fact they have lower CFRs than the other groups. For this particular base transmission rate this result agrees with the results presented in [5].

Further research and computations are required to reach a more conclusive and specific conclusion with regard to the best vaccination programs. Nonetheless, these preliminary results provide additional insight into finding the best ordered priority vaccination programs.

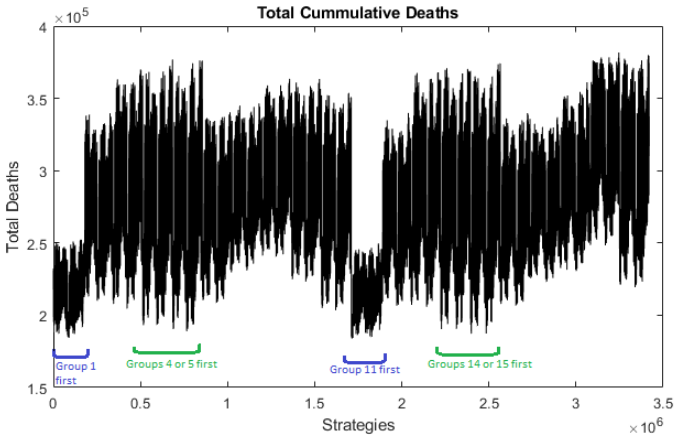


Fig. 1: Total number of fatalities for random ordered priority vaccination regimens using the base transmission rate.

4 Conclusions

In this work we explored a large variety of ordered-priority vaccination programs for COVID-19 vaccine allocation. We presented a background of previous studies related to the allocation of vaccines and in particular for COVID-19 pandemics. We developed a mathematical model in order to investigate how the allocation of vaccines impacts the number of deaths of the COVID-19 pandemic and different outcomes related to this pandemic. The model is based on non-autonomous systems of nonlinear differential equations and considers age, gender, willingness to vaccinate and comorbidity status. We tested a large number of vaccination programs for the early COVID-19 pandemic. We used the particular scenario of the USA. However, the model can be used for other regions. The results of this work provide information about the complexity of designing an optimal vaccine strategy and show the significance of developing effective vaccination programs in order to save human lives, and to be better prepared for future pandemics.

Acknowledgments

Funded with the Aid to promote postdoctoral research from the UPV (PAIDPD-22). This research is also supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103451.

References

1. Asif Afzal, C Ahamed Saleel, Suvanjan Bhattacharyya, N Satish, Olusegun David Samuel, and Irfan Anjum Badruddin. Merits and limitations of mathematical modeling and computational simulations in mitigation of COVID-19 pandemic: A comprehensive review. Archives of Computational Methods in Engineering, pages 1–27, 2021.
2. Jose Aguilar and Gilberto González. Data extrapolation using genetic programming to matrices singular values estimation. In 2006 IEEE International Conference on Evolutionary Computation, pages 3227–3230. IEEE, 2006.
3. Linda Juel Ahrenfeldt, Martina Otavova, Kaare Christensen, and Rune Lindahl-Jacobsen. Sex and age differences in COVID-19 mortality in Europe. Wiener klinische Wochenschrift, 133:393–398, 2021.
4. Yeva Aleksanyan and Jason P Weinman. Women, men and COVID-19. Social Science & Medicine, 294:114698, 2022.
5. Kate M Bubar, Kyle Reinholt, Stephen M Kissler, Marc Lipsitch, Sarah Cobey, Yonatan H Grad, and Daniel B Larremore. Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. Science, 371(6532):916–921, 2021.
6. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html, 2022.
7. Centers for Disease Control and Prevention. https://covid.cdc.gov/covid-datatracker/#vaccination-trends, 2022.

8. Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College, London, 2020.
9. Brody H Foy, Brian Wahl, Kayur Mehta, Anita Shet, Gautam I Menon, and Carl Britto. Comparing COVID-19 vaccine allocation strategies in India: A mathematical modelling study. International Journal of Infectious Diseases, 103:431–438, 2021.
10. Luca Gallo, Mattia Frasca, Vito Latora, and Giovanni Russo. Lack of practical identifiability may hamper reliable predictions in COVID-19 epidemic models. Science advances, 8(3):eabg5234, 2022.
11. Catherine Gebhard, Vera Regitz-Zagrosek, Hannelore K Neuhauser, Rosemary Morgan, and Sabra L Klein. Impact of sex and gender on COVID-19 outcomes in Europe. Biology of sex differences, 11:1–13, 2020.
12. Selene Ghisolfi, Ingvild Almas, Justin C Sandefur, Tillman von Carnap, Jesse Heitner, and Tessa Bold. Predicted COVID-19 fatality rates based on age, sex, comorbidities and health system capacity. BMJ global health, 5(9):e003094, 2020.
13. J Gianatti, PA Lotito, JG Neder, PM Núñez, and LA Parente. Optimal vaccination policies for COVID-19 considering vaccine doses delays. Trends in Computational and Applied Mathematics, 24:121–139, 2023.
14. Aurelie Godbout, Melanie Drolet, Myrto Mondor, Marc Simard, Chantal Sauvageu, Gaston De Serres, and Marc Brisson. Time trends in social contacts of individuals according to comorbidity and vaccination status, before and during the COVID-19 pandemic. BMC Med, 20(199):e0246970, 2022.
15. Gilberto Gonzalez-Parra. Analysis of delayed vaccination regimens: A mathematical modeling approach. Epidemiologia, 2(3):271–293, 2021.
16. Gilberto González-Parra and Abraham J Arenas. Mathematical modeling of SARSCoV-2 Omicron wave under vaccination effects. Computation, 11(2):36, 2023.
17. Gilberto González-Parra, Myladis R Cogollo, and Abraham J Arenas. Mathematical modeling to study optimal allocation of vaccines against COVID-19 using an age-structured population. Axioms, 11(3):109, 2022.
18. Gilberto González-Parra, Miguel Díaz-Rodríguez, and Abraham J Arenas. Mathematical modeling to study the impact of immigration on the dynamics of the COVID-19 pandemic: A case study for Venezuela. Spatial and Spatio-temporal Epidemiology, 43:100532, 2022.
19. Gilberto González-Parra, Rafael-J Villanueva, Javier Ruiz-Baragaño, and Jose-A Moraño. Modelling influenza A (H1N1) 2009 epidemics using a random network in a distributed computing environment. Acta tropica, 143:29–35, 2015.
20. Akram Hernández-Vásquez, Diego Azañedo, Rodrigo Vargas-Fernández, and Guido Bendezu-Quispe. Association of comorbidities with pneumonia and death among covid-19 patients in mexico: a nationwide cross-sectional study. Journal of Preventive Medicine and Public Health, 53(4):211, 2020.
21. Robert Hinch, William JM Probert, Anel Nurtay, Michelle Kendall, Chris Wymant, Matthew Hall, Katrina Lythgoe, Ana Bulas Cruz, Lele Zhao, Andrea Stewart, et al. OpenABM-Covid19—an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. PLoS computational biology, 17(7):e1009146, 2021.
22. Alexandra B Hogan, Peter Winskill, Oliver J Watson, Patrick GT Walker, Charles Whittaker, Marc Baguelin, Nicholas F Brazeau, Giovanni D Charles, Katy AM

Gaythorpe, Arran Hamlet, et al. Within-country age-based prioritisation, global allocation, and public health impact of a vaccine against SARS-CoV-2: A mathematical modelling analysis. Vaccine, 39(22):2995–3006, 2021.

23. Héctor Pifarré I Arolas, Enrique Acosta, and Mikko Myrskyl¨a. Optimal vaccination age varies across countries. Proceedings of the National Academy of Sciences, 118(29), 2021.
24. Md Rafiul Islam, Tamer Oraby, Audrey McCombs, Mohammad Mihrab Chowdhury, Mohammad Al-Mamun, Michael G Tyshenko, and Claus Kadelka. Evaluation of the united states covid-19 vaccine allocation strategy. PloS one, 16(11):e0259700, 2021.
25. Siri R Kadire, Robert M Wachter, and Nicole Lurie. Delayed second dose versus standard regimen for Covid-19 vaccination. New England Journal of Medicine, 384(9):e28, 2021.
26. Mark Kimathi, Samuel Mwalili, Viona Ojiambo, and Duncan Kioi Gathungu. Agestructured model for COVID-19 effectiveness of social distancing and contact reduction in Kenya. Infectious Disease Modelling, 6:15–23, 2021.
27. Sabra L Klein, Santosh Dhakal, Rebecca L Ursin, Sharvari Deshpande, Kathryn Sandberg, and Franck Mauvais-Jarvis. Biological sex impacts covid-19 outcomes. PLoS pathogens, 16(6):e1008570, 2020.
28. Vansh Kodesia, Ankur Suri, and Sarita Azad. An optimal vaccination strategy for pandemic management and its impact on economic recovery. CURRENT SCIENCE, 124(3):319, 2023.
29. Nicholas Lam, Paul Docherty, and Rua Murray. Practical identifiability of parametrised models: A review of benefits and limitations of various approaches. Mathematics and Computers in Simulation, 2022.
30. Carl A Latkin, Lauren Dayton, Grace Yi, Brian Colon, and Xiangrong Kong. Mask usage, social distancing, racial, and gender correlates of covid-19 vaccine intentions among adults in the us. PloS one, 16(2):e0246970, 2021.
31. Andrew T Levin, William P Hanage, Nana Owusu-Boaitey, Kensington B Cochran, Seamus P Walsh, and Gideon Meyerowitz-Katz. Assessing the age specificity of infection fatality rates for COVID-19 systematic review, meta-analysis, and public policy implications. European journal of epidemiology, 35(12):1123–1138, 2020.
32. Kaihui Liu and Yijun Lou. Optimizing COVID-19 vaccination programs during vaccine shortages: A review of mathematical models. Infectious Disease Modelling, 2022.
33. Giulia Luebben, Gilberto González-Parra, and Bishop Cervantes. Study of optimal vaccination strategies for early COVID-19 pandemic using an age-structured mathematical model: A case study of the USA. Mathematical Biosciences and Engineering, 20(6):10828–10865, 2023.
34. C Raina MacIntyre, Valentina Costantino, and Mallory Trent. Modelling of COVID-19 vaccination strategies and herd immunity, in scenarios of limited and full vaccine supply in NSW, Australia. Vaccine, 40(17):2506–2513, 2022.
35. David Martínez-Rodríguez, Gilberto Gonzalez-Parra, and Rafael-J Villanueva. Analysis of key factors of a SARS-CoV-2 vaccination program: A mathematical modeling approach. Epidemiologia, 2(2):140–161, 2021.
36. David Mc Evoy, Conor G McAloon, Aine B Collins, Kevin Hunt, Francis Butler, Andrew W Byrne, Miriam Casey, Ann Barber, John M Griffin, Elizabeth A Lane, et al. The relative infectiousness of asymptomatic SARS-CoV-2 infected persons compared with symptomatic individuals: A rapid scoping review. medRxiv, 2020.

37. Seyed M Moghadas, Thomas N Vilches, Kevin Zhang, Shokoofeh Nourbakhsh, Pratha Sah, Meagan C Fitzpatrick, and Alison P Galvani. Evaluation of COVID-19 vaccination strategies with a delayed second dose. PLoS Biology, 19(4):e3001211, 2021.
38. Sam Moore, EdwardMHill, Michael J Tildesley, Louise Dyson, and Matt J Keeling. Vaccination and non-pharmaceutical interventions for covid-19: a mathematical modelling study. The Lancet Infectious Diseases, 2021.
39. Rajeev Motwani and Prabhakar Raghavan. Randomized algorithms. Cambridge university press, 1995.
40. Eun Bi Noh, Hae-Kweun Nam, and Hocheol Lee. Which group should be vaccinated first?: A systematic review. Infection & chemotherapy, 53(2):261–270, 2021.
41. World Health Organization. Who coronavirus (covid-19) dashboard. https://covid19.who.int/, 2021.
42. Elizabeth M Painter, Emily N Ussery, Anita Patel, Michelle M Hughes, Elizabeth R Zell, Danielle L Moulia, Lynn Gibbs Scharf, Michael Lynch, Matthew D Ritchey, Robin L Toblin, et al. Demographic characteristics of persons vaccinated during the first month of the COVID-19 vaccination program—United States, december 14, 2020–january 14, 2021. Morbidity and mortality weekly report, 70(5):174, 2021.
43. A David Paltiel, Jason L Schwartz, Amy Zheng, and Rochelle P Walensky. Clinical outcomes of a covid-19 vaccine: Implementation over efficacy: Study examines how definitions and thresholds of vaccine efficacy, coupled with different levels of implementation effectiveness and background epidemic severity, translate into outcomes. Health Affairs, 40(1):42–52, 2021.
44. Kiesha Prem, Kevin van Zandvoort, Petra Klepac, Rosalind M Eggo, Nicholas G Davies, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Alex R Cook, and Mark Jit. Projecting contact matrices in 177 geographical regions: an update and comparison with empirical data for the COVID-19 era. PLoS computational biology, 17(7):e1009098, 2021.
45. et al. Romero-Brufau, Santiago. Public health impact of delaying second dose of BNT162b2 or mRNA-1273 covid-19 vaccine: simulation agent based modeling study. BMJ, 373, 2021.
46. Sangeeta Saha, Guruprasad Samanta, and Juan J Nieto. Impact of optimal vaccination and social distancing on COVID-19 pandemic. Mathematics and Computers in Simulation, 200:285–314, 2022.
47. Eunha Shim. Optimal allocation of the limited COVID-19 vaccine supply in South Korea. Journal of clinical medicine, 10(4):591, 2021.
48. Sara Y Tartof, Lei Qian, Vennis Hong, Rong Wei, Ron F Nadjafi, Heidi Fischer, Zhuoxin Li, Sally F Shaw, Susan L Caparosa, Claudia L Nau, et al. Obesity and mortality among patients diagnosed with covid-19: results from an integrated health care organization. Annals of internal medicine, 173(10):773–781, 2020.
49. Ashleigh R Tuite, Lin Zhu, David N Fisman, and Joshua A Salomon. Alternative dose allocation strategies to increase benefits from constrained COVID-19 vaccine supply. Annals of internal medicine, 2021.
50. Maaike van Gerwen, Mathilda Alsen, Christine Little, Joshua Barlow, Eric Genden, Leonard Naymagon, and Douglas Tremblay. Risk factors and outcomes of covid-19 in new york city; a retrospective cohort study. Journal of medical virology, 93(2):907–915, 2021.
51. Rochelle P Walensky, Henry T Walke, and Anthony S Fauci. SARS-CoV-2 variants of concern in the United States—challenges and opportunities. JAMA, 325(11):1037–1038, 2021.

52. Derek Weycker, John Edelsberg, M Elizabeth Halloran, Ira M Longini Jr, Azhar Nizam, Vincent Ciuryla, and Gerry Oster. Population-wide benefits of routine vaccination of children against influenza. Vaccine, 23(10):1284–1293, 2005.
53. Sheng Zhang, Joan Ponce, Zhen Zhang, Guang Lin, and George Karniadakis. An integrated framework for building trustworthy data-driven epidemiological models: Application to the COVID-19 outbreak in New York City. PLoS computational biology, 17(9):e1009334, 2021.
54. Chao Zuo, Zeyang Meng, Fenping Zhu, Yuzhi Zheng, and Yuting Ling. Assessing vaccination prioritization strategies for COVID-19 in South Africa based on agespecific compartment model. Frontiers in public health, 10, 2022.

A fractional-order discrete-time epidemic model with vaccination

Carmen Coll, Damián Ginestar, Alicia Herrero, and Elena Sánchez

Institut Universitari de Matemàtica Multidisciplinar, Universitat Politècnica de València,46022 València, Spain,
mccoll@mat.upv.es, dginesta@mat.upv.es, aherrero@mat.upv.es, esanchezj@mat.upv.es

Abstract. In this paper, we apply the vaccination control strategy to the fractional-order discrete-time SIC epidemic model given in [1]. We use two different scenarios. The first one takes into account the initial variables and involves new parameters related to vaccination and its effectiveness in preventing contagion. In the second scenario, we add a new variable to the model representing the vaccinated individuals. For both cases, the basic reproductive number is obtained to study the behaviour of the disease. However, the main aim of the paper is to analyse the effect of the different parameters on the evolution of the disease when we use vaccination as a control strategy to relieve a certain indirectly transmitted disease. A quantitative relationship would allow us to relate this strategy with the reduction of the burden of the disease. The sensitivity indexes [2] will help to study this effect and will be useful to relate the efficiency with the cost-effectiveness of the vaccination action when reducing the impact of the disease. The obtained results are applied to an epidemiological process developed in a pig farm.

Keywords: fractional order, epidemic model, vaccination, sensitivity analysis

1 Models with vaccination

Let us consider an infectious disease transmitted through the environment, that is, by contact of the population with the contaminant produced by the infected individuals. Population has a constant size and is divided into susceptible individuals with survival rate $0 < p < 1$ and infected individuals with survival rate $0 < q < 1$. In addition, we consider $0 < s < 1$ the survival rate of the contaminant, $0 < \sigma < p$ the infection rate of susceptible individuals and $\beta > 0$ the amount of contaminant.

Using a fractional order derivative discrete-time approach with k memory steps being $0 < \alpha < 1$ the fractional order, in [1], a *Susceptible-Infected-Contaminant* (SIC) discrete-time mathematical model was provided. There, the truncated discrete-time fractional order (DTFO) operator [3] was used and some properties of the model were studied.

Now, we are going to consider that the disease is not eradicated and we plan a control vaccination action, which consists of vaccinating some of the susceptible individuals. We denote by $0 \leq v \leq p$ the vaccinated population rate.

On one hand, we consider that the vaccine can produce full immunity or null immunity. In this case we denote by ξ the percentage of immunized susceptible individuals with the vaccine. Following the steps in [1], we propose the following discrete-time nonlinear model

$$x(t+1) = A_e x(t) - \sum_{j=0}^{k} a_j^{\alpha} x(t+1-j) + (-a_e + b_e)e_1 + e_2, \tag{1}$$

with

$$A_e = \begin{pmatrix} q & 0 & 0 \\ 0 & q & 0 \\ 0 & \beta & s \end{pmatrix}, \; a_e = \sigma(1 - v\xi)x_1(t)x_3(t), \; b_e = (1 - q + \Sigma_k^{\alpha})P, \tag{2}$$

and $x(t) = (x_j(t))_{j=1,2,3}^T$ with $x_1(t)$ representing the susceptible individuals, $x_2(t)$ the infected individuals and $x_3(t)$ the contaminant, and $(e_j)_{j=1,\ldots,n}$ are the canonical basis for the space $\mathbb{R}^n$, in this case $n = 3$. Futhermore, $\Sigma_k^{\alpha} = \sum_{j=0}^{k} a_j^{\alpha}$ being

$$a_j^{\alpha} = \begin{cases} 1 & j = 0 \\ (-1)^j \; \dfrac{\alpha(\alpha-1)...(\alpha-j+1)}{j!} & j > 0 \end{cases}. \tag{3}$$

On the other hand, if the vaccine does not immunize the individuals but reduces the infection rate, we construct a new model introducing a new variable related to the vaccinated individuals and a new parameter $0 < \eta < \sigma$ representing the infection rate in the vaccinated individuals. The proposed model has a four dimensions state variable $x(t) = (x_j(t))_{j=1,2,3,4}^T$ with $x_1(t)$ representing the susceptible individuals, $x_2(t)$ the vaccinated individuals, $x_3(t)$ the infected individuals and $x_4(t)$ the contaminant, and the canonical basis is for $\mathbb{R}^4$. In this case, the model is given by

$$x(t+1) = A_i x(t) - \sum_{j=0}^{k} a_j^{\alpha} x(t+1-j) + (-a_i + b_e) - b_i e_2 + (a_i + b_i)e_3, \tag{4}$$

where

$$A_i = \begin{pmatrix} q - v & q - p & 0 & 0 \\ v & p & 0 & 0 \\ 0 & 0 & q & 0 \\ 0 & 0 & \beta & s \end{pmatrix}, \; a_i = \sigma x_1(t)x_4(t), \; b_i = \eta x_2(t)x_4(t). \tag{5}$$

2 Analysis of the vaccination rate

In this section, we are going to consider an initial model without vaccination in which the disease remains. This occurs when the basic reproduction number R_0 is greater than 1. When non vaccination is considered, in [1], the authors obtained an explicit expression of R_0 given by

$$R_0 = \frac{\beta\sigma P}{(1-q+\Sigma_k^\alpha)(1-s+\Sigma_k^\alpha)}.$$

When the vaccine is taken into account in the process, the basic reproduction number is affected by the parameters involved in the vaccination control action. Now, in the first model, (1)-(2), we obtain that its basic reproduction number is

$$R_e = \sqrt{\frac{\beta\sigma(1-v\xi)P}{(1-q+\Sigma_k^\alpha)(1-s+\Sigma_k^\alpha)}}.$$

This expression allow us get a lower bound on the vaccination rate v, which is given in the next Proposition.

Proposition 1. *If $R_0 > 1$ and $\xi > 1 - \dfrac{1}{R_0^2}$, then the disease tends to disappear in the model with vaccine (1)-(2) if and only if the vaccination rate satisfies*

$$v > \frac{1}{\xi}\left(1 - \frac{1}{R_0^2}\right)$$

In the second model, (4)-(5), linearizing around of the disease-free equilibrium point, which is given by $x^* = \dfrac{P}{1-p+\Sigma_k^\alpha+v}\,(1-p+\Sigma_k^\alpha, v, 0, 0)$, the basic reproduction number results

$$R_i = \sqrt{\frac{\beta(\sigma x_1^* + \eta x_2^*)}{(1-q+\Sigma_k^\alpha)(1-s+\Sigma_k^\alpha)}}.$$

In this case, the lower bound on the vaccination rate v is given in the next Proposition.

Proposition 2. *If $R_0 > 1$ and $\eta < \dfrac{\sigma}{R_0^2}$, then the disease tends to disappear in the model with vaccine (4)-(5) if and only if the vaccination rate satisfies*

$$v > (1-p+\Sigma_k^\alpha)\frac{R_0^2 - 1}{1 - \frac{\eta}{\sigma}R_0^2}.$$

3 Sensitivity analysis applied to a pig farm

Sensitivity analysis aids in identifying the key parameters of the model. We are particularly interested in estimating how the corresponding parameter should be changed using this approach in order to reduce the basic reproduction number by a specified percentage. A reduction in the basic reproduction number value corresponds to a decrease in the number of individuals with the infection. If the basic reproduction number is less than 1, a decrease in its value correlates to a faster approach to equilibrium, which means that the disease will vanish more quickly. Additionally, a lower basic reproduction number indicates that the disease does not spread as virulently when it is bigger than one.

We make the assumption in our study that the models represent a pig farm where an infectious epidemic is indirectly spread by ingestion of the contaminant. We used some of the data from [4] and [5], which match to the specific case of a Salmonella infection, to evaluate the parameters of our models. Thus, $p = 0.9995$, $q = 0.99$, $s = 0.98$, $\sigma = 0.24 \times 10^{-9}$ Bacteria^{-1}, $\beta = 2.25 \times 10^{4}$ Bacteria.Indiv^{-1} colony-forming unit (c.f.u.).

We compute the sensitivity indices of the basic reproduction number of both models. Concretely, we consider a population size $P = 100$ and a fractional order $\alpha = 0.99$ with $k = 10$ memory steps.

We perform some numerical sensitivity analyzes of the basic reproduction number for both models with respect to parameters v, ξ in the first model, and concerning parameters v, η, in the second model. It is known, [2], that the normalized sensitivity index for a quantity Q with respect to a parameter h is defined by $\Phi(Q/h) = \dfrac{h}{Q}\dfrac{\partial Q}{\partial h}$. Then, for the model (1)-(2), we obtain

$$\Phi(R_e/\xi) = -\frac{(1-\xi)v}{2(1-\xi v)}, \quad \Phi(R_e/v) = -\frac{v\xi}{2(1-\xi v)}.$$

We observe that taking $v = 0.564$ and $\xi = 0.6$, the disease remains since the corresponding $R_e = 1.241$ is greater than 1. Applying the above sensitivity indices we can reduce this basic reproduction number by about 10% increasing the vaccination rate to $v = 0.784$ or the percentage of effectivity to $\xi = 0.835$.

Analogously, we make a sensitivity analysis of the basic reproduction number of the model (4)-(5). In this case, the sensitivity indices are

$$\Phi(R_i/\eta) = \frac{\eta\Lambda}{2}, \quad \Phi(R_i/v) = \frac{(1-p+\Sigma_k^{\alpha})(\eta-\sigma)\Lambda}{2(1-p+\Sigma_k^{\alpha}+v)}, \quad \Lambda = \frac{v}{v\eta+\sigma(1-p+\Sigma_k^{\alpha})}.$$

We see how appropriate variation of the values of parameters η and v implies the reduction of the value of the basic reproduction number.

If $v = 0.01$ and $\eta = 9 \times 10^{-11}$ the basic reproduction number is $R_i = 1.032$. We can reduce by 10% this value by increasing the vaccination rate to $v = 0.023$ or by decreasing the infection rate of the vaccinated individuals to $\eta = 6.466 \times 10^{-11}$.

Figures 1 and 2 show the evolution of the infected population when one of these parameters is changed according to the new proposed values.

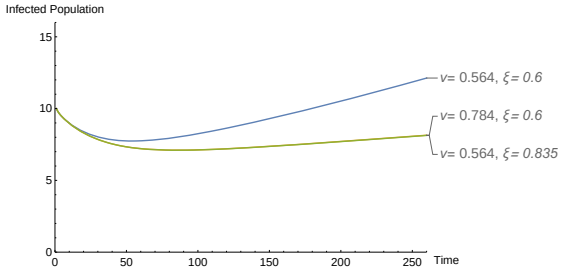


Fig. 1: Comparison of the evolution of infected population for model (1)-(2) with 10 memory steps and $\alpha = 0.99$ when varying the parameters v and ξ according to the sensitivity indices.

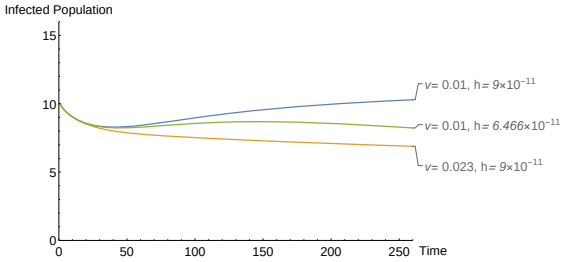


Fig. 2: Comparison of the evolution of infected population for model (4)-(5) with 10 memory steps and $\alpha = 0.99$ when varying the parameters v and η according to the sensitivity indices.

References

1. Coll, C., Ginestar, D., Herrero, A., Sánchez, E.: The discrete fractional order difference applied to an epidemic model with indirect transmission. Appl. Math. Model. 103, 636-648 (2022). doi:10.1016/j.apm.2021.11.002
2. Siriprapaiwan, S., Moore, E.J., Koonprasert, S.: Generalized reproductions numbers, sensitivity analysis and critical immunity levels of an SEQIJR disease model with immunization and vaying total population size. Math. Comput. Simul. 146, 70-89 (2018). doi:10.1016/j.matcom.2017.10.006
3. Dzielinski, A., Sierociuk, D.: Stability of discrete fractional order state-space systems. In: Procceding of 2-IFAC Workshop on Fractional Differentiation and its Applications, Porto, Portugal, July 19-21, (2006)
4. Berriman, A.D.C., Clancy, D., Clough, H.E., Christley, R.M.: Semi-stochastic models for Salmonella infection within finishing pig units in the UK. Math. Biosci. 245, 148–156 (2013) doi:10.1016/j.mbs.2013.06.004
5. Berrmann, A.D.C.: Mathematical modelling of the dynamics and control of Salmonella on UK pig farms. PH.D. thesis, University of Liverpool, UK (2012)

Efficiency analysis of public hospitals in Colombia between 2017-2021 and the influence of different variables

Ricardo Losada Sáenz[1], Isabel Barrachina Martínez[2], and María Caballer Tarazona[2]

[1] Universitat Politècnica de València,46022 València, Spain,
[2] INECO, Universitat Politècnica de València and Facultat d´Economia Universitat de València 46022 València, Spain
RLOSSAE@upv.edu.es

Abstract. The objective of this research is to measure health efficiency among 23 public institutions selected with complexity level 3, that is, IPS that has the greatest infrastructure and highly qualified professionals. The Data Envelopment Analysis (DEA) methodology is used to measure the effi-ciency between the IPS of different cities in Colombia that serve the population, mostly with the lowest economic resources in the country. There are barriers to access health services. Although in theory the entire population has a right to services, the requirements to have external consultations, emergencies, surgical interventions and health care services are quite high, due to a high volume of administrative proce-dures [1], a situation that makes it difficult for users to use the health service. The results show that none of the IPS is efficient during the 5 years of the study, there are efficiencies in several years and in IPS of several cities in the country.

Keywords: Efficiency, data envelopment analysis – public clinics and hospitals.

1 Introduction

The study carried out focuses its attention on the processes of health care services in the Colombian Health System, specifically related to the installed capacity (Number of beds), number of external outpatient consultation, number of hospital emergencies, number of hospital discharges and rate of readmission of patients in less than 15 days, services that are provide in clinics and hospitals, probably some variables can negatively influence efficiency, use of resources and institutional results. These variables are included in the research, building a data panel that is analyzed through the DEA method, in this way clinics and hospitals that are efficient and inefficient can be identified.

The selection of clinics and hospitals (IPS) is established by determining institutions with similar characteristics with level of complexity number 3, that

is, 23 clinics and hospitals that have high technology, specialized professionals, medical processes of highly specialized surgeries, large infrastructures. and significant volumes of patient care. The results allow us to identify the health service providers (IPS), public clinics and hospitals that are closest to the efficient frontier, as well as the IPS that are furthest away from the efficient frontier.

2 Methods

The objective of the study is analyse the efficiency of public clinics and hospitals in Colombia between 2017 and 2021 and determine the influential variables.

The methodology used was data envelopment analysis with input orientation, using the basic radial model with variable returns to scale. Data Envelopment Analysis [2] introduces the linear programming model, the objective is to construct a panel of data to identify efficient and inefficient units (Darairo and Simar, 2007), according to Cordero, J.M.; García, A.G.; Cortes, E.L.; Polo, C (2021) [3] based on the following formula:

$$\hat{\Psi} = \left\{ (x,\ y) \in \mathbb{R}_{+}^{p+q} \ | y \leq \sum_{i=1}^{n} \gamma_i \cdot\ y_i;\ x \geq \sum_{i=1}^{n} \gamma_i \cdot x_i \text{ for } (\gamma_1, \ldots, \gamma_n) \right.$$
$$\left. \text{s.t.} \quad \sum_{i=1}^{n} \gamma_i = 1;\ \gamma_i \geq 0,\ i = 1,\ \ldots, n \right\}. \quad (1)$$

Banker R. D.; Charnes, A.; Cooper, W. W. (1984) [4] assumes variable returns to scale. The identification of efficient units is determined according to Cordero, J.M.; García, A.G.; Cortes, E.L.; Polo, C (2021) with units that are equal to 1 as efficient and those that are less than 1 as inefficient:

$$\hat{\theta}_{DEA}\,(x,\ y) = \inf \left\{ \theta |\ (\theta x,\ y) \in\ \hat{\Psi} \right\} \quad (2)$$

The Ministry of Health and Social Protection formally delivers the information, via formal right of request, with the detail of the inputs and outputs required for the period 2017 to 2021. Input and output variables were established in the research, the input variables being the number of hospital beds (non-discretionary variable) and the cost of the general social security system, the output variables are as follows number of external consultations, hospital emergencies, hospital discharges and rate of readmission in less than 15 days (unwanted variable) [5]. The data panel included the production and resource values of the IPS throughout the period under study.

The selection is made by taking into account the level of complexity of each IPS, where level 1 is basic health care services up to the highest level (Level 3). The selected IPS are of complexity level 3, characterized by having more technology, specialized and subspecialized personnel (Rev. Gerenc. Polit. Salud, Bogotá, Colombia, 16 (32): 51-65, January- June 2017). Regarding the cost of the general social security system the National Superintendence of Health of

Colombia produces a report entitled "Financial Information Catalog. Report for oversight purposes of entities classified in the IFRS Group 1, 2 and 3 for health service provider institutions (IPS)" [6], details several items which include the costs of the general social security system of each health service provider institution (IPS), System costs include Medicines, materials, fees, administrative services, maintenance, repairs, emergencies, outpatient care, hospitalization, operating rooms, diagnostic support, therapeutic support, high cost , pharmacy etc. Costs and expenses can be predicted considering the policies and strategies of each country according to Tarazona, V.C., Guadalajara-Olmeda, N.; Vivas-Consuelo, D. (2019) [7]. All entities supervised by the National Superintendence of Health are required to submit the information in this report on a monthly, quarterly, semiannual and annual basis according to Resolution 1043 of 2006 and the External Circular of the National Superintendent of Health (2016). The study uses data cut-off as of December of each year from 2017 through 2021 [8].

3 Results

The total number of level 3 public IPS under study are 23 clinics and hospitals (Figure 1), chosen based on the largest size of the number of beds. The graph below shows the distribution of beds by the volume of the institutions:

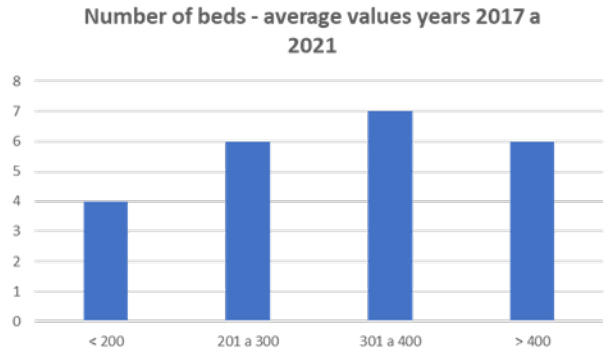


Fig. 1: Number of beds.

The number of efficient IPS varied each year between 2 and 7 (Figure 2). It was noted that 4 IPS obtained efficient results in 3 out of the 5 years of the study. 7 IPS achieved a higher average efficiency rating over the whole 5 years. 22% of the IPS are close to the efficient frontier and need to reduce all their inputs by between 2% and 13%, while 69% of the IPS need to reduce all inputs by between 18% and 55% in order to be on the efficient frontier. Finally, the 9% of IPS furthest from the efficient frontier need to reduce all their inputs by between 70% and 83%.

Regarding the evolution of efficiency between 2017 and 2021, it was noted that the average efficiency of all the IPS increased between 2017 and 2019 but decreased between 2020 and 2021.

Nombre sede IPS	2017	2018	2019	2020	2021	Promedio	3 años eficientes
HOSPITAL UNIVERSITARIO DEL CARIBE	0,64002	0,70732	0,85447	0,59609	0,53639	0,666858	
HOSPITAL UNIVERSITARIO ERASMO MEOZ	1	0,99544	1	0,91186	1	0,98146	1
HOSPITAL MANUEL URIBE ANGEL	0,71106	0,85159	0,86799	1	0,38091	0,76231	
HOSPITAL UNIVERSITARIO DEL VALLE EVARISTO GARCIA	0,3818	0,30681	0,29685	0,21929	0,299	0,30075	
HOSPITAL UNIVERSITARIO HERNANDO MONCALEANO PERDOMO DE NEIVA	0,48202	0,60731	0,24286	0,45049	0,50508	0,457552	
HOSPITAL UNIVERSITARIO SAN JORGE	0,63405	0,67175	0,74522	0,74355	0,74212	0,707338	
HOSPITAL DEPARTAMENTAL UNIVERSITARIO DEL QUINDIO SAN JUAN DE DIOS	0,7215	0,83789	1	0,60183	0,86049	0,804342	
HOSPITAL GENERAL DE MEDELLIN LUZ CASTRO DE GUTIERREZ	0,68315	0,87993	0,85063	0,58888	0,36772	0,674062	
HOSPITAL UNIVERSITARIO SAN JOSE DE POPAYAN	0,45978	0,7587	0,95936	0,64733	0,66702	0,698438	
Unidad de Servicios de Salud Occidente de Kennedy HOSPITAL KENNEDY ESE	0,92559	1	1	1	0,89837	0,964792	1
HOSPITAL LA MARIA	0,9067	1	0,56035	0,49603	0,31332	0,65528	
HOSPITAL UNIVERSITARIO DE SANTANDER	0,43839	0,48814	0,56344	0,41645	0,46268	0,47382	
INSTITUTO NACIONAL DE CANCEROLOGIA	0,15534	0,19746	0,13066	0,14436	0,20038	0,16564	
HOSPITAL UNIVERSITARIO DEPARTAMENTAL DE NARIÑO	0,63537	0,54993	0,61145	0,44033	0,5102	0,549456	
UNIDAD DE SERVICIOS DE SALUD SIMÓN BOLÍVAR	1	0,9427	1	0,58232	0,5809	0,821184	
HOSPITAL UNIVERSITARIO DE LA SAMARITANA	0,52541	0,60113	0,65737	0,50684	0,5554	0,56923	
UNIDAD DE SERVICIOS DE SALUD SANTA CLARA	0,71136	0,80207	0,82284	0,61072	0,66139	0,721676	
HOSPITAL UNIVERSITARIO JULIO MENDEZ BARRENECHE	0,79415	0,89055	0,76545	0,75235	0,78809	0,798118	
HOSPITAL FEDERICO LLERAS ACOSTA	0,37346	0,47421	0,50135	0,42664	0,46178	0,447488	
HOSPITAL UNIVERSITARIO SAN RAFAEL DE TUNJA	0,60574	0,75874	0,88573	0,46033	0,52423	0,646954	
UNIDAD DE SERVICIOS DE SALUD MEISSEN	1	0,9169	1	0,67509	1	0,918398	1
HOSPITAL DEPARTAMENTAL UNIVERSITARIO SANTA SOFIA DE CALDAS	1	0,83938	1	0,82144	0,69447	0,871058	
HOSPITAL REGIONAL MANUELA BELTRAN SOCORRO	1	1	1	0,92833	0,84771	0,955208	1
Valores promedio	0,68629957	0,74251957	0,75287043	0,60958913	0,60250652		
IPS Eficiente por año	5	3	7	2	2		4

Fig. 2: Evolution of efficiency years 2017 to 2021.

The costs of the general social security system increase (Figure 3), there is evidence of an increase in external outpatient consultations in the years 2017 to 2019, the years 2020 and 2021 decrease, with a percentage variation of 14.4%. The number of hospital emergencies shows a notable decrease for the years 2020 and 2021(Figure 4). The number of hospital discharges behaves normally with few increases, with the exception of 2019 and the proportion of readmissions of patients hospitalized in less than 15 days increases every year.

22% of IPS are close to the efficient frontier and need to reduce all their inputs between 2% and 13%, 69% of IPS need to reduce all inputs between 18% and 55% to be on the efficient frontier and 9% of IPS is very far from the efficient frontier need to reduce all their inputs between 70% and 83%.

The evolution of average efficiency in clinics and hospitals between 2017 and 2021 increased between 2017 and 2019, but decreased between 2020 and 2021(Table 1). The cause of decreased efficiency in 2019 and 2021 is likely to be the covid 19 pandemic.

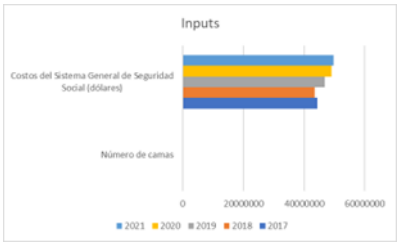


Fig. 3: Inputs.

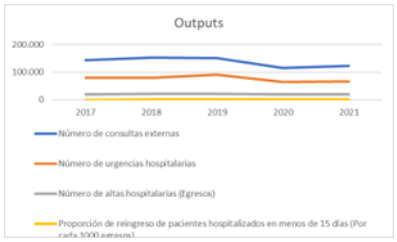


Fig. 4: Outputs.

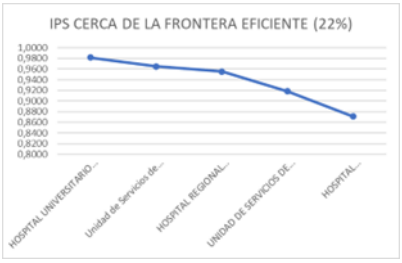


Fig. 5: IPS Near the efficient frontier.

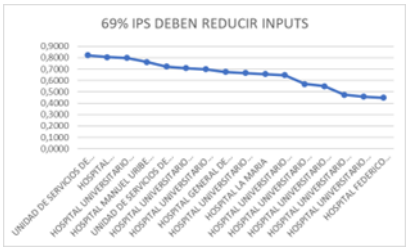


Fig. 6: 69% IPS should reduce their Inputs.

Year	average efficiency
2017	0,6863
2018	0,7425
2019	0,7529
2020	0,6096
2021	0,6025

Fig. 7: Average Efficiency Data.

Fig. 8: Average efficiency

The public resources allocated to each IPS have a percentage variation of 12.3%, i.e. the budget increased between 2017 and 2021 and finally the number of beds increases with a percentage change of 5.3%, especially in 2020 and 2021.

4 Conclusions

- Of the 23 public hospitals (IPS), none remained on the efficient frontier during the 5 years of the study.
- The DEA method identifies efficient IPS and inefficient IPS.
- The results probably show that the covid 19 pandemic affected efficiency by decreasing it in all IPS.
- The costs of the general social security system increase, including in pandemic years.
- Efficiency increased during the first three years of the study, decreased in the last two years.

- The study can help understand the impact of efficiency and inefficiency on some important variables in health care processes in public clinics and hospitals.

References

1. Pan American Health Organization, OPS Washington, D.C. (2018): "Health Indicators Conceptual and operational considerations ", CIP data are available at http://iris.paho.org, páginas 29-43.
2. Charnes, A. W., Cooper W.W., y Rhodes, E., 1978): "Measuring Efficiency of Decision Making Units", European Journal of Operational Research, páginas 429-222.
3. Cordero José M , García-García Agustin , Lau-Cortés Enrique and Polo Cristina. "Assessing Panamanian hospitals' performance with alternative frontier methods". International Transactions in Operational Research published by John Wiley & Sons Ltd on behalf of International Federation of Operational Research Societies (2021), Res. 30 (2023) 394–420.
4. Banker R. D.; Charnes, A.; Cooper, W. W. (1984). "Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis". Management and Science, vol. 30, pp.1078-1092.
5. Pan American Health Organization, OPS Washington, D.C. (2018): "Health Indicators Conceptual and operational considerations ", CIP data are available at http://iris.paho.org, páginas 29-43.
6. Superintendencia Nacional de Salud (2022): "Catálogo de información financiera con fines de supervisión – Entidades clasificadas en NIIF Grupo 1, 2 y 3 para las instituciones prestadoras de servicios de salud". https://www.supersalud.gov.co/es-co/nuestra-entidad/cifras-y-estadisticas.
7. Tarazona, V.C., Guadalajara-Olmeda, N.; Vivas-Consuelo, D. (2019): "Predicting healthcare expenditure by multimorbidity groups". Health Policy, 123, 427–434, doi:10.1016/j.healthpol.2019.02.002.
8. Prada-Ríos SI, Pérez-Castaño AM, Rivera-Triviño AF. (2017) "Clasificación de instituciones prestadores de servicios de salud según el sistema de cuentas de la salud de la Organización para la Cooperación y el Desarrollo Económico: el caso de Colombia". Rev Gerenc Polít Salud; 16 (32): 51-65. https://doi.org/10.11144/Javeriana.rgps16-32.cips.

Balanced models from unbalanced data: an illustrative case in cardiovascular risk

Beatriz de Otto, Ignacio Pedrosa, Pelayo Quirós and Jimena Pascual

CTIC Technology Centre. W3C LATAM Chapter, Gijón, Spain
ignacio.pedrosa@fundacionctic.org

Abstract. Artificial Intelligence plays a crucial role in making impactful decisions for individuals and society. The effectiveness of AI and machine learning heavily relies on the quality of the training data. Hence, ensuring fairness becomes vital during the training process itself. The presence of class imbalance is a common issue in clinical data analysis, leading to datasets with extreme class imbalance that affects classifier performance. This paper introduces an innovative method to select a threshold that minimizes the difference between sensitivity and specificity in classifiers, using a risk estimator for ischemic heart disease as a case study. Data from the Behavioral Risk Factor Surveillance System survey is used, comprising approximately 400,000 respondents, with the positive class representing about 10% of the surveyed population.
The proposed approach involves an alternative strategy for handling class imbalance by identifying an appropriate classification threshold based on the intensity of the imbalance, rather than using the standard 0.5 threshold. This new threshold selection criterion provides a clear and explicit way to handle type I and type II errors, assigning equal importance in the absence of expert knowledge about their relative costs. Additionally, analyzing score distributions separated by classes aids result interpretation and offers valuable context to end users.
In summary, this research emphasizes the importance of integrating fairness into AI training processes and proposes a threshold selection method tailored to class imbalance intensity. The approach demonstrates promising results in addressing the challenges of imbalanced data, particularly in the context of healthcare risk assessment.

Keywords: AI-based models, fairness, imbalanced-data, healthcare.

1 Introduction

1.1 Bias and fairness in AI-based systems

Artificial Intelligence (AI)-based systems are nowadays extensively utilized to make decisions that hold far-reaching implications for individuals and society as a whole. The impact of these decisions can affect everyone, anywhere, and at any time, raising concerns about potential human rights issues. Related to this issue, AI and machine learning are limited by the quality of data on which they are

trained. The generalizability of AI algorithms across subgroups is critically dependent on factors like representativeness of included populations, missing data, and outliers. The process by which the data are generated may be more important and particular to AI. If AI algorithms use data that are generated through a biased process, then the output may be similarly biased producing a result that differs from the true underlying estimate. This is a significant challenge when using clinical data sources [1]. Consequently, it becomes imperative to transcend conventional AI algorithms solely optimized for predictive performance and instead incorporate ethical principles into their design, training, and deployment. By doing so, we can ensure the promotion of social good while still harnessing the immense potential of AI technology. Considering potential biases, this is not a new problem rather "*bias is as old as human civilization" and "it is human nature for members of the dominant majority to be oblivious to the experiences of other groups.*" [2]. Nonetheless, the application of AI-based decision-making has the capacity to amplify pre-existing biases and introduce novel classifications and criteria, carrying significant potential for new types of biases. The rising apprehensions surrounding this issue have prompted a reassessment of AI-based systems, advocating for new approaches that prioritize the fairness of their decisions. Therefore, fairness should be integrated into the training process itself [3]. In addition to this potential statistical bias, it is essential to take into account social biases as far as could be caused by a statistically biased algorithm or by other human factors, including implicit or explicit bias. For example, clinicians may incorrectly discount the diagnosis of myocardial infarction in older women because these patients are more likely to present with atypical symptoms. At this point, one of the critical aspects linked to both data quality and human factors that may include biases is data imbalance and the techniques employed for its handling.

1.2 The common challenge of imbalanced data

Many real-world domains are, by definition, class imbalanced by virtue of having a majority class that naturally has many more instances than its minority class. Subsequently, class imbalance in classification models is also a common issue within clinical data analysis [4]. These phenomena lead to datasets with extreme levels of class imbalance, which impacts the performance of the classifiers. Models trained with such datasets will exhibit bias towards the most prevalent majority classes because of their higher prior probabilities, while often ignoring the minority classes. However, despite being a subject of interest for over two decades, it remains a profound area of research aimed at achieving improved accuracy [5]. Regarding machine learning, class imbalance techniques are typically categorized into two groups: data-level techniques and algorithm-level methods. Data-level techniques concentrate on manipulating the distribution of the training dataset to lighten the imbalance innate in the original data. The two fundamental paradigms within this category are: (1) random oversampling (ROS), which duplicates samples from minority classes, and (2) random undersampling (RUS), which discards samples from majority classes. While the simplicity and

efficiency of combining ROS and RUS methods might appear attractive, when featuring extreme levels of class imbalance, repeatedly oversampling from the same data within the minority classes can lead the model to memorize irrelevant features that may hold no actual utility for those respective classes [6]. Some of the most frequently mentioned drawbacks of these methods for correcting imbalances in the training sample include a) the loss of useful information associated with the observations discarded in the case of under-sampling; and b) the risk of overfitting the model to irrelevant features of the observations from the minority class in the case of oversampling, which becomes excessively homogeneous as a result of replicating a small number of available observations. The SMOTE technique [7] offers a way to avoid excessive homogeneity by generating artificial observations through combining features from available records for the minority class, with the threat that these new instances may be unrealistic. Another challenge arises with these strategies as far as the model requires testing and validation on a test sample exhibiting the same level of class imbalance as observed in the phenomenon being modeled. This guarantees that performance metrics can be taken as estimates of the model's real-world performance when deployed on actual data. Rather than using the common approaches of undersampling or oversampling, identifying an appropriate threshold corrects the effect of class imbalance without disrupting the training sample. In this paper we consider the potential negative outcomes from poor data quality on healthcare and the effect when applying an innovative strategy to tackle class imbalance by replacing the standard 0.5 threshold. Thus, the challenge addressed is to develop a model capable of identifying significant risks of ischemic heart disease based on incomplete and non-clinical detailed information. The objective is to identify individuals early on and prompt them to seek expert diagnostic evaluation from healthcare services. The developed model aims to estimate the risk of ischemic heart disease (i.e. heart attack and angina pectoris). To reach this goal, within this paper an alternative method is propose to choose the threshold that minimises the difference between sensitivity and specificity of a classifier using an estimator of risk of ischemic heart disease as use case because: 1) in the absence of expert medical knowledge that allows an informed assessment of the costs associated with each type of error, the chosen threshold assigns the same importance to false positives and false negatives, and 2) when sensitivity and specificity are very similar, the overall accuracy rate, also takes a very close value to both, reducing the disparity between the most common performance metrics evaluation and facilitating the comparison between different specifications of the same model.

2 Method

2.1 Participants

This method has been applied to a model for predicting the risk of ischemic heart disease (i.e. heart attack and angina) in the US population using low-quality data from almost 400,000 people surveyed within the Behavioral Risk Factor Surveillance System survey. Positive class (i.e. individuals who reported

having experienced a heart attack or angina pectoris) accounts approximately for 10% of the participants, while also highlighting significant differences between genders.

Considering this type of sampling, the sample exhibits various biases that impact the quality of the data: 1) survivor bias, because it does not include those who died from cardiovascular accidents, 2) protopathic bias, as far as only information on lifestyle after the diagnosis of heart disease is recorded, which often is modified as a result of the ischemic episode or its sequelae [8], and 3) a bias associated with only knowing the age of the respondent at the time of the survey, and not when the coronary accident happened, which makes it impossible to adopt an actuarial approach to the problem.

2.2 Procedure

The dataset used in this study was collected through the Behavioral Risk Factor Surveillance System survey (US). This telephone-based survey focuses on public health and gathers non-specialized information regarding diagnosed diseases, subjective perception of the respondent's health status, dietary habits, exercise routines, and more. However, it does not provide information about clinical tests (e.g. blood pressure, blood analysis, etc.).

2.3 Data Analysis

Firstly, an exploratory analysis was conducted to analyse the data distribution. After identifying the need of established separated models by sex, several trained classifiers were applied including: logistic regression, K-Nearest Neighbors, Stochastic Gradient Descend (SGD), decision trees, random forest, Adaptive Boosting (Adaboost), Gradient Tree Boosting (with tree and linear booster), Extreme Gradient Boosting and Category Boosting, with and without adjustment of their internal weights for algorithms that allow it, and with the selection of their hyperparameters through grid search in a cross-validation process.

Subsequently, two alternative strategies were adopted. Firstly, the adjustment of the internal weights of the classification models that allow for this correction to compensate for the effect of imbalance. Secondly, the search for specific classification thresholds for each trained model so that they produce balanced metrics in the validation set.

3 Results

The exploratory analysis shown the convenience of developing separated models by sex, so that the weight of the explanatory factors may be different. This was confirmed by the results of a trial and error process of different specifications of the tested classification models (Fig 1).

Women and men also show different patterns in terms of the explanatory factors of ischemic disease and the intensity of correlations. As shown in Figure

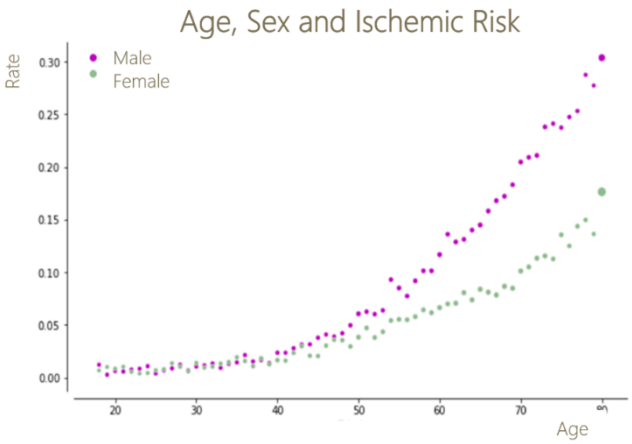


Fig. 1: Relative frequency of ischemic heart disease in women and men according to age

1, in both cases the risk of ischemic heart disease increases, but notably more sharply for men, especially from the age of 50.

Among the explanatory variables, sex stands out because the rest of the factors have a different effect for men and women. Therefore, separate models have been trained.

The remaining explanatory variables were selected through an iterative process of trying different combinations of regressors. The chosen ones include the factors identified by medical science as predictors of cardiovascular risk: age, diabetes, tobacco, overweight and alcohol consumption, etc., and some others that managed to improve the quality of the results in the trials for the selection of regressors: self-reported health status, diagnoses of arthritis, gout, lupus, and fibromyalgia, stroke history, COPD and regular medical follow-up.

Several imbalance handling strategies have been tested: undersampling, adjustment of the internal weights of the algorithms that offer this possibility, and classification threshold tuning. On one hand, addressing the imbalance in the training subset leads to models that do not generalize effectively when evaluated on the original imbalanced population. On the other hand, internal weights of the models were adjusted to balance the influence of the two classes during the training process. The latter are those identified with the prefix '*b_*' in Table 1. The results obtained are show in Table 1.

Finally, several criteria for selecting the classification threshold were tested. The imbalance handling method that yields the best results for this problem, which also stems from a dataset with modest predictive capacity, is the threshold selection that minimizes the difference between recall and specificity.

The undersampling strategy is characterized by artificially reducing the size of the negative observation subset so that the two classes have the same representation in the training set. The disadvantage of this procedure is that the

error rates obtained in the rebalanced subset are not generalizable to the general population and the model thus trained fails when faced with the imbalance that characterizes the population of inter-est. Indeed, when the negative class is resized to restore its relative size with respect to the positive class, true and false positives are expected to increase in a similar proportion. When the imbalance is intense and the false positive rate is high (as happens when the available data are not informative enough, as is the current use case), the proportion of false positives in the total population significantly increases, sinking the precision metric. Furthermore, if the specificity is lower than the recall, accuracy is also reduced.

The adjustment of internal weights produces, in most cases, more balanced models regarding recall and specificity. The effect of this adjustment varies depending on the algorithm used, and we have no control over the handling of different types of error. This is shown in the associated ROC curves (Fig 2): for models without internal weight adjustment, the standard 0.5 threshold offers specificities close to 1 and recall close to 0; while in models with weight adjustment, the same threshold provides results closer to the diagonal of the ROC curve, on which specificity and recall are equal.

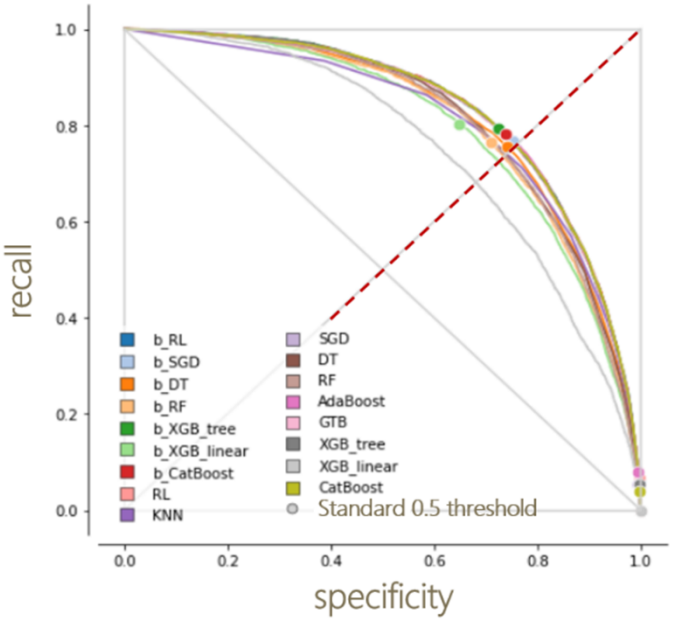


Fig. 2: ROC curves and 0.5 threshold for the trained models

As for the selection of the classification threshold, we propose choosing the one that minimizes the difference between recall and specificity. This implies that without expert information regarding the costs of Type I and Type II errors, we

Table 1: Metrics of models at 0.5 and optimal thresholds.

	umbral	accuracy	recall	especificidad	precision	roc_auc_score
b_RL	0.5	0.933486	0.060117	0.996435	0.548589	0.834778
b_SGD	0.5	0.753320	0.766747	0.752352	0.182442	0.834940
b_DT	0.5	0.741842	0.755067	0.740888	0.173577	0.820081
b_RF	0.5	0.712996	0.763312	0.709369	0.159169	0.810202
b_XGB_tree	0.5	0.728608	0.792168	0.724027	0.171424	0.836149
b_XGB_linear	0.5	0.659669	0.802130	0.649401	0.141558	0.800079
b_CatBoost	0.5	0.742280	0.781175	0.739477	0.177712	0.836061
RL	0.5	0.933532	0.066644	0.996014	0.546479	0.834741
KNN	0.5	0.932608	0.054277	0.995915	0.489164	0.809805
SGD	0.5	0.933347	0.046719	0.997252	0.550607	0.834820
DT	0.5	0.932770	0.000000	1.000000	0.000000	0.818856
RF	0.5	0.932770	0.000000	1.000000	0.000000	0.814663
AdaBoost	0.5	0.933416	0.078667	0.995023	0.532558	0.835095
GTB	0.5	0.933162	0.040536	0.997499	0.538813	0.834509
XGB_tree	0.5	0.933162	0.052216	0.996657	0.529617	0.836015
XGB_linear	0.5	0.932770	0.000000	1.000000	0.000000	0.752511
CatBoost	0.5	0.933162	0.038131	0.997673	0.541463	0.835377

	umbral	accuracy	recall	especificidad	precision	roc_auc_score
b_RL	0.119527	0.758762	0.747578	0.760113	0.273558	0.836142
b_SGD	0.525372	0.759092	0.746048	0.760668	0.273611	0.836117
b_DT	0.542192	0.735149	0.736614	0.734972	0.251414	0.816084
b_RF	0.546458	0.732813	0.724120	0.733863	0.247430	0.807908
b_XGB_tree	0.561091	0.759614	0.745028	0.761377	0.273929	0.837613
b_XGB_linear	0.556284	0.732126	0.723865	0.733124	0.246848	0.806106
b_CatBoost	0.556861	0.760906	0.746048	0.762701	0.275311	0.838636
RL	0.119527	0.758762	0.747578	0.760113	0.273558	0.836142
KNN	0.150000	0.758212	0.722081	0.762578	0.268742	0.820667
SGD	0.121526	0.759174	0.747833	0.760545	0.273984	0.836276
DT	0.145282	0.752467	0.739674	0.754013	0.266514	0.826702
RF	0.145269	0.743423	0.734064	0.744554	0.257744	0.815784
AdaBoost	0.479838	0.758789	0.744518	0.760514	0.273076	0.835251
GTB	0.121084	0.744963	0.729985	0.746773	0.258347	0.820249
XGB_tree	0.137267	0.760136	0.744008	0.762085	0.274248	0.838363
XGB_linear	0.138317	0.713241	0.710607	0.713559	0.230636	0.783910
CatBoost	0.127745	0.760934	0.748343	0.762455	0.275716	0.838871

treat both types of errors with equal importance. Geometrically, the threshold as close as possible to the diagonal of the ROC curve is chosen. This criterion is especially suitable when there is no expert knowledge to assess the relative cost of the two types of error: not detecting a patient at high risk or unnecessarily alarming a patient at low risk. In geometric terms, the aim is to find the threshold closer to the diagonal of the ROC curve.

One advantage of this threshold is that when recall and specificity are approximately equal, the accuracy rate is also nearly the same . This desirable effect means that, in such a scenario, the precision rate depends only on the accuracy and the extent of the imbalance. The greater the accuracy and the less the intensity of the imbalance, the greater the precision. Formally:

$$\text{Precision} = \frac{\text{sensitivity} \cdot \#\text{ positive obs.}}{\text{sensitivity} \cdot \#\text{ positive obs.} + (1 - \text{specificity}) \cdot \#\text{ negative obs.}} \tag{1}$$

$$\sim \frac{\text{accuracy} \cdot \#\text{ positive obs.}}{\text{accuracy} \cdot \#\text{ positive obs.} + (1 - \text{accuracy}) \cdot \#\text{ negative obs.}} = \frac{1}{1 + \frac{1-\text{accuracy}}{\text{accuracy}} \cdot \frac{\#\text{ negative obs.}}{\#\text{ positive obs.}}}. \tag{2}$$

These two properties greatly facilitate the comparison of performance metrics from different models, eliminating large disparities between those that offer high specificity with low recall and vice versa. On the one hand, the three metrics representing success rates in the population (accuracy) or in a subpopulation (the positive class in the case of recall and the negative class in the case of specificity) are very similar for each model. On the other hand, the precision rate can be disregarded, which will be given by the accuracy of each model and the intensity of class imbalance, common to all models. In summary, the choice of one model over another becomes a univariate decision problem (except for irreducible distances in metrics).

4 Discussion

AI is making its way into clinical practice. AI can create or perpetuate biases that may worsen patient outcomes. However, by strategically deploying AI and carefully selecting underlying data, algorithm developers can mitigate AI bias. Addressing bias could allow AI to reach its fullest potential by helping to improve diagnosis and prediction while protecting patients. Additionally, clinicians may have a propensity to trust suggestions from AI decision support systems, which summarize large numbers of inputs into automated real-time predictions, while inadvertently discounting relevant information from non-automated systems —so-called automation complacency.

Although much of the discussion about AI and bias has focused on its potential for harm, strategies exist to mitigate such bias. In this vein, the results

allow us to identify a novel method aimed at mitigating the potential negative effects of imbalanced data in a specific challenge, such as the one presented in this study around risk estimation of ischemic heart disease.

The models evaluated, each using a threshold with this property, present balanced metrics that simplify their comparison and aid in selecting the most appropriate one. The results clearly show a stable inverse relationship between accuracy and precision.

As for the utility of the model for the end user, we are particularly concerned that the information associated with a specific prediction is comprehensible. Typically, the predictions of a classification model are offered in two forms: the binary prediction (positive or negative, in this case, associated with a high or low risk of ischemic heart disease respectively) or a numerical prediction ranging between 0 and 1, which is usually interpreted as a measure of the probability that the example belongs to the positive class. In this use case, this information offered in isolation is not sufficient by itself, as the classification threshold is much lower than the standard 0.5 that is usually adopted as a reference to separate positive and negative predictions. Moreover, the distribution of these scores for the sample population varies greatly between models and, with them, the ideal threshold in each case. However, the distribution of scores for the selected model provides useful contextual information for the user. Specifically, it allows them to identify their relative risk referred to that associated with the general population and to the positive and negative classes separately.

In particular, for example, for a user with a score equal to the classification thresh-old of the chosen model, the proportion of observations of the negative class with a score lower than theirs is equal to the proportion of observations of the positive class with a score higher. That is, there is the same proportion of low-risk population with a score below theirs as there is of high-risk population with a score higher. In the rest of the cases, the proportions of low and high risk population with scores below and above their own provides a useful context for the user to assess the level of relative risk associated with their situation.

In summary, the analysis of the distribution of scores provided by the model offers contextual information that complements other interpretability strategies for black box models that are also applicable in this case: Shap values for the user-declared features, the distribution of these values in the population, permutation-based methods, etc.

Conclusions are specific for models characterized by a predictive capacity limited by the poor quality of the available data and a strong class imbalance, two features that characterize research in healthcare. In this scenario, imbalance treatment strategies based on correcting the training subsample do not generalize effectively to the population.

The proposed threshold selection criterion is characterized by providing a clear and explicit criterion for the treatment of type I and type II errors, to which, in the absence of other expert knowledge about their relative costs, the same importance is attributed. In addition, the analysis of score distributions

separated by classes facilitates the interpretation of results and provides additional context to the end user.

In conclusion, the adoption of AI is poised to bring about a paradigm shift in healthcare, where its performance in medical tasks, including clinical diagnosis. AI holds substantial potential in healthcare, empowering clinicians to deliver more precise and timely diagnoses and design effective treatment strategies. Nonetheless, the legitimate concern of exacerbating pre-existing healthcare disparities with the implementation of AI models remains. Regrettably, in numerous diagnostic and prognostic clinical applications, the "ground truth" utilized for fairness assessment metrics might already encompass inherent biases and be intertwined with suboptimal outcomes that cannot be solely attributed to clinical features. Consequently, the medical AI community must transcend the mere evaluation of AI models' clinical readiness using methods and metrics reliant on potentially biased and continuously evolving clinical ground truth.

Acknowledgment

This paper was conducted within the framework of the IBERUS project. Technological Network of Biomedical Engineering applied to degenerative pathologies of the neuromusculoskeletal system in clinical and outpatient settings (CER-20211003), CERVERA Network financed by the Ministry of Science and Innovation through the Center for Industrial Technological Development (CDTI), charged to the General State Budgets 2021 and the Recovery, Transformation and Resilience Plan.

References

1. Parikh, R. B., Teeple, S., & Navathe, A. S.: Addressing Bias in Artificial Intelligence in Health Care. JAMA (2019). doi:10.1001/jama.2019.18058
2. Fortune, http://fortune.com/longform/ai-bias-problem/, last accessed 2023/25/07.
3. Mehrabi, N., Morstatter, F., Saxena. N., Lerman, K., Galstyan. A.: A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54(6), Article 115 (2022). https://doi.org/10.1145/3457607
4. Tasci E., Zhuge Y., Camphausen K., Krauze AV.: Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets. Cancers, 14(12), 2897 (2022). https://doi.org/10.3390/cancers14122897
5. Sharma, S., Gosain, A., Jain, S.: A Review of the Over-sampling Techniques in Class Imbalance Problem. In: Khanna, A., Gupta, D., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds.) INTERNATIONAL CONFERENCE ON INNOVATIVE COMPUTING AND COMMUNICATIONS. ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING, vol. 1387. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-2594-7_38
6. De Angeli, K., Gao, S., Danciu, I., Durbin, E. B., Wu, X. C., Stroup, A., Doherty, J., Schwartz, S., Wiggins, C., Damesyn, M., Coyle, L., Penberthy, L., Tourassi, G. D., & Yoon, H. J.: Class imbalance in out-of-distribution datasets: Improving

the robustness of the TextCNN for the classification of rare cancer types. Journal of biomedical informatics 125, 103957 (2022). https://doi.org/10.1016/j.jbi.2021.103957

7. Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905 (2018).
8. Hernández-Avila, M., Garrido, F., Salazar-Martínez, E.: Sesgos en estudios epidemiológicos. Salud Pública de Mexico 42(5), 438-446 (2000). https://scielosp.org/pdf/spm/2000.v42n5/438-446/es.

Part V

Recent advances in iterative processes for solving nonlinear problems

Computing Lyapunov exponents for the study of the dynamical behaviour of Chebyshev's method on polynomials

Víctor Álvarez-Aparicio[1], José Manuel Gutiérrez-Jiménez[1], Luis Javier Hernández-Paricio[1], and María Teresa Rivas-Rodríguez[1]

Universidad de La Rioja, Depto. de Matemáticas y Computación, Edificio CCT, C/Madre de Dios, 53, Logroño, 26006, Spain,
valvarezaparicio@gmail.com

Abstract. It is well known (see [3], for instance) that Chebyshev's method presents some interesting dynamical properties, like superattracting n-cycles or superattracting extraneous fixed points. The main focus of this work, framed in the context of polynomial root-finding algorithms, will be to give a numerical and graphical study of the dynamical behaviour of Chebyshev's method based on the computation of the Lyapunov exponents of the discrete dynamical system induced by the iteration of a rational map.
In [1], a novel method to compute the basins of attraction induced by the iteration of a rational map was introduced and implemented. The present work can be considered as a continuation of this previous article. Also, the implementation in Julia Language (which can be found in [2]) of the algorithms that allows us to compute the Lyapunov exponents and the basins of attraction induced by a rational map solves some frequent computational problems, like overflows or indeterminations.

Keywords: Chebyshev's method, Lyapunov exponents, dynamical study, polynomials, root-finding algorithms, numerical methods

1 Introduction

Chebyshev's method is, to this day, one of the most used and studied root-finding numerical methods for non-linear equations. In this work, which can be considered as a continuation to [1], we undertake a dynamical study of this method when applied to cubic polynomials. In order to do this, we will consider the novel techniques and algorithms presented in [1] to compute the Lyapunov constants associated with the rational map induced by Chebyshev's method when applied to a polynomial. By studying these Lyapunov constants (closely related with the well-known Lyapunov exponents of the discrete-time dynamical system induced by the iteration of the rational map) we will be able to extract useful information regarding the dynamics of Chebyshev's method for the particular case of cubic polynomials, as we will observe in the following sections. It is important

to note that the techniques we will consider for our study are topological and geometrical in nature, involving broadly-studied constructions such as the Hopf fibration $S^3 \to P^1(\mathbb{C})$ or the complex projective line $P^1(\mathbb{C})$. In this proceeding, we will not delve into the technical details of our considered framework, since every detail can be checked in [1].

It is also worth mentioning that the algorithms used for this study are implemented in Julia Language and are able to avoid some computational problems that often arise in Numerical Analysis, such as overflows, underflows and mathematical indeterminations. The implementation of the algorithms, along with a user guide and many examples, can be found in the GitHub repository [2].

2 Theoretical framework

As we have already stated, for the sake of simplicity we will not delve into the technicalities behind the considered theoretical framework, and since every detail can be found in [1]. However, there are some important observations to be made regarding the way our algorithms represent rational maps and how do we compute its associated Lyapunov constants.

In order to avoid some computational problems like overflows or indeterminations, we represent each rational map $f : P^1(\mathbb{C}) \to P^1(\mathbb{C})$ defined over the Riemann sphere (in particular, over the complex projective line $P^1(\mathbb{C})$) as an irreducible r-homogeneous pair of bivariate polynomials $R = (F, G)$ (see [1]). We consider the usual 3-sphere $S^3 = \{(z, t) \in \mathbb{C}^2 \mid |z| + |t| = 1\}$, which we take as a subspace of $\mathbb{C}^2$. Given the following commutative diagram induced by the Hopf fibration $S^3 \to P^1(\mathbb{C})$

$$\begin{array}{ccc} \mathbb{C}^2 \setminus \{(0,0)\} & \xrightarrow{R} & \mathbb{C}^2 \setminus \{(0,0)\} \\ \downarrow p & & \downarrow p \\ S^3 & \xrightarrow{R^S} & S^3 \\ \downarrow q & & \downarrow q \\ P^1(\mathbb{C}) & \xrightarrow{R^P} & P^1(\mathbb{C}) \end{array}$$

one has that the pair $R = (F, G)$ (and consequently, the rational map f) can be studied through its associated Hopf-endomorphism (R^S, R^P). This way of representing a rational map f is very suitable for a computational environment, and allows our algorithms to avoid the mentioned computational problems, and to compute the basins of attraction induced by the iteration of f over the whole Riemann sphere $\hat{\mathbb{C}} \cong P^1(\mathbb{C})$, including the infinity point ∞.

We will also consider the spherical derivative of a rational map $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, instead of the usual notion of derivative, in order to define the standard notions of super-attracting, attracting, indifferent and repulsive fixed points and cycles.

The spherical derivative $f^{\#} : \hat{\mathbb{C}} \to \mathbb{R}_+ = \{r \in \mathbb{R} \mid r \geq 0\}$ of a rational map f is given by

$$f^{\#}(z_0) = |f'(z_0)| \frac{1 + |z_0|^2}{1 + |f(z_0)|^2}, \tag{1}$$

where $f'(z_0)$ denotes the usual derivative of f in $z_0 \in \hat{\mathbb{C}}$, for every $z_0 \in \hat{\mathbb{C}}$.

Note that if f' is not defined in z_0, we can define $f^{\#}(z_0)$ as the limit of the expression above when $z \to z_0$, and it can be proven that this limit always exist, $f^{\#}$ is well defined and it is continuous. Also, since $f^{\#}$ is continuous and $\hat{\mathbb{C}}$ is compact, $f^{\#}$ is bounded.

Hence, we will say that $z_0 \in \hat{\mathbb{C}}$ is a critical point of f if $f^{\#}(z_0) = 0$, and we will also consider the usual notions of super-attracting, attracting, indifferent and repelling points and cycles, using the spherical derivative $f^{\#}$.

We consider this particular notion of the derivative of a rational map since a computationally-suitable expression of the spherical derivative of a rational map defined over the complex projective line $P^1(\mathbb{C})$ can be proven (see [1]) and this notion is also useful when addressing the connection of the presented methodology with Lyapunov exponents (as it was shown in [1]).

As a final note on the theoretical framework considered for this dynamical study, we will briefly introduce the notions that allow us to make calculations regarding the basins of attraction induced by the iteration of a rational function and the Lyapunov exponents of the induced discrete-time dynamical system.

Let X be a topological space, and let $f : X \to X$ and $\phi : X \to \mathbb{R}_+$ be continuous maps. We define the function $L_f(\phi) : X \to [0, +\infty]$ given by the expression

$$L_f(\phi)(x) = \lim_{n \to +\infty} \left(\prod_{k=0}^{n-1} \phi(f^k(x)) \right)^{\frac{1}{n}}, \tag{2}$$

which we will call the Lyapunov function of f associated with ϕ.

It is important to note that Lyapunov functions are used frequently in the context of Dynamical Systems to study local stability. In our case, despite that the function $L_f(\phi)$ is not a Lyapunov function in that context, we will also call it a Lyapunov function, since it will also be used to study local stability and dependency on initial conditions. Of course, in order to be able to compute this function, we will work with finite approximations.

As it was shown in [1], these Lyapunov functions, under certain conditions (that hold for our study), are constant in each basin of attraction induced by the iteration of the map of f. We will refer to these constants as the Lyapunov constant associated with each basin of attraction. This way, by computing these Lyapunov constants, we are able to distinguish between the different basins of attraction induced by f.

For our study we will consider $X = P^1(\mathbb{C})$, f a rational map (induced by Chebyshev's method applied to a cubic polynomial) represented by its induced Hopf-endomorphism, and $\phi = f^{\#}$ the spherical derivative of f (defined over $P^1(\mathbb{C})$). In this particular case, the logarithm of a Lyapunov constant is precisely

a Lyapunov exponent of the discrete-time dynamical system induced by the iteration of the rational map f (see [1]).

3 Dynamical study of Chebyshev's method on cubic polynomials

3.1 Preliminaries

Our main goal in this study is to analyze the dynamics that Chebyshev's method induces when applied to a cubic polynomial, using the previously mentioned theory about Lyapunov functions and Lyapunov constants. In this regard, different algorithms have been implemented in Julia Language (see the GitHub repository [2]) to compute the Lyapunov constants of an arbitrary rational map, and to construct Chebyshev's parameter plane using the notions that make up our theoretical framework.

We recall that, if $f(z)$ is a differentiable function, Chebyshev's method applied to f is given by the iteration of the induced map:

$$C_f(z) = z - \left(1 + \frac{1}{2}L_f(z)\right)\frac{f(z)}{f'(z)}, \tag{3}$$

with $L_f(z) = \frac{f(z)f''(z)}{(f'(z))^2}$. It is important to address that, despite the notation, in the previous expression the operator L_f is not related in any way to the described Lyapunov functions, and it is used simply to define Chebyshev's method in a more tractable way. Also, note that if $p(z)$ is a polynomial, $C_p(z)$ is a rational map that has the roots of p as superattracting fixed points.

In order to undertake a dynamical study of these characteristics, we will be using two well-known theorems: The Scaling Theorem for Chebyshev's method, and Fatou's Theorem.

The first theorem states the following: Let $T(z) = \alpha z + \beta$, where $\alpha \neq 0$, be an affine map in $\mathbb{C}$, and let $\lambda \in \mathbb{C} \setminus \{0\}$. Let $p(z)$ be a complex polynomial, and define $q(z) = \lambda(p \circ T)(z)$. Then, the rational maps C_p and C_q induced by Chebyshev's method are conjugate by T; that is, $C_p \circ T = T \circ C_q$. This essentially tells us that if two polynomials differ only by an affine map and a complex nonzero constant, then they share the same dynamics. Therefore, in order to study the dynamics of every cubic polynomial, it suffices to restrict ourselves to the family of polynomials of the form $p_\lambda(z) = (z - \lambda)(z^2 - 1)$ (since one can always fix two of the roots of a cubic polynomial to the points 1 and -1 through an affine map).

Fatou's Theorem states that, if f is a rational map then, if f has some attracting k-cycle, then at least one of its critical points converges to it. Therefore, in order to study whether if a rational map presents attracting cyclic behaviour or some kind of attracting extraneous fixed point under iteration, it suffices to

study the orbits of its critical points. As it was remarked in [3], the critical points of the function C_p induced by Chebyshev's method are precisely the roots of

$$(C_p)'(z) = \frac{(3 - L_{p'}(z))(L_p(z))^2}{2}, \tag{4}$$

where $L_{p'}(z) = \frac{p'(z)p'''(z)}{(p''(z))^2}$. Moreover, in the particular case that p_λ is a polynomial of the form $p_\lambda(z) = (z - \lambda)(z^2 - 1)$, we have that (see [3]) for almost every value of λ the only free critical points (that is, critical points which are not the roots of p_λ) are

$$\gamma_1(\lambda) = \frac{5\lambda - \sqrt{-5(\lambda^2 + 3)}}{15} \quad \text{and} \quad \gamma_2(\lambda) = \frac{5\lambda + \sqrt{-5(\lambda^2 + 3)}}{15}. \tag{5}$$

Hence, to study the dynamics of Chebyshev's method when applied to a polynomial of the form $p_\lambda(z) = (z - \lambda)(z^2 - 1)$ (for almost every value of λ) it suffices to study the orbits of the points γ_1 and γ_2 under the iteration of C_{p_λ}.

3.2 Algorithm

In order to study the dynamics of Chebyshev's method when applied to cubic polynomials we will consider the following procedure:

1. Since every cubic polynomial share the same dynamics (under Chebyshev's method) with a polynomial of the form $p_\lambda(z) = (z - \lambda)(z^2 - 1)$, we will construct a grid of complex points λ (a parameter plane), each of which represents the polynomial p_λ.
2. We will apply Chebyshev's method to each polynomial p_λ in the grid, obtaining a rational map C_{p_λ}. Then, we will use the algorithms implemented in [2] and described in [1] in order to study the orbits of the free critical points γ_1 and γ_2 (using Lyapunov constants).
3. If the orbits of both γ_1 and γ_2 converge to a super-attracting fixed point of C_{p_λ} (that is, a root of p_λ), we will assign a color to the point λ, depending on which root each orbit converges to (note that there are 9 possibilities). If either the orbit of γ_1 or the orbit of γ_2 converges to an attracting k-cycle, we will assign λ a different color, and another one if some orbit converges to an attracting extraneous fixed point. Then, we will graphic the numerical results of our considered parameter plane.

3.3 Graphical results

In this last section we present some graphical results obtained by the described procedure.

In Figure 1 the region $[-2.5, 2.5] \times [-2.5, 2.5]$ of our parameter plane is visualized. As we can see, in the plot on the left 12 different colors appear. The color 0 (red) corresponds to regions for which the polynomials p_λ present some kind of attracting k-cyclic behaviour when Chebyshev's method is applied. The

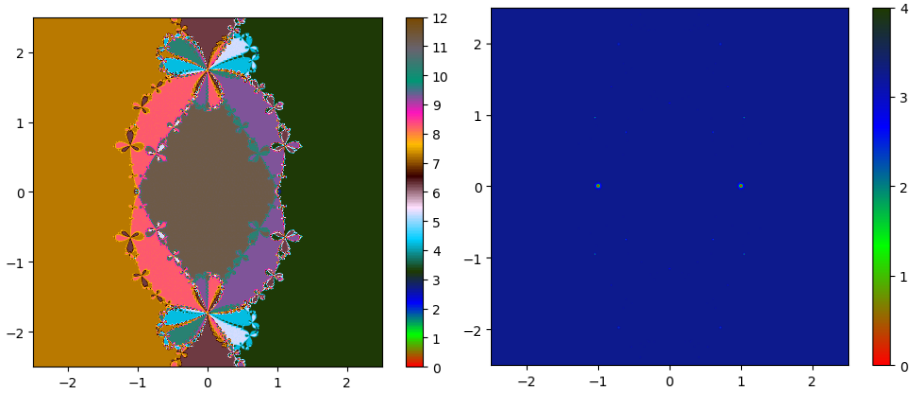


Fig. 1: Parameter plane of Chebyshev's method on cubic polynomials

color 2 (blue) corresponds to regions for which the induced rational map C_{p_λ} presents an attracting extraneous fixed point. The color 1 (green) corresponds to regions for which the induced rational map C_{p_λ} have both attracting cyclic behaviour and an attracting extraneous fixed point. Every other color corresponds to regions for which the orbits of the two free critical points γ_1 and γ_2 of C_{p_λ} converge to one of the 3 roots of p_λ, distinguishing the 9 possibilities. In the plot on the right the exact same thing is depicted, but in this case the 9 possibilities of the orbits of the free critical points converging to one of the roots of the polynomial are represented by the same color 3 (dark blue), in order to visualize more clearly those regions in which cyclic behaviour appears. As one can observe, in the considered rectangle of the parameter plane, these regions are relatively small.

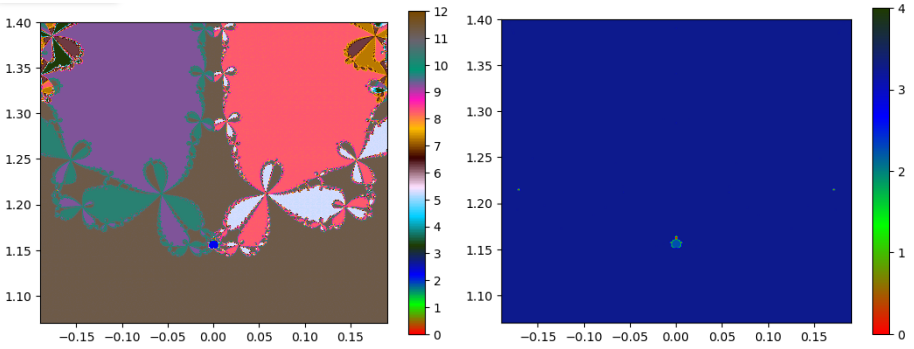


Fig. 2: Parameter plane of Chebyshev's method on cubic polynomials

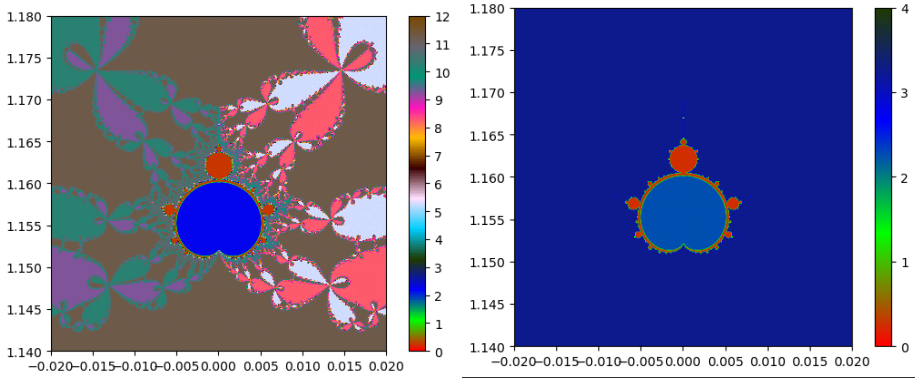


Fig. 3: Parameter plane of Chebyshev's method on cubic polynomials

In Figures 2 and 3 we can observe with more detail one of the regions for which cyclic behaviour appear. In particular, we see that the main cardioid of the Mandelbrot-like set in the center of each plot corresponds to polynomials that present an attracting extraneous fixed point under the iteration of Chebyshev's method, and the smaller bulbs of this set correspond to polynomials that present different kinds of cyclic behaviour.

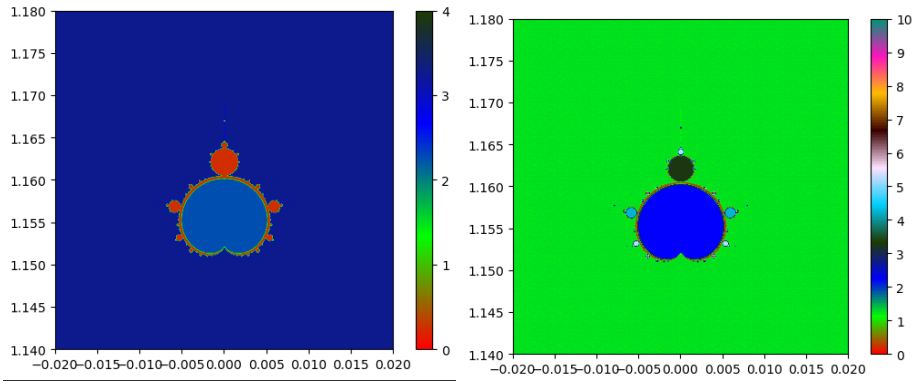


Fig. 4: Distribution of the cyclic behaviour in the parameter plane

One might also be interested in studying which particular kind of cyclic behaviour does each polynomial induce; that is, to distinguish between polynomials that induce attracting 2-cycles, 3-cycles,... In this regard, on the right plot in Figure 4 we have modified slightly the coloring strategy in order to make this distinction clear. In this plot, the color 1 (green) corresponds to polynomials that do not present any kind of attracting cyclic behaviour, neither any attracting extraneous fixed point, when Chebyshev's method is applied. Colors 2, 3, 4,

and so on, correspond to polynomials that induce an attracting extraneous fixed point, an attracting 2-cycle, an attracting 3-cycle, and so on, respectively. The color 0 (red) corresponds to the case where the considered maximum number of iterations of the algorithm was not enough for the orbits of the free critical points to converge.

We can note that the distribution of the attracting k-cyclic behaviour on the depicted Mandelbrot-like set closely resembles the distribution of the hyperbolic components of the classic and widely studied Mandelbrot set. Despite the fact that this resemblance is clear from a numerical point of view, currently there is not a known theoretical justification that supports it.

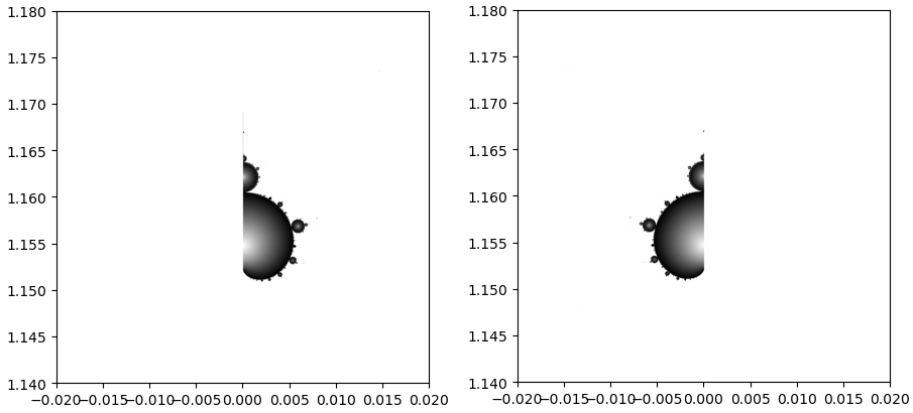


Fig. 5: Computed Lyapunov constants in a region with cyclic behaviour

Since, as we have mentioned, for these calculations in the parameter plane we are computing the Lyapunov constants associated with the basins of attraction in which each free critical point γ_1 and γ_2 of C_{p_λ} lies, we can also visualize the value of these Lyapunov constants in the parameter plane. In Figure 5, the left plot represents the computed Lyapunov constants associated with the free critical point γ_1, and the right plot represents the computed Lyapunov constants associated with γ_2. In these plots, the color white represents that the computed Lyapunov constant is 0; that is, that the orbit of the respective free critical point converges to a super-attracting fixed point or cycle. The color black represents that the computed Lyapunov constant is 1, and the different shades of grey correspond to the values between 0 and 1. Hence, the darker a point in the parameter plane is, the less attracting is the basin of attraction in which the respective free critical point lies.

Considering the same rectangle as in Figures 3 and 4, in Figure 5 we see that a vertical symmetry appears between the respective plots for γ_1 and γ_2. This is due to the vertical symmetry that the parameter plane itself exhibits, and that can be seen in Figure 1. In this case, one can observe that, in the Mandelbrot-

like set that can be seen in Figure 3 for which there is cyclic behaviour and extraneous fixed points, its right half is due to the convergence of the orbit of γ_1 to an attracting cycle or attracting extraneous fixed point, and its left half is due to the convergence of the orbit of γ_2.

This particular symmetry in the Mandelbrot-like regions does not always appear. For example, in Figure 6 we can see another Mandelbrot-like set corresponding to the appearance of attracting cycles or attracting extraneous fixed points in the rectangle $[-0.74, -0.68] \times [1.95, 2.01]$ in the parameter plane. In this case, we see that the free critical point that it is converging to an attracting cycle or attracting extraneous fixed point is γ_1 on the entire Mandelbrot-like region.

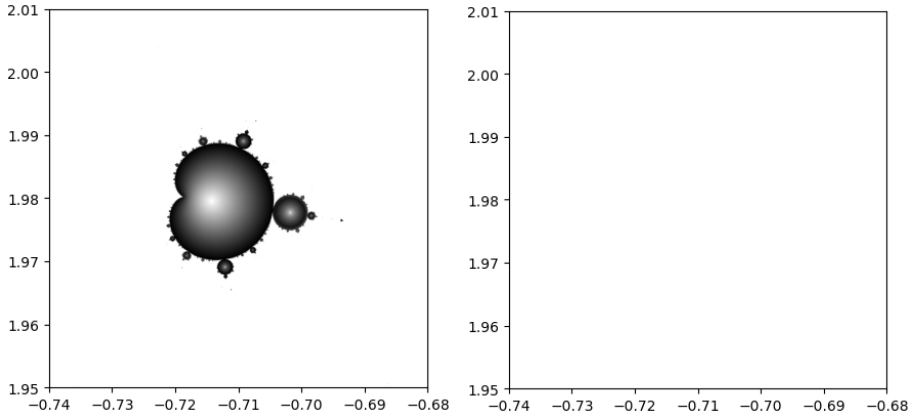


Fig. 6: Computed Lyapunov constants in a region with cyclic behaviour

In conclusion, we believe that the techniques used for this study are a useful tool from which much information regarding the dynamics of a root-finding numerical method such as Chebyshev's can be extracted. The development of these techniques might lead to some advances in this direction for this and other root-finding numerical methods.

References

1. V. Álvarez, J.M. García, L.J. Hernández and M.T. Rivas, *Algorithms for computing attraction basins of a rational self-map of the Hopf fibration based on Lyapunov exponents*, *Topology and its Applications*, 2023.
2. V. Álvarez-Aparicio, *Lyapunov Cycle Detector (LCD.jl)*, github.com/LCD, MIT license, v.1.4, 2022.
3. J.M. Gutiérrez and J.L. Varona, *Superattracting extraneous fixed points and n-cycles for Chebyshev's method on cubic polynomials*, Qualitative Theory of Dynamical Systems, 19:1-23, 2020.

Numerical approximation method for hybrid nonlinear Caputo fractional differential equations with boundary value conditions

K. Ben Amara[1] and M. I. Berenguer[2]

[1] Department of Mathematics. Faculty of Sciences of Sfax. University of Sfax. Road Soukra Km $3.5B.P.1171, 3000$, Sfax, Tunisia,
[2] Department of Applied Mathematics, E.T.S. de Ingeniería de Edificación and Institute of Mathematics (IMAG), University of Granada, Spain
khaled.benamara.etud@fss.usf.tn, maribel@ugr.es

Abstract. This paper presents a novel numerical approach to investigate the existence, uniqueness and approximation of solution for a broad range of fractional differential equations with boundary conditions. The proposed method combines the use of biorthogonal bases and Boyd-Wong fixed point theory.

Keywords: Banach algebras, biorthogonal bases, fixed point theory, fractional differential equations.

1 Introduction

Fractional differential equations have gained significant attention in recent years due to their ability to model complex phenomena in various scientific and engineering fields. Fractional calculus, a branch of mathematics dealing with derivatives and integrals of non-integer order, is employed to formulate them.

This paper presents a comprehensive study on the existence of solutions for large class of fractional differential equations in Banach algebras and proposes an efficient approximation method to numerically solve them.

Specifically, this study focuses with a rigorous analysis of the existence of an unique solution for the class of fractional differential equations of the type

$$\begin{cases} D^{\alpha}\left(\dfrac{x(t)-\sum_{j=1}^{m}\mathcal{I}^{\beta_j}h_j(t,x(t))}{f(t,x(t))}\right)=g(t,x(t)), t\in J, \\ \\ x(0)=a \text{ and } x(1)=b, \end{cases} \tag{1}$$

where $J=[0,1]$, a,b are real constants, D^{α} is the Caputo derivative of order α, $1<\alpha\leq 2$, I^{β_j} is the Riemann-Liouville fractional integral of order $\beta_j>0$, $j=1,\ldots,m$, and $f: J\times\mathbb{R}\to\mathbb{R}\setminus\{0\}$, $h_j: J\times\mathbb{R}\to\mathbb{R},$, $j=1,\ldots,m$, are given functions.

The Caputo fractional derivative and Riemann-Liouville fractional integral operators are defined as follows (see [3], [4] for details).

Definition 1. *Let $f : [0,\infty) \to \mathbb{R}$ be an $(n-1)$-times absolutely continuous mapping. The Caputo derivative of fractional order q of the function f is given by*

$$D^q f(t) = \frac{1}{\Gamma(n-q)} \int_0^t (t-s)^{n-q-1} f^{(n)}(s) ds, n-1 < q < n, n = [q]+1,$$

where $[q]$ denotes the integer part of the real number q.

Definition 2. *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a continuous function. The Riemann-Liouville fractional integral of order $q > 0$ is given by*

$$I^q f(t) = \frac{1}{\Gamma(q)} \int_0^t \frac{f(s)}{(t-s)^{1-q}} ds, t \in \mathbb{R}_+$$

provided the right-hand side is point-wise defined on $(0,\infty)$. Here the mapping Γ is defined by $\Gamma(n) = \displaystyle\int_0^\infty t^{n-1} e^{-t} dt$, $n > 0$.

Using the theory of fractional calculus and the Boyd-Wong fixed point theorem, sufficient conditions for the existence and uniqueness of solution for (1) are established. In addition we propose a method of approximating the solution and a numerical experiment as a sample of the validity of the effectiveness of the proposed method.

2 Existence, uniqueness and approximation of solutions.

According to an useful result due to Ahmad et al in [1], the problem of existence of a continuous solution for (1) can be transformed to a fixed point problem on $C(J)$ of the form:

$$x = A(x) \cdot B(x) + C(x). \tag{2}$$

where the space $C(J)$ of all continuous functions on J is endowed with the supremum norm $\|\cdot\|$.

Therefore, to study this problem we use the Boyd-Wong Theorem [2] which ensure that every nonlinear contraction mapping in a Banach space E, $S : E \to E$, i.e., there exists a continuous function $\Xi : [0,\infty) \to [0,\infty)$ such that $\Xi(r) < r$ if $r > 0$, and

$$\|S(x) - S(y)\| \leq \Xi(\|x-y\|), \quad \forall x, y \in E,$$

has a unique fixed point in E and moreover for any $x \in E$, the sequence $\{S^n(x), n \in \mathbb{N}\}$ converges to this point.

The conditions for the existence of a unique continuous solution in certain B_R (closed balls of $C(J)$) are rigorously derived in terms of a generalized Lipschitz condition with respect to the space variable called $\mathcal{D}$-Lipschitzian, i.e., $F : J \times E \to E$ is called $\mathcal{D}$-Lipschitzian with respect to the space variable, if there exist a continuous function $L : J \to J$ and a nondecreasing, continuous function $\Xi : [0, \infty) \to [0, \infty)$ with $\Xi(0) = 0$ such that

$$\|F(t, x) - F(t, y)\| \leq L(t)\Xi(\|x - y\|), \quad \forall t \in J, x, y \in E.$$

It allows us to express the solution of (1) as the limit of the sequence of elements $\{(A \cdot B + C)^n(X_0), n \in \mathbb{N}\}$, with $X_0 \in B_R$. Obviously, if it were possible to explicitly calculate, for each iteration, the expression $(A \cdot B + C)^n(X_0)$, then for each n we would have an approximation of the fixed point. But, as a practical matter, such an explicit calculation is only possible in very particular cases. For this reason, we need to construct another approximation of the fixed point which is simple to calculate in practice. For the problem at hand, since B and C are given by integral operators, we choice to approximate only the power terms of the operators B and C which are difficult to compute in general, unlike operator A which is easy to calculate and does not need to approximate their power terms. So, we will begin with an initial function $X_0 \in C(J)$ and construct two sequences of operators $\left\{\widetilde{B}_m\right\}_{m\in\mathbb{N}}$ and $\left\{\widetilde{C}_m\right\}_{m\in\mathbb{N}}$ in order to obtain successive $N_m \circ \ldots \circ N_1(X_0)$ approximations of the fixed point $\tilde{x}$ of the product $A \cdot B + C$.

To build these approximations we use fundamental biorthogonal systems in suitable Banach spaces. Recall that a biorthogonal system in a Banach space E is a system $\{u_n, \xi_n\}_{n\geq 1}$ of $E \times E^*$, where E^* denotes the topological dual space of E and $u_n(\xi_m) = \delta_{nm}$ (δ is Kronecker's delta). A Schauder basis of E is a sequence $\{u_n\}_{n\in\mathbb{N}} \subset E$ such that for every $x \in E$, there is a unique sequence $\{a_n\}_n \subset \mathbb{R}$ such that

$$x = \sum_{n\geq 1} a_n u_n.$$

Note that a Schauder basis is always a fundamental biorthogonal system, i.e., $\overline{span}\{u_n\} = E$, under the interpretation of the coordinate functionals as biorthogonal functionals.

Furthermore, posed approximation method is performed, demonstrating its accuracy and efficiency.

3 A numerical example

This subsection is devoted to illustrate the above model with a numerical example.

Example 1. We consider the following example

$$\begin{cases} D^{5/3}\left(\dfrac{x(t)-\sum_{j=1}^{2} I^{\frac{2j-1}{2}} h_j(t,x(t))}{f(t,x(t))}\right) = g(t,x(t)), t \in J, \\ x(0) = \frac{1}{10} \text{ and } x(1) = \frac{2}{10}. \end{cases} \tag{3}$$

where $g(t,x) = \frac{1}{2}x^2$, $h_j(t,x) = \frac{1}{2}x^2$ for $j = 1,2$ and and $f(t,x)$ is such that $\tilde{x}(t) = \frac{1}{10}(t+1)$ is the solution of the problem in $B_{\frac{1}{2}}$ (the closed ball of $C(J)$ centered in 0 with radius $\frac{1}{2}$).

Considering the usual biorthogonal systems and with initial $X_0(t) = \frac{1}{5}(t+1)$, we obtain for $m = 2$,

$$\|x^* - \tilde{x}\| = 0.00636518$$

where x^* represents the approximate solution of (3).

4 Conclusion

In conclusion, this work provides a comprehensive investigation into the existence of a unique continuous solution for the fractional differential equation (1) and introduces a novel approximation method for their numerical solution. The established existence results ensure the presence of viable solutions under appropriate conditions, while the proposed approximation method overcomes the challenges associated with fractional derivatives. The contributions of this research are expected to significantly enhance the understanding and numerical treatment of fractional differential equations, facilitating their application in real-world problems across diverse applied scientific and engineering disciplines.

References

1. B. Ahmad, S.K. Ntouyas, J. Tariboon, *Nonlocal hybrid boundary value problems of Caputo fractional integro-differential equations.* Acta Math. Sci. 36 (2016).
2. Boyd, D. W.; Wong, J. S. W. *On nonlinear contractions.* Proc. Amer. Math. Soc. 20 (1969), 458–464.
3. Kilbas AA , Srivastava HM , Trujillo JJ. *Theory and applications of fractional differential equations.* Amsterdam: North-Holland Mathematics Studies, 204. Elsevier Science B.V.; 2006.
4. Podlubny I. *Fractional Differential Equations.* San Diego: Academic Press, 1999

An inverse problem for Fredholm-type integro-differential equations with application to pollution emission modelling

M. I. Berenguer[1], D. Gámez[1], H. Kunze[2], D. La Torre[3], and M. Ruiz Galán[1]

[1] Dpt. of Applied Mathematics and IMAG, University of Granada, Spain,
[2] Dpt. of Mathematics and Statistics, University of Guelph, Canada,
[3] SKEMA Business School, Université Côte d'Azur, France.
maribel@ugr.es, domingo@ugr.es, hkunze@uoguelph.ca,
davide.latorre@skema.edu mruizg@ugr.es

Abstract. In this paper we present a novel and innovative approach aimed at approximating the solution to a specific class of integro-differential equations of the Fredholm type. In addition, we delve into an exhaustive analysis of the associated inverse problem through the implementation of a collage-type result. We show the practicality and relevance of our proposed method and we apply it to the modelling of pollutant emissions, an area of great concern in today's society.

Keywords: Inverse problem, integro–differential equations, pollution emission model.

1 Introduction

In this work we study the existence, uniqueness and solution approximation, as well as the associated inverse problem of the following integro-differential problem of the Fredholm type:

$$\begin{cases} \dfrac{\partial u(x,t)}{\partial t} = g(x,t)u(x,t) + \displaystyle\int_a^b f(x,s,u(s,t))d\mu(s), (x,t) \in [a,b]\times[0,\tau], \\ u(x,0) = u_0(x) \end{cases} . \quad (1)$$

where μ is a generic probability measure with compact support over $[a,b]$, the functions $g \in C([a,b]\times[0,\tau])$, $f \in C([a,b]^2 \times \mathbb{R})$ and $u_0 \in C([a,b])$ are given and $u \in C([a,b]\times[0,\tau])$ is the unknown function to be determined.

As a particular case of the previous model, we will consider the following model of pollution emmision in a one-dimensional environment, that is a spatial model in which the agents are located along a linear city. In this context, the pollution level $P(x,t)$ at time $t \in [0,\tau]$ and location $x \in [a,b]$ can be modelled by a reaction-diffusion equation subject to an initial condition at $t = 0$, which when heavy pollutants are considered or we are just interested in radioactive

pollution, take the form (see [9–11]):

$$\begin{cases} \dfrac{\partial P}{\partial t}(x,t) = S(x,t)P(x,t) - \delta_P P(x,t) + \displaystyle\int_a^b \phi(s,x)P(s,t)d\mu, \ (x,t) \in [a,b] \times [0,\tau], \\ \\ P(x,0) = P_0(x), x \in [a,b] \end{cases} \tag{2}$$

with μ a probability measure with compact support over the interval $[a,b]$.

2 Uniqueness of the solution and its approximation

Under the hypothesis that $g \in C([a,b]\times[0,\tau])$, $u_0 \in C([a,b])$ and $f \in C([a,b]^2 \times \mathbb{R})$ satisfies a Lipschitz condition with respect to the last variable, problem (1) has a unique solution that can be obtained as the limit of a sequence of Picard iterants. It is enough to use the Caccioppoli-Banach Fixed Point Theorem and take into account the Fundamental Theorem of Calculus, which allows us to state that u is solution of (1) if, and only if u is a fixed point of the operator

$$\mathbf{T} : C([a,b] \times [0,\tau]) \longrightarrow C([a,b] \times [0,\tau])$$

defined for $(x,t) \in [a,b] \times [0,\tau]$ and $u \in C([a,b] \times [0,\tau])$ as

$$(\mathbf{T}u)(x,t) := u_0(x) + \int_0^t g(x,r)u(x,r)dr + \int_0^t \int_a^b f(x,s,u(s,r))d\mu(s)dr, \tag{3}$$

where "dr" denotes Lebesgue measure.

The use of Shauder bases in the Banach spaces $C([a,b] \times [0,\tau])$ and $C([a,b]^2 \times [0,\tau])$ allow us to build operators, easy to compute, that approximate the operator $\mathbf{T}$, to later derive the construction of a sequence that approximates the Picard iterant sequence, thus obtaining a novel method to approximate the only solution of the problem (1).

3 The inverse problem with application to pollution model

In the study of inverse problems, the determination of certain parameters associated with a given model has been approached from different perspectives. One of them is based on the Collage theorem, which is a simple but powerful consequence of Banach's fixed point theorem (see [3–8]). This technique reads as follows. Let $(X, \|\cdot\|)$ be a Banach space and $\Lambda \subset \mathbb{R}^n$ be a nonempty compact set. For each $\lambda \in \Lambda$, assume that $\mathbf{T}_\lambda : X \to X$ is a contractive operator with contractivity factor $c_\lambda \in (0,1)$ and that $u_\lambda^\bullet$ is the unique fixed point of $\mathbf{T}_\lambda$. Let $u^* \in X$ be a target element, that is, u^* is the fixed point (or an approximation) of the operators $\mathbf{T}_\lambda$ and we are interested in estimating the parameter $\lambda^* \in \Lambda$

such that $\mathbf{T}_{\lambda^*}(u^*) = u^*$.

According to the Collage theorem,

$$\|u_\lambda^\bullet - u^*\| \leq \frac{1}{1-c_\lambda}\|u^* - \mathbf{T}_\lambda(u^*)\|.$$

In this context, $\|u^* - \mathbf{T}_\lambda(u^*)\|$ is called Collage distance. If we suppose that

$$\sup_{\lambda\in\Lambda} c_\lambda < 1,$$

our interest becomes finding

$$\lambda^* \in \operatorname{argmin}_{\lambda\in\Lambda}\|u^* - \mathbf{T}_\lambda(u^*)\|.$$

By modifying the proof of the Collage Theorem, the following more general result than the Collage Theorem can be obtained under the hypotheses of the Caccioppoli-Banach Fixed Point Theorem:

Theorem 1. *(see [2]) Let $(X, \|\cdot\|)$ be a Banach space, let $\boldsymbol{T}: X \longrightarrow X$ and let $\{\rho_n\}_{n\geq 1}$ be a sequence of nonnegative real numbers such that the series $\sum_{n\geq 1}\rho_n$ is convergent and for all $u_1, u_2 \in X$ and for all $n \geq 1$,*

$$\|\boldsymbol{T}^n(u_1) - \boldsymbol{T}^n(u_2)\| \leq \rho_n\|u_1 - u_2\|.$$

Let $u^\bullet$ be the fixed point of $\boldsymbol{T}$ and $\widetilde{u} \in X$ be the target. Then there exits n_0 such that $\rho_{n_0} < 1$ and this implies

$$\|u^\bullet - u^*\| \leq \frac{\sum_{i=0}^{n_0-1}\rho_i}{1-\rho_{n_0}}\|\boldsymbol{T}(u^*) - u^*\|. \tag{4}$$

The result can be used, analogously to how we have explained above for problem (1) when $f \in C([a,b]^2 \times \mathbb{R})$ satisfies a Lipschitz condition with respect to the last variable. Under this assumption, known f, we have numerically implemented the inverse problem for (1) to determine g with format $g(x,t) = a_1 + a_2x + a_3t$. Also, we have tested the inverse problem for (2) in the case where we have collected data of P, we know both, δ_P and ϕ, and we are interested in determining an estimate S.

4 Future lines

Future lines of research include the analysis of more general models of macroeconomic geography in which demography and pollution dynamics affect each other. These problems are modelled with systems of two partial differential equations: one describing the evolution of pollution over time and space and the other describing the evolution of the human population.

References

1. Almezel, S., Ansari, Q.H., Khamsi, M.A., Topics in fixed point theory, Springer, Cham, 2014.
2. Arana-Jiménez, M., Berenguer, M.I., Gámez, D., Garralda-Guillem, A.I., Ruiz Galán, M. , A perturbed collage theorem and its application to inverse interval integral problems, Communications in Nonlinear Science and Numerical Simulation 90, 105365, 10 pages, (2020).
3. Berenguer, M. I., Kunze, H., La Torre, D., Ruiz Galán, M., Galerkin schemes and inverse boundary value problems in reflexive Banach spaces, J. Comput. Appl. Math. 275, 100–112 (2015).
4. Berenguer, M. I., Kunze, H., La Torre, D., Ruiz Galán, M. , Galerkin method for constrained variational equations and a collage-based approach to related inverse problems, J. Comput. Appl. Math. 292, 67–75 (2016).
5. Capasso, V., Kunze, H., La Torre, D., Vrscay, E. R., Solving inverse problems for differential equations by a "generalized collage" method and application to a mean field stochastic model, Nonlinear Analysis: Real World Applications 15, 276-289 (2014).
6. Deng, X., Liao, Q., Parameter Estimation for Partial Differential Equations by Collage-Based Numerical Approximation, Mathematical Problems in Engineering, 510934 (2009).
7. Kirsch A., An introduction to the mathematical theory of inverse problems, Springer, 2011.
8. Kunze, H., La Torre, D., Vrscay, E.R., Random fixed point equations and inverse problems using collage method for contraction mappings, Journal of Mathematical Analysis and Applications, 334, 1116–1129 (2007).
9. La Torre, D., Liuzzi D., Marsiglio, S. Pollution diffusion and abatement activities across space and over time, Mathematical Social Sciences 78, 48–63 (2015).
10. La Torre, D., Liuzzi, D., Marsiglio, S., Pollution control under uncertainty and sustainability concern, Environmental and Resource Economics 67, 885-903 (2017).
11. La Torre, D., Liuzzi, D., Marsiglio, S., Population and geography do matter for sustainable development, Environment and Development Economics 24, 201-223 (2019).

Optimal multipoint fractional methods for solving nonlinear problems

Giro Candelario[1], Alicia Cordero[2], Juan R. Torregrosa[2], and María P. Vassileva[1]

[1] Instituto Tecnológico de Santo Domingo, Área de Ciencias Básicas y Ambientales, Santo Domingo, Dominican Republic,
giro.candelario@intec.edu.do

[2] Universitat Politècnica de València, Institute for Multidisciplinary Mathematics, València, Spain

Abstract. In this work, we introduce the design of higher-order Newton-type schemes with conformable derivatives and higher-order classical derivatives for the solution of nonlinear equations. We also propose a general technique to obtain the conformable version of any classical procedure, and we use it to design the conformable version of some multipoint classical methods. The convergence analysis of these schemes is made, obtaining higher or equal order of convergence compared to classical procedures. The numerical experiments hold the results acquired in the theory, and these methods present a good stability.

Keywords: Nonlinear equations · Conformable derivative · Higher-order Newton's methods · Multipoint schemes · Optimal procedures · Stability

1 Introduction

In the literature, some iterative methods with fractional and fractal derivatives were proposed to find the solution $\bar{x} \in \mathbb{R}$ of a function $f : I \subseteq \mathbb{R} \to \mathbb{R}$, being f a continuous function in I (see [1, 2, 2–6]). These schemes do not preserve the theoretical order of convergence in the practice, unlike iterative procedures with conformable derivatives [8, 9]. Our aim is to increase the order of convergence of conformable methods, introducing higher-order Newton-type schemes with conformable derivatives, and design a technique to obtain the conformable version of any procedure with integer order derivatives.
The left conformable derivative of $f : [a, \infty) \to \mathbb{R}$ of order $\alpha \in (0, 1]$, starting from a, $x, a, \alpha \in \mathbb{R}$, $a < x$, is [10, 11]

$$(T_\alpha^a f)(x) = \lim_{\varepsilon \to 0} \frac{f(x + \varepsilon(x-a)^{1-\alpha}) - f(x)}{\varepsilon}. \tag{1}$$

Next, we present a Taylor power series of f by using conformable derivatives evaluated at a_1 [12]:

$$f(x) = f(a_1) + \frac{(T_\alpha^a f)(a_1)\delta_1}{\alpha} + \frac{(T_\alpha^a f)^{(2)}(a_1)\delta_1^2}{2\alpha^2} + R_2(x, a_1, a), \tag{2}$$

where $\delta_1 = H^\alpha - L^\alpha$, being $H = x - a$, and $L = a_1 - a$.
In next section, the design of three higher-order Newton-type schemes is proposed, and a general technique is provided in order to obtain the conformable version of any classical method. Later, we use this technique to design multipoint schemes with conformable derivatives, and make the convergence analysis of all procedures. In Section 3, we talk about the numerical performance of these methods. Finally, in Section 4 we show the conclusions.

2 Design and convergence analysis of the methods

Let us consider the error equation of conformable Newton method proposed in [8] up to order four:

$$\begin{aligned} e_{k+1} = &\left(C_2 + \frac{1-\alpha}{2(\bar{x}-a)}\right) e_k^2 + \left(2C_3 - 2C_2^2 + \frac{(\alpha-1)C_2}{\bar{x}-a} + \frac{(1-\alpha)(\alpha-2)}{3(\bar{x}-a)^2}\right) e_k^3 \\ &+ \left(3C_3 - 7C_2C_3 + 4C_2^3 + \frac{(1-\alpha)(5C_2^2 - C_3)}{2(\bar{x}-a)} + \frac{(2\alpha^2 - 5\alpha + 3)C_2}{2(\bar{x}-a)^2}\right. \\ &\left. + \frac{(1-\alpha)(2\alpha^2 - 7\alpha + 7)}{8(\bar{x}-a)^3}\right) e_k^4 + O\left(e_k^5\right). \end{aligned} \tag{3}$$

So, we can design the following Newton-type schemes [13]:

$$x_{k+1} = a + \left((x_k - a)^{\alpha_k} - \alpha_k \frac{f(x_k)}{(T^a_{\alpha_k} f)(x_k)}\right)^{1/\alpha_k}, \quad k = 0, 1, 2, \ldots, \tag{4}$$

being $\alpha_k = 1 + (x_k - a)\dfrac{f''(x_k)}{f'(x_k)}$, $k = 0, 1, 2, \ldots$, where α_k is isolated from (3), and is denoted by NeL3,

$$x_{k+1} = a_k + \left((x_k - a_k)^{\alpha} - \alpha \frac{f(x_k)}{(T^{a_k}_{\alpha} f)(x_k)}\right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots, \tag{5}$$

being $a_k = x_k + (1-\alpha)\dfrac{f'(x_k)}{f''(x_k)}$, $k = 0, 1, 2, \ldots$, where a_k is isolated from (2), and is denoted by NeA3, and

$$x_{k+1} = a_k + \left((x_k - a_k)^{\alpha_k} - \alpha_k \frac{f(x_k)}{(T^{a_k}_{\alpha_k} f)(x_k)}\right)^{1/\alpha_k}, \quad k = 0, 1, 2, \ldots, \tag{6}$$

being $\alpha_k = 1 + \dfrac{f''(x_k)^2}{f''(x_k)^2 - f'(x_k)f'''(x_k)}$, and $a_k = x_k + \dfrac{f'(x_k)f''(x_k)}{f'(x_k)f'''(x_k) - f''(x_k)^2}$, $k = 0, 1, 2, \ldots$, where α_k and a_k are isolated from (2), and is denoted by NeLA4.
The general technique proposed in [13] states that, for a classical procedure

$$\phi(x) = x - f(x)g(x), \tag{7}$$

the conformable version is given by

$$\phi(x) = a + \left((x-a)^{\alpha} - \alpha f(x) g_{\alpha}(x)\right)^{1/\alpha}. \tag{8}$$

So, we are able to design some conformable multipoint methods [13]:

1. Traub's scheme [14, 15] in conformable version:

$$\psi_1(x) = a + \left((\phi_2(x)-a)^{\alpha} - \alpha \frac{f[\phi_1(x)]}{(T_{\alpha}^{a} f)(x)}\right)^{1/\alpha}, \tag{9}$$

   where

$$\phi_1(x) = a + \left((x-a)^{\alpha} - \alpha \frac{f(x)}{(T_{\alpha}^{a} f)(x)}\right)^{1/\alpha}, \tag{10}$$

   denoted by TeCO.
2. Chun-Kim's procedure [15, 16] in conformable version:

$$\psi_2(x) = a + \left((x-a)^{\alpha} - \frac{\alpha}{2}\left[3 - \frac{(T_{\alpha}^{a} f)[\phi_1(x)]}{(T_{\alpha}^{a} f)(x)}\right]\frac{f(x)}{(T_{\alpha}^{a} f)(x)}\right)^{1/\alpha}, \tag{11}$$

   denoted by CKeCO.
3. Ostrowski's method [14, 15] in conformable version:

$$\psi_3(x) = a + \left((\phi_2(x)-a)^{\alpha} - \alpha\left[\frac{f(x)}{f(x)-2f[\phi_1(x)]}\right]\frac{f[\phi_1(x)]}{(T_{\alpha}^{a} f)(x)}\right)^{1/\alpha}, \tag{12}$$

   denoted by OeCO.
4. Chun's scheme [15] in conformable version:

$$\psi_4(x) = a + \left((\phi_2(x)-a)^{\alpha} - \alpha\left[\frac{f(x)+2f[\phi_2(x)]}{f(x)}\right]\frac{f[\phi_2(x)]}{(T_{\alpha}^{a} f)(x)}\right)^{1/\alpha}, \tag{13}$$

   denoted by CeCO.

The next results assure the convergence of the procedures previously proposed:

Theorem 1. *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a continuous function in the interval I containing the zero $\bar{x}$ of $f(x)$. Let $(T_{\alpha_k}^{a} f)(x)$ be the conformable derivative of $f(x)$ starting from a, with order α_k. Let us suppose that $(T_{\alpha_k}^{a} f)(x)$ is continuous and not null at $\bar{x}$. If an initial approximation x_0 is sufficiently close to $\bar{x}$, then the local order of convergence of the conformable Newton-type scheme (NeL3)*

$$x_{k+1} = a + \left((x_k-a)^{\alpha_k} - \alpha_k \frac{f(x_k)}{(T_{\alpha_k}^{a} f)(x_k)}\right)^{1/\alpha_k}, \quad k = 0, 1, 2, \ldots,$$

where

$$\alpha_k = 1 + (x_k - a)\frac{f''(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots,$$

is at least 3*, and its error equation is*

$$e_{k+1} = \frac{1}{3}\left(2C_2^2 - 3C_3 - \frac{C_2}{\bar{x}-a}\right)e_k^3 + O\left(e_k^4\right),$$

being $C_j = \dfrac{f^{(j)}(\bar{x})}{j!f'(\bar{x})}$*, for* $j \geq 2$*, such that* $a < x_k$*,* $\forall k$*.*

Theorem 2. *Let* $f : I \subseteq \mathbb{R} \to \mathbb{R}$ *be a continuous function in the interval* I *containing the zero* $\bar{x}$ *of* $f(x)$*. Let* $(T_\alpha^{a_k} f)(x)$ *be the conformable derivative of* $f(x)$ *starting from* a_k*, with order* α*, for any* $\alpha \in (0,1)$*. Let us suppose that* $(T_\alpha^{a_k} f)(x)$ *is continuous and not null at* $\bar{x}$*. If an initial approximation* x_0 *is sufficiently close to* $\bar{x}$*, then the local order of convergence of the conformable Newton-type scheme (NeA3)*

$$x_{k+1} = a_k + \left((x_k - a_k)^\alpha - \alpha\frac{f(x_k)}{(T_\alpha^{a_k} f)(x_k)}\right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

where

$$a_k = x_k + (1-\alpha)\frac{f'(x_k)}{f''(x_k)}, \quad k = 0, 1, 2, \ldots,$$

is at least 3 *for* $0 < \alpha < 1$*, and the error equation is*

$$e_{k+1} = \left(\frac{2}{3}\frac{(2-\alpha)C_2^2}{1-\alpha} - C_3\right)e_k^3 + O\left(e_k^4\right).$$

Theorem 3. *Let* $f : I \subseteq \mathbb{R} \to \mathbb{R}$ *be a continuous function in the interval* I *containing the zero* $\bar{x}$ *of* $f(x)$*. Let* $(T_{\alpha_k}^{a_k} f)(x)$ *be the conformable derivative of* $f(x)$ *starting from* a_k*, with order* α_k*. Let us suppose that* $(T_{\alpha_k}^{a_k} f)(x)$ *is continuous and not null at* $\bar{x}$*. If an initial approximation* x_0 *is sufficiently close to* $\bar{x}$*, then the local order of convergence of the conformable Newton-type procedure (NeLA4)*

$$x_{k+1} = a_k + \left((x_k - a_k)^{\alpha_k} - \alpha_k\frac{f(x_k)}{(T_{\alpha_k}^{a_k} f)(x_k)}\right)^{1/\alpha_k}, \qquad k = 0, 1, 2, \ldots,$$

where

$$\alpha_k = 1 + \frac{f''(x_k)^2}{f''(x_k)^2 - f'(x_k)f'''(x_k)}, \quad k = 0, 1, 2, \ldots,$$

and

$$a_k = x_k + \frac{f'(x_k)f''(x_k)}{f'(x_k)f'''(x_k) - f''(x_k)^2}, \quad k = 0, 1, 2, \ldots,$$

is at least 4*, and the error equation is*

$$e_{k+1} = 2\left(C_2C_3 - 3\frac{C_3^2}{C_2} + 2C_4\right)e_k^4 + O\left(e_k^5\right).$$

Theorem 4. *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a continuous function in the interval I containing the zero $\bar{x}$ of $f(x)$. Let $(T_\alpha^a f)(x)$ be the conformable derivative of $f(x)$ starting from a, with order α, for any $\alpha \in (0,1]$. Let us suppose that $(T_\alpha^a f)(x)$ is continuous and not null at $\bar{x}$. If an initial approximation x_0 is sufficiently close to $\bar{x}$, then the local order of convergence of the conformable Traub-type procedure (TeCO)*

$$x_{k+1} = a + \left((y_k - a)^\alpha - \alpha \frac{f(y_k)}{(T_\alpha^a f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

where

$$y_k = a + \left((x_k - a)^\alpha - \alpha \frac{f(x_k)}{(T_\alpha^a f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

is at least 3*, and the error equation is*

$$e_{k+1} = \left(2C_2^2 + 2\frac{(1-\alpha)C_2}{\bar{x} - a} + \frac{1}{2}\frac{(1-\alpha)^2}{(\bar{x} - a)^2} \right) e_k^3 + O\left(e_k^4\right).$$

Theorem 5. *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a continuous function in the interval I containing the zero $\bar{x}$ of $f(x)$. Let $(T_\alpha^a f)(x)$ be the conformable derivative of $f(x)$ starting from a, with order α, for any $\alpha \in (0,1]$. Let us suppose that $(T_\alpha^a f)(x)$ is continuous and not null at $\bar{x}$. If an initial approximation x_0 is sufficiently close to $\bar{x}$, then the local order of convergence of the conformable Chun-Kim-type scheme (CKeCO)*

$$x_{k+1} = a + \left((x_k - a)^\alpha - \frac{\alpha}{2}\left[3 - \frac{(T_\alpha^a f)(y_k)}{(T_\alpha^a f)(x_k)} \right] \frac{f(x_k)}{(T_\alpha^a f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

where

$$y_k = a + \left((x_k - a)^\alpha - \alpha \frac{f(x_k)}{(T_\alpha^a f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

is at least 3*, and the error equation is*

$$e_{k+1} = \left(2C_2^2 + \frac{1}{2}C_3 + \frac{5}{2}\frac{(1-\alpha)C_2}{\bar{x} - a} + \frac{1}{12}\frac{(1-\alpha)(7-8\alpha)}{(\bar{x} - a)^2} \right) e_k^3 + O\left(e_k^4\right).$$

Theorem 6. *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a continuous function in the interval I containing the zero $\bar{x}$ of $f(x)$. Let $(T_\alpha^a f)(x)$ be the conformable derivative of $f(x)$ starting from a, with order α, for any $\alpha \in (0,1]$. Let us suppose that $(T_\alpha^a f)(x)$ is continuous and not null at $\bar{x}$. If an initial approximation x_0 is sufficiently close to $\bar{x}$, then the local order of convergence of the conformable Ostrowski-type method (OeCO)*

$$x_{k+1} = a + \left((y_k - a)^\alpha - \alpha \left[\frac{f(x_k)}{f(x_k) - 2f(y_k)} \right] \frac{f(y_k)}{(T_\alpha^a f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

where

$$y_k = a + \left((x_k - a)^\alpha - \alpha \frac{f(x_k)}{(T_\alpha^a f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

is at least 4*, and the error equation is*

$$e_{k+1} = \left(C_2^3 - C_2 C_3 + \frac{1}{2} \frac{(1-\alpha)(C_2^2 - C_3)}{\bar{x} - a} + \frac{1}{12} \frac{(1-\alpha^2)^2 C_2}{(\bar{x} - a)^2} + \frac{1}{24} \frac{(1-\alpha)(1-\alpha^2)}{(\bar{x} - a)^3} \right) e_k^4 + O\left(e_k^5\right).$$

Theorem 7. *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a continuous function in the interval I containing the zero $\bar{x}$ of $f(x)$. Let $(T_\alpha^a f)(x)$ be the conformable derivative of $f(x)$ starting from a, with order α, for any $\alpha \in (0, 1]$. Let us suppose that $(T_\alpha^a f)(x)$ is continuous and not null at $\bar{x}$. If an initial approximation x_0 is sufficiently close to $\bar{x}$, then the local order of convergence of the conformable Chun-type method (CeCO)*

$$x_{k+1} = a + \left((y_k - a)^\alpha - \alpha \left[\frac{f(x_k) + 2f(y_k)}{f(x_k)} \right] \frac{f(y_k)}{(T_\alpha^a f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

where

$$y_k = a + \left((x_k - a)^\alpha - \alpha \frac{f(x_k)}{(T_\alpha^a f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots,$$

is at least 4*, and the error equation is*

$$e_{k+1} = \left(5C_2^3 - C_2 C_3 + \frac{(1-\alpha)(13C_2^2 - C_3)}{2(\bar{x} - a)} + \frac{(24(1-\alpha)^2 + (1-\alpha)(13 - 11\alpha))C_2}{12(\bar{x} - a)^2} + \frac{(1-\alpha)^2(13 - 11\alpha)}{24(\bar{x} - a)^3} \right) e_k^4 + O\left(e_k^5\right).$$

The numerical performance of given procedures is dicussed in next section.

3 Numerical results

We used Matlab R2020a with double precision arithmetics, $|x_{k+1} - x_k| < 10^{-8}$ or $|f(x_{k+1})| < 10^{-8}$ as stopping criterium, and a maximum of 500 iterations. We used the Approximated Computational Order of Convergence (ρ) [17]:

$$\rho = \frac{\ln(|x_{k+1} - x_k|/|x_k - x_{k-1}|)}{\ln(|x_k - x_{k-1}|/|x_{k-1} - x_{k-2}|)}, \quad k = 2, 3, 4, \ldots$$

We compare the results obtained with each mathod, including the classical version of conformable multipoint schemes, and we also compare our procedures with the Toolbox fsolve from Matlab. In methods NeL3, TeCO, CKeCO, OeCO and CeCO we fix $a = -10$, because a is not calculated in each iteration.

Table 1: Results of one-point methods and *fsolve* for $f(x)$, with initial estimate $x_0 = 1$

NeA3 method					
α	$\bar{x}$	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1} - x_k\rvert$	iter	ρ
1	$\bar{x}_4$	$4.36 \cdot 10^{-11}$	$6.85 \cdot 10^{-7}$	13	2.01
0.9	$\bar{x}_4$	$3.55 \cdot 10^{-15}$	$1.02 \cdot 10^{-9}$	7	2.92
0.8	$\bar{x}_3$	$3.91 \cdot 10^{-13}$	$3.58 \cdot 10^{-8}$	15	2.89
0.7	$\bar{x}_6$	$5.33 \cdot 10^{-15}$	$5.35 \cdot 10^{-10}$	38	3.01
0.6	$\bar{x}_4$	$1.71 \cdot 10^{-13}$	$8.73 \cdot 10^{-6}$	22	3.37
0.5	$\bar{x}_3$	$6.18 \cdot 10^{-13}$	$6.70 \cdot 10^{-10}$	54	2.96
0.4	$\bar{x}_3$	$3.91 \cdot 10^{-13}$	$1.69 \cdot 10^{-7}$	31	2.91
0.3	$\bar{x}_4$	$1.03 \cdot 10^{-9}$	$1.88 \cdot 10^{-4}$	19	3.57
0.2	$\bar{x}_3$	$4.02 \cdot 10^{-9}$	$8.18 \cdot 10^{-5}$	35	2.81
0.1	$\bar{x}_4$	$9.07 \cdot 10^{-10}$	$1.94 \cdot 10^{-4}$	168	3.53
NeL3 method					
α_k	$\bar{x}$	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1} - x_k\rvert$	iter	ρ
-	$\bar{x}_3$	$4.13 \cdot 10^{-13}$	$4.10 \cdot 10^{-6}$	11	2.89
NeLA4 method					
α_k	$\bar{x}$	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1} - x_k\rvert$	iter	ρ
-	$\bar{x}_1$	$6.53 \cdot 10^{-15}$	$1.51 \cdot 10^{-7}$	3	4.00
fsolve					
-	$\bar{x}$	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1} - x_k\rvert$	iter	ρ
-	-	-	-	-	-

Our test function is $f_1(x) = -12.84x^6 - 25.6x^5 + 16.55x^4 - 2.21x^3 + 26.71x^2 - 4.29x - 15.21$, whose real and complex roots are $\bar{x}_1 = 0.82366 + 0.24769i$, $\bar{x}_2 = 0.82366 - 0.24769i$, $\bar{x}_3 = -2.62297$, $\bar{x}_4 = -0.584$, $\bar{x}_5 = -0.21705 + 0.99911i$ and $\bar{x}_6 = -0.21705 - 0.99911i$.
In Table 1, NeA3 requires less or more iterations than classical Newton method, and ρ is three. NeLA3 and NeL3 require less iterations than classical Newton's

Table 2: Results of multipoint methods for $f(x)$, with initial estimate $x_0 = 1$

	NeA3 method					NeA3 method				
α	$\bar{x}$	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1}-x_k\rvert$	iter	ρ	$\bar{x}$	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1}-x_k\rvert$	iter	ρ
1	$\bar{x}_3$	$6.18\cdot10^{-13}$	$4.49\cdot10^{-7}$	115	2.89	-	-	-	>500	-
0.9	$\bar{x}_4$	$4.04\cdot10^{-9}$	$2.43\cdot10^{-4}$	69	2.80	-	-	-	>500	-
0.8	$\bar{x}_4$	$5.30\cdot10^{-11}$	$5.73\cdot10^{-5}$	61	2.83	$\bar{x}_4$	$9.95\cdot10^{-14}$	$9.91\cdot10^{-10}$	190	3.00
0.7	$\bar{x}_2$	$1.16\cdot10^{-13}$	$3.57\cdot10^{-6}$	329	0.00	-	-	-	>500	-
0.6	$\bar{x}_4$	$1.07\cdot10^{-14}$	$1.12\cdot10^{-6}$	119	2.90	-	-	-	>500	-
0.5	$\bar{x}_4$	$3.24\cdot10^{-10}$	$1.05\cdot10^{-4}$	213	2.82	-	-	-	>500	-
0.4	$\bar{x}_4$	$1.17\cdot10^{-13}$	$7.31\cdot10^{-9}$	104	2.95	$\bar{x}_5$	$1.41\cdot10^{-11}$	$1.88\cdot10^{-5}$	484	2.80
0.3	-	-	-	>500	-	$\bar{x}_4$	$2.11\cdot10^{-13}$	$1.11\cdot10^{-5}$	490	0.00
0.2	-	-	-	>500	-	-	-	-	>500	-
0.1	-	-	-	>500	-	-	-	-	-	-

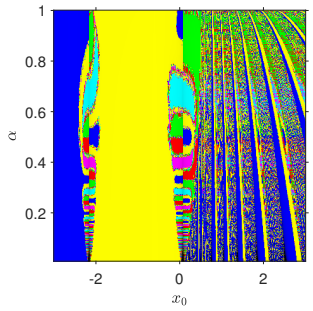
	OeCO method					CeCO method				
α	$\bar{x}$	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1}-x_k\rvert$	iter	ρ	$\bar{x}$	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1}-x_k\rvert$	iter	ρ
1	$\bar{x}_4$	$1.07\cdot10^{-14}$	$2.77\cdot10^{-5}$	3	2.41	$\bar{x}_3$	$6.18\cdot10^{-13}$	$4.76\cdot10^{-9}$	47	3.73
0.9	$\bar{x}_4$	$9.95\cdot10^{-14}$	$1.97\cdot10^{-5}$	3	2.43	$\bar{x}_3$	$3.11\cdot10^{-10}$	$2.76\cdot10^{-4}$	75	3.21
0.8	$\bar{x}_4$	$9.95\cdot10^{-14}$	$1.36\cdot10^{-5}$	3	2.44	$\bar{x}_3$	$6.18\cdot10^{-13}$	$9.77\cdot10^{-5}$	66	3.28
0.7	$\bar{x}_4$	$1.17\cdot10^{-13}$	$9.14\cdot10^{-6}$	3	2.46	$\bar{x}_3$	$2.26\cdot10^{-12}$	$1.89\cdot10^{-10}$	87	3.81
0.6	$\bar{x}_4$	$9.95\cdot10^{-14}$	$5.93\cdot10^{-6}$	3	2.48	$\bar{x}_3$	$1.81\cdot10^{-10}$	$2.42\cdot10^{-4}$	108	3.22
0.5	$\bar{x}_4$	$1.17\cdot10^{-13}$	$3.68\cdot10^{-6}$	3	2.49	$\bar{x}_3$	$4.92\cdot10^{-11}$	$1.76\cdot10^{-4}$	87	3.24
0.4	$\bar{x}_4$	$1.17\cdot10^{-13}$	$2.16\cdot10^{-6}$	3	2.51	$\bar{x}_4$	$1.17\cdot10^{-13}$	$1.08\cdot10^{-7}$	212	3.68
0.3	$\bar{x}_4$	$1.07\cdot10^{-14}$	$1.18\cdot10^{-6}$	3	2.53	$\bar{x}_4$	$1.07\cdot10^{-14}$	$1.75\cdot10^{-7}$	53	3.67
0.2	$\bar{x}_4$	$2.24\cdot10^{-13}$	$5.90\cdot10^{-7}$	3	2.55	$\bar{x}_4$	$2.24\cdot10^{-13}$	$7.13\cdot10^{-7}$	45	3.61
0.1	$\bar{x}_4$	$4.21\cdot10^{-13}$	$2.58\cdot10^{-7}$	3	2.57	-	-	-	-	-

method, and the computational order of convergence is greater. We can note that fsolve does not find solution with this initial estimate.

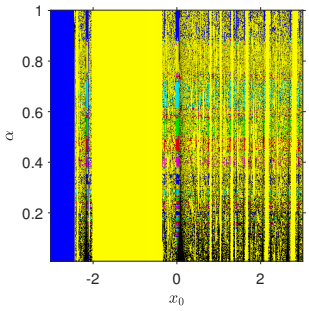
In Table 2 we can observe a similar behavior0. We can see that TeCO can get a solution in less iterations than classical Traub's procedure, and the computational order of convergence is three. We can see that classical Chun-Kim's methos does not encounter any solution, unlike CKeCO. OeCO and CeCO have a similar behavior compared to their classical partners.

Now, we analyze the dependence on initial estimates of these methods. For this aim, we use convergence planes [18]. These planes are constructed with a 400×400 grid; we consider initial estimate in the horizontal axis, and the value of $\alpha \in (0, 1]$ in the vertical axis. Each solution is represented by a distinct color. It is painted in black when no solution has been is found in 500 iterations, and the tolerance is 10^{-3}. The percentage of converging pairs (x_0, α) is also calculated.

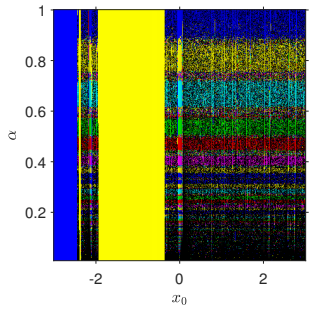
In Figure 1, we can note that NeA3 and OeCO get around 100% of convergence,



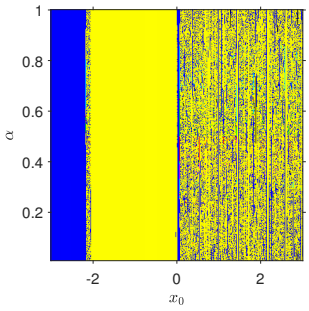
(a) NeA3, 99.76% of convergence



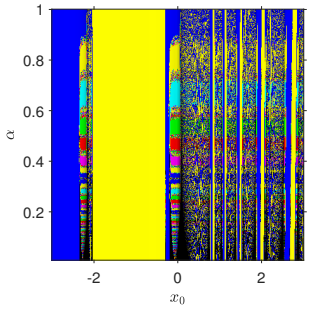
(b) TeCO, 94.63% of convergence



(c) CKeCO, 76.76% of convergence



(d) OeCO, 99.46% of convergence



(e) CeCO, 94.04% of convergence

Fig. 1: Convergence planes for $f(x)$

whereas TeCO and CeCO obtain around 94% of convergence, and CKeCO is around 77% of convergence. Notice that all roots are obtained in all planes.

4 Conclusion

In this work, conformable higher-order Newton-type methods were designed. These schemes improved the convergence of Newton's procedure with integer derivative. Also, it has been designed a general technique in order to get the conformable version of any classical method. The numerical experiments support the theory, obtaining the expected order of convergence in each case, and the dependence on initial estimates was visualized by means of convergence planes. These schemes can find solution when the classical partners fail, and can converge in fewer iterations. They can get complex roots with real initial estimates, and it is possible to obtain distinct roots by choosing distinct values of the fractional order α. Also, these procedures have shown good stability, because of the wideness of basins of attraction of the roots, and because all the were found in all planes.

References

1. Akgül, A., Cordero, A., Torregrosa, J.R.: A fractional Newton method with 2αth-order of convergence and its stability. Applied Mathematics Letters 98, 344–351 (2019)
2. Candelario, G., Cordero, A., Torregrosa, J.R.: Multipoint Fractional Iterative Methods with $(2\alpha + 1)$th-Order of Convergence for Solving Nonlinear Problems. Mathematics 8(3), 452 (2020). https://doi.org/10.3390/math8030452
3. Gdawiec, K., Kotarski, W., Lisowska, A.: Newton's method with fractional derivatives and various iteration processes via visual analysis. Numerical Algorithms 86, 953–1010 (2021). https://doi.org/10.1007/s11075-020-00919-4
4. Torres-Hernandez, A., Brambila-Paz, F., Iturrarán-Viveros, U., Caballero-Cruz, R.: Fractional Newton–Raphson Method Accelerated with Aitken's Method. Axioms 10, 47 (2021). https://doi.org/10.3390/axioms10020047
5. Nayak, S.K., Parida, P.K.: The dynamical analysis of a low computational cost family of higher-order fractional iterative method. International Journal of Computer Mathematics 100:6, 1395–1417 (2023). https://doi.org/10.1080/00207160.2023.2190034
6. Bayrak, M.A., Demir, A., Ozbilge, E.: On Fractional Newton-Type Method for Nonlinear Problems. Journal of Mathematics 2022 (2022). https://doi.org/10.1155/2022/7070253
7. Akgül, A., Grow, D.: Fractal Newton Methods. Mathematics (2023). https://doi.org/10.3390/math11102277
8. Candelario, G., Cordero, A., Torregrosa, J.R., Vassileva, M.P.: An optimal and low computational cost fractional Newton-type method for solving nonlinear equations. Applied Mathematics Letters 124, 107650 (2022). https://doi.org/10.1016/j.aml.2021.107650
9. Candelario, G., Cordero, A., Torregrosa, J.R., Vassileva, M.P.: Generalized conformable fractional Newton-type method for solving nonlinear systems. Numerical Algorithms, (2023). https://doi.org/10.1007/s11075-022-01463-z
10. Khalil, R., Al Horani, M., Yousef, A., Sababheh, M.: A new definition of fractional derivative. Journal of Computational and Applied Mathematics 264, 65–70 (2014)

11. Abdeljawad, T.: On conformable fractional calculus. Journal of Computational and Applied Mathematics 279, 57–66 (2014)
12. Toprakseven, S.: Numerical Solutions of Conformable Fractional Differential Equations by Taylor and Finite Difference Methods. Journal of Natural and Applied Sciences 23, 850–863 (2019)
13. Candelario, G., Cordero, A., Torregrosa, J.R., Vassileva, M.P: Solving Nonlinear Transcendental Equations by Iterative Methods with Conformable Derivatives: A General Approach. Mathematics, (2023). http://dx.doi.org/10.3390/math11112568
14. Traub, J.F.: Iterative Methods for the Solution of Equations. Prentice-Hall, New Jersey (1964)
15. Petković, M.S., Neta, B., Petković, L.D., Džunić, J.: Multipoint Methods for Solving Nonlinear Equations. Elsevier, USA (2013)
16. Chun, C., Kim, Y.I.: Several New Third-Order Iterative Methods for Solving Nonlinear Equations. Acta Appl. Math. 109, 1053–1063 (2010)
17. Cordero, A., Torregrosa, J.R.: Variants of Newton's method using fifth order quadrature formulas. Applied Mathematics and Computation 190, 686–698 (2007)
18. Magreñan, A.Á.: A new tool to study real dynamics: The convergence plane. Applied Mathematics and Computation 248, 215–224 (2014)

On doubly stochastic combined matrices

Begoña Cantó[1], Rafael Cantó[1], and Ana M. Urbano[1]

Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València,
Camí de Vera s/n, 46071 València, Spain

Abstract. In this work, doubly stochastic combined matrices are considered. These matrices have applications in several areas such as Statistic or Control Theory in order to determine the behaviour of the processes. In these areas, the matrices are known as Relative Gain Array and they are characterized by having a small size equal to 2, 3 or 4. In particular, this work answers two questions on combined matrices. The first is the existence of a real matrix of order 3 such that its combined matrix is a given doubly stochastic matrix. The second is the construction of this real matrix. Finally, we apply these results to doubly stochastic Hankel matrices of order 3.

Keywords: combined matrix, doubly stochastic matrix, nonnegative matrix, Hankel matrix

1 Introduction

A real matrix $A = (a_{ij})$ of order n is a doubly stochastic matrix if its entries are nonnegative numbers and the sum of its entries of any row or column is equal to 1, that is $0 \leq a_{ij} \leq 1,\ \forall i, j = 1, 2, \ldots, n,\ \sum_{i=1}^{n} u_{ij} = 1,\ \forall j = 1, 2, \ldots, n$ and $\sum_{j=1}^{n} u_{ij} = 1,\ \forall i = 1, 2, \ldots, n$. These matrices are useful in Statistic and Probability.

A combined matrix of a nonsingular matrix $A = (a_{ij})$ is the matrix $\mathcal{C}(A) = (c_{ij}) = A \circ A^{-T}$ where $\circ$ means the Hadamard (entrywise) product, and A^{-T} means the inverse transpose, $(A^{-1})^T$, of A. These matrices are called Relative Gain Array (RGA) in Control Theory and they are used to determine the best input-output pairing for multivariable process control systems, see [1–3].

Several authors have studied combined matrices of nonsingular matrices. For example, problems of characterization of a combined matrix when the elements of its main diagonal are known are studied in [4, 5], the structure of matrices whose combined matrix has a given structure is obtained in [6, 7], and some algorithms for constructing doubly stochastic matrices are considered in [8].

Usually, the RGA matrices used in Control Theory have a small size, being their size equal to 2, 3 or 4. There are some results on matrices of order 3 given in [4, 9, 10].

In this work we consider real matrices of order 3 and our aim is answer the following questions: *If we have a doubly stochastic matrix of order* 3*, is there*

a real matrix such that its combined matrix is this doubly stochastic matrix? Moreover, if it exists, what is its structure?

2 Results

Firstly, we give the following properties of combined matrices that we use to answer the questions given in the Introduction, see [11].

Lemma 1. *The combined matrix $\mathcal{C}(A) = (c_{ij})$ of a nonsingular matrix $A = (a_{ij})$ satisfies*

(a) If D_1 and D_2 are two nonsingular diagonal matrices then $\mathcal{C}(A) = \mathcal{C}(D_1 A D_2)$.
(b) If P and Q are two permutation matrices and G is any triangular matrix then
  (b.1) $\mathcal{C}(PAQ) = P\mathcal{C}(A)Q$.
  (b.2) $\mathcal{C}(PGQ) = P\mathcal{C}(G)Q = PQ$.
  (b.3) $\mathcal{C}(PQP^T) = I$.
(c) If $c_{ij} \neq 0$ then $a_{ij} \neq 0$, $\forall\ i, j$.
(d) The sum of the entries of any row or column of $\mathcal{C}(A)$ is 1.
(e) $\mathcal{C}(A^{-T}) = \mathcal{C}(A)$.

We answer the first question when the doubly stochastic matrix has not zero entries in its first row and its first column.

Theorem 1. *Let $U = (u_{ij})$ be a doubly stochastic matrix of order 3 without zero entries in its first row and its first column. Then, there exists a nonsingular matrix $A = (a_{ij})$, such that $\mathcal{C}(A) = U$, if and only if the polynomial*

$$P(x) = u_{11}x^2 + (u_{22} - u_{13}u_{31} - u_{33}u_{11})x + u_{33}u_{13}u_{31} \tag{1}$$

has at least a real root different from $\dfrac{-u_{13}u_{31}}{u_{21} + u_{31}}$.

Note that if $P(x)$ has a unique root, there exists a real matrix A such that $\mathcal{C}(A) = U$. By Lemma 1 any other matrix such that its combined matrix is U belongs to the set

$$F_U(A) = \{A,\ A^{-T},\ \widetilde{D}_1 A \widetilde{D}_2,\ \widetilde{D}_1 A^{-T} \widetilde{D}_2\}, \tag{2}$$

where $\widetilde{D}_1$ and $\widetilde{D}_2$ are nonsingular diagonal matrices. If P(x) has two roots, there exist two real matrices A_i, $i = 1, 2$, with the same combined matrix U. In this case any other matrix such that its combined is U belongs to one of the sets $F_U(A_1)$ or $F_U(A_2)$.

The following example clarifies the above result.

Example 1. Consider the doubly stochastic matrix

$$U = \begin{pmatrix} 0.5000 & 0.4375 & 0.0625 \\ 0.4375 & 0.1250 & 0.4375 \\ 0.0625 & 0.4375 & 0.5000 \end{pmatrix}.$$

The second degree polynomial given in (1) is

$$P(x) = u_{11}x^2 + (u_{22} - u_{13}u_{31} - u_{33}u_{11})x + u_{33}u_{13}u_{31} = 0.5x^2 - 0.128x + 0.002,$$

with positive discriminant,

$$\Delta = (u_{22} - u_{13}u_{31} - u_{33}u_{11})^2 - 4u_{11}u_{33}u_{13}u_{31} = 0.0127 > 0.$$

Therefore, it has two real roots $x_1 = 0.2416$ and $x_2 = 0.0162$, both different from $\dfrac{-u_{13}u_{31}}{u_{21} + u_{31}} = -0.0078$. According to Theorem 1, for each root x_i, there exists a nonsingular matrix A_i such that $\mathcal{C}(A_i) = U$, $i = 1, 2$.

The following example is taken from [7]. In that paper, the authors could not obtain a matrix A such that its combined matrix was U, and we reach the same conclusion by applying Theorem 1.

Example 2. Consider the matrix given in [7],

$$U = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}. \tag{3}$$

The polynomial in x given in (1) is

$$P(x) = \frac{1}{3}\, x^2 + \frac{1}{9}\, x + \frac{1}{27},$$

whose discriminant is negative

$$\Delta = (u_{22} - u_{13}u_{31} - u_{33}u_{11})^2 - 4u_{11}u_{33}u_{13}u_{31} = -1/27 < 0.$$

According to Theorem 1, there is no matrix A whose combined matrix is U, because the polynomial $P(x)$ does not have real roots.

Now, we answer the second question posed in this work. Suppose that we have a doubly stochastic matrix U such that there exists a matrix A whose combined matrix is U. The following theorem shows how to obtain the matrix A from a diagonally equivalent matrix T.

Theorem 2. *Let $U = (u_{ij})$ be a doubly stochastic matrix of order 3 without zero entries in its first row and its first column. If there exists a nonsingular matrix A such that $\mathcal{C}(A) = U$, then the matrix A is given by $A = D_1 T D_2$, where D_1 and D_2 are nonsingular diagonal matrices and the matrix T diagonally equivalent to A, is given by*

$$T = \frac{1}{d}\begin{pmatrix} d & d & d \\ d & u_{11}s_{33} - u_{13}u_{31} & u_{11}s_{33} + u_{31}(1-u_{13}) \\ d & u_{11}s_{33} + u_{13}(1-u_{31}) & u_{11}s_{33} - u_{13}u_{31} + u_{13} - u_{21} \end{pmatrix},$$

where $d = -(u_{21} + u_{31})s_{33} - u_{13}u_{31}$ and

$$S = (s_{ij}) = T^{-1} = \begin{pmatrix} u_{11} & u_{21} & u_{31} \\ u_{12} & u_{13} - u_{21} + s_{33} & -u_{31} - s_{33} \\ u_{13} & -u_{13} - s_{33} & s_{33} \end{pmatrix},$$

with s_{33} a real root of the polynomial $P(x)$ given in (1) different from $\dfrac{-u_{13}u_{31}}{u_{21} + u_{31}}$.

The following example clarifies the above result.

Example 3. Consider the doubly stochastic matrix given in Example 1. For each root x_i, $i = 1, 2$, we obtain the matrices S_i and T_i, such that $\mathcal{C}(T_i) = U$. Using, for example, $D_1 = \operatorname{diag}(1, 2, 3)$ and $D_2 = \operatorname{diag}(4, 2, 1)$, we calculate the matrices $A_i = D_1 T_i D_2$ which satisfy that $\mathcal{C}(A_i) = U$.

• For $s_{33} = x_1 = 0.2416$, we have

$$S_1 = \begin{pmatrix} 0.5000 & 0.4375 & 0.0625 \\ 0.4375 & -0.1334 & -0.3041 \\ 0.0625 & -0.3041 & 0.2416 \end{pmatrix}, \qquad T_1 = S_1^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -0.9374 & -1.4384 \\ 1 & -1.4384 & 2.0691 \end{pmatrix},$$

and

$$A_1 = D_1 T_1 D_2 = \begin{pmatrix} 4.0000 & 2.0000 & 1.0000 \\ 8.0000 & -3.7496 & -2.8768 \\ 12.0000 & -8.6304 & 6.2073 \end{pmatrix}.$$

• For $s_{33} = x_2 = 0.0162$ we have

$$S_2 = \begin{pmatrix} 0.5000 & 0.4375 & 0.0625 \\ 0.4375 & -0.3588 & -0.0787 \\ 0.0625 & -0.0787 & 0.0162 \end{pmatrix}, \qquad T_2 = S_2^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -0.3493 & -5.5549 \\ 1 & -5.5549 & 30.8844 \end{pmatrix},$$

and

$$A_2 = D_1 T_2 D_2 = \begin{pmatrix} 4.0000 & 2.0000 & 1.0000 \\ 8.0000 & -1.3972 & -11.1098 \\ 12.0000 & -33.3295 & 92.6533 \end{pmatrix}.$$

Note that any other matrix such that its combined is U belongs to one of the sets $F_U(A_1)$ or $F_U(A_2)$ given in (2).

Note that, if U is a symmetric matrix by construction the matrices S_i and T_i are also symmetric. Moreover if $D_1 = D_2$ then the matrices A_i will be symmetric, but if $D_1 \neq D_2$ then they will not be symmetric.

3 Application to Hankel matrices

An easy way to construct doubly stochastic matrices of order n is from nonnegative Hankel matrices of the same order. We recall that a Hankel matrix, denoted by $H = (h_{ij})$ is a square matrix whose values are constant along the ascending diagonals, i.e., whose entries satisfy the relation $h_{ij} = h_{i-1,j+1}$ for $i = 2, 3, \ldots, n$ and $j = 1, 2, \ldots, n-1$. In particular, the positive Hankel matrix of order 3 has the following form

$$H = \begin{pmatrix} a & b & c \\ b & c & a \\ c & a & b \end{pmatrix}, \qquad \forall a,\, b,\, c > 0.$$

Since the sum of the entries of its rows and its columns is the same, $a+b+c$, we construct a doubly stochastic matrix multiplying the matrix H by the inverse of the sum $a + b + c$, i.e.

$$U = \frac{1}{a+b+c} \begin{pmatrix} a & b & c \\ b & c & a \\ c & a & b \end{pmatrix} = \begin{pmatrix} \dfrac{a}{a+b+c} & \dfrac{b}{a+b+c} & \dfrac{c}{a+b+c} \\ \dfrac{b}{a+b+c} & \dfrac{c}{a+b+c} & \dfrac{a}{a+b+c} \\ \dfrac{c}{a+b+c} & \dfrac{a}{a+b+c} & \dfrac{b}{a+b+c} \end{pmatrix}.$$

The matrix U is the doubly stochastic matrix corresponding to the Hankel matrix H. Note that U is also a Hankel matrix and therefore it is a symmetric matrix. If we permute rows or columns of this matrix then we obtain a nonsymmetric doubly stochastic matrix. Thus, we conclude that we obtain symmetric or nonsymmetric doubly stochastic matrices from nonnegative Hankel matrices.

Now, we consider a doubly stochastic Hankel matrix U obtained from a positive Hankel matrix H. We pose the following question: *If we have a matrix*

U of order 3, is there a real matrix A such that its combined matrix is U? The following example shows that the answer is negative.

Example 4. Consider the real numbers 1, 2 and 3 and construct the following Hankel matrix H and its corresponding doubly stochastic Hankel matrix U,

$$H = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 3 & 2 \\ 3 & 2 & 1 \end{pmatrix}, \qquad U = \frac{1}{6} H = \begin{pmatrix} 1/3 & 1/6 & 1/2 \\ 1/6 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/6 \end{pmatrix}.$$

The polynomial $P(x)$ given in (1) associated with U is

$$P(x) = \frac{1}{3}\, x^2 + \frac{1}{18}\, x + \frac{1}{24},$$

whose discriminant is negative

$$\Delta = (u_{22} - u_{13}u_{31} - u_{33}u_{11})^2 - 4u_{11}u_{33}u_{13}u_{31} = -0.0177469 < 0.$$

From Theorem 1, there is no a real matrix A whose combined matrix is U.

Note that, the matrix given in (3) is a doubly stochastic matrix constructed from the Hankel matrix of order 3 with all its entries equal to one.

The negative answer to the previous question leads us to another question: *Does the distribution of the numbers on the main diagonal of a Hankel matrix H influence the fact that the discriminant of the corresponding polynomial $P(x)$ given in (1) is positive or negative?* The answer is shown in the following result.

Proposition 1. *Let $H = (h_{ij})$ be a Hankel matrix of order 3 and let U be its corresponding doubly stochastic Hankel matrix. The sign of the discriminant of the polynomial $P(x)$, associated with U, is independent of the distribution of the entries h_{ii}, $i = 1, 2, 3$.*

Proof. Let $H = (h_{ij})$ be a Hankel matrix of order 3 and let U be the corresponding doubly stochastic Hankel matrix. To change the order of the entries on the main diagonal of H is equivalent to make a permutation similarity in rows and columns of H and therefore in U. By Lemma 1, if there exists a matrix A such that $\mathcal{C}(A) = U$, then any matrix similar by permutation to A satisfies that its combined matrix is the matrix obtained from the same permutation similarity over U. So, the sign of the discriminant of the polynomial $P(x)$, associated with U, is independent of the distribution of the entries h_{ii}, $i = 1, 2, 3$. □

Consider a set of three positive real numbers $\{a, b, c\}$. We construct the positive Hankel matrix of order 3 with these numbers, and the corresponding doubly stochastic Hankel matrix U. *Is there any condition on the set $\{a, b, c\}$ such that a nonsingular matrix A with $\mathcal{C}(A) = U$ can be obtained?* We answer this question in the following result.

Proposition 2. *Let a, b and c positive real numbers, let $H = (b_{ij})$ be a Hankel matrix of order 3 and let U be the corresponding doubly stochastic Hankel matrix. Then, there exists a nonsingular matrix $A = (a_{ij})$, such that $\mathcal{C}(A) = U$ if:*

(i) $a = b \neq c$, *with* $c \leq a/4$.

(ii) $a \neq b$ *and* $c \in \left]0, \dfrac{ab}{\left(\sqrt{a}+\sqrt{b}\right)^2}\right] \cup \left[\dfrac{ab}{\left(\sqrt{a}-\sqrt{b}\right)^2}, +\infty\right[$.

Proof. We consider three positive real numbers a, b and c. Regardless of the distribution, if they are on the main diagonal of a Hankel matrix H, the discriminant of the polynomial $P(x)$ given in (1) is

$$\triangle = \frac{a^2b(b-2c) + b^2c(c-2a) + c^2a(a-2b)}{(a+b+c)^4}.$$

So, in order to ensure $\triangle \geq 0$, a, b and c must satisfy the following inequality,

$$a^2b(b-2c) + b^2c(c-2a) + c^2a(a-2b) \geq 0. \tag{4}$$

If a and b are known positive real numbers, to obtain the values for c that equation (4) is satisfied, we distinguish the following cases:

(1) $a = b$

(1.1) $a = b = c$

In this case the Hankel matrix H and the corresponding doubly stochastic Hankel matrix U are

$$H = \begin{pmatrix} a & a & a \\ a & a & a \\ a & a & a \end{pmatrix}, \qquad U = \frac{1}{3a}H = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

As we have seen in Example 3, for this matrix U there is no matrix A whose combined matrix is U.

(1.2) $a = b \neq c$

Now, we have that

$$a^2b(b-2c) + b^2c(c-2a) + c^2a(a-2b) = a^3(a-4c).$$

Therefore,

$$\triangle \geq 0 \;\Leftrightarrow\; c \leq \frac{a}{4}.$$

If $c = \dfrac{a}{4}$ then we obtain a nonsingular matrix A, such that $\mathcal{C}(A) = U$.

If $c < \dfrac{a}{4}$ we obtain two nonsingular matrices A_i, such that $\mathcal{C}(A_i) = U$, $i = 1, 2$.

(2) $a \neq b$
Equation (4) can be written in the following form

$$(a^2 - 2ab + b^2)c^2 - (2a^2b + 2ab^2)c + a^2b^2 \geq 0.$$

This inequality is satisfied if

$$c \in \left]0, \quad \frac{ab}{\left(\sqrt{a}+\sqrt{b}\right)^2}\right] \cup \left[\frac{ab}{\left(\sqrt{a}-\sqrt{b}\right)^2}, \quad +\infty\right[.$$

If $c = \dfrac{ab}{\left(\sqrt{a}+\sqrt{b}\right)^2}$ or $c = \dfrac{ab}{\left(\sqrt{a}-\sqrt{b}\right)^2}$ then there exists a matrix A such that $\mathcal{C}(A) = U$. In other case, there exist two nonsingular matrices A_i, such that $\mathcal{C}(A_i) = U$, $i = 1, 2$.

We finalize this work with the following example.

Example 5.

- Consider the real numbers $a = b = 8$ and $c = 1$. As $a = b$ and $c \leq a/4$ by Proposition 2, there exists a matrix A such that $\mathcal{C}(A) = U$. Without loss of generality, we assume that the Hankel matrix H and the corresponding doubly stochastic Hankel matrix U are

$$H = \begin{pmatrix} 8 & 8 & 1 \\ 8 & 1 & 8 \\ 1 & 8 & 8 \end{pmatrix}, \quad U = \frac{1}{17}H = \begin{pmatrix} 0.4706 & 0.4706 & 0.0588 \\ 0.4706 & 0.0588 & 0.4706 \\ 0.0588 & 0.4706 & 0.4706 \end{pmatrix}.$$

For U the polynomial $P(x)$ given in (1) is

$$P(x) = 0.4706\,x^2 - 0.1661\,x + 0.0016$$

whose discriminant is $\triangle = 0.0245 > 0$ and the two real roots are $x_1 = 0.3428$ and $x_2 = 0.0101$. Applying Theorem 2 we have

$$x_1 = 0.3428 \;\rightarrow\; T_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -0.8536 & -1.1716 \\ 1 & -1.1716 & 1.3726 \end{pmatrix} \quad \text{and} \quad A_1 = D_1T_1D_2$$

$$x_2 = 0.0101 \;\rightarrow\; T_2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -0.1464 & -6.8284 \\ 1 & -6.8284 & 46.6274 \end{pmatrix} \quad \text{and} \quad A_2 = D_1T_2D_2,$$

where D_1 and D_2 are arbitrary nonsingular diagonal matrices.

If the distribution on the main diagonal of a, b and c is different, i.e. $\tilde{H} = PHP^{-1}$, where P is a permutation matrix, then the matrices whose combined matrix is U are $\tilde{A}_1 = PA_1P^{-1}$ and $\tilde{A}_2 = PA_2P^{-1}$.

- Now, consider $a = 9$ and $b = 4$. If $c \in]0,\ 1.44] \cup [36,\ +\infty[$ by Proposition 2, the doubly stochastic matrix U satisfies that there exists a matrix A such that $\mathcal{C}(A) = U$.

  Suppose that $c = 36$, then the Hankel matrix H and the corresponding doubly stochastic Hankel matrix U are

$$H = \begin{pmatrix} 9 & 4 & 36 \\ 4 & 36 & 9 \\ 36 & 9 & 4 \end{pmatrix}, \quad U = \frac{1}{17} H = \begin{pmatrix} 0.1837 & 0.0816 & 0.7347 \\ 0.0816 & 0.7347 & 0.1837 \\ 0.7347 & 0.1837 & 0.0816 \end{pmatrix}.$$

  For U the polynomial $P(x)$ given in (1) is

$$P(x) = 0.1837\,x^2 + 0.1799\,x + 0.0441$$

  whose discriminant is $\triangle = 0$ and the unique real root of $P(x)$ is $x_1 = -0.4898$. Applying Theorem 2 we have

$$T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 4.5000 & -0.7500 \\ 1 & -0.7500 & -0.1667 \end{pmatrix} \quad \text{and} \quad A = D_1 T D_2$$

  where D_1 and D_2 are arbitrary nonsingular diagonal matrices. If the distribution of the main diagonal is different, we obtain $\widetilde{A}$ as in the previous case.

4 Conclusions

In this work we consider a doubly stochastic matrix U of order 3, without zero entries in its first row and its first column, and we present a result to assure that there exists a real matrix A of order 3 such that its combined matrix is U. In addition, we give a method to obtain A and the set of all real matrices such that their combined matrix is also U. We apply these results to doubly stochastic matrices obtained from Hankel matrices of order 3.

Acknowledgements

This research was partially supported by Consellería de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana under Spanish grant CIAICO/2021/162.

References

1. Bristol, R.: On a new measure of interaction for multivariable process control. Automatica 11(1), 123–134 (1966). doi:10.1109/TAC.1966.1098266
2. Chiu, M.-S.: A methodology for the synthesis of robust decentralized control systems [dissertation]. Atlanta GA: Georgia Institute of Technology, (1991). doi.org/10.1016/0005-1098(92)90153-7
3. Hovd, M., Skogestad, S.: Sequencial Desig of Descentralized Controllers. Automatica 30(10):1601–1607 (1994). doi.org/10.1016/0005-1098(94)90099-X
4. Bru,R., Gassó, M.-T., Giménez, I., Santana, M.: Diagonal entries of the combined matrix of a totally negative matrix. Linear and Multilinear Algebra 65(10):1971–1984 (2017). doi.org/10.1080/03081087.2016.1261079
5. Fiedler, M.,Markham, T.-L.: Combined matrices in special classes of matrices. Linear Algebra and its Applications 435:1945–1955 (2011). doi.org/10.1016/j.laa.2011.03.054
6. Bru, R., Gassó, M.-T., Santana, M.: Combined matrices and conditioning. Applied Mathematics and Computation 412:1–8 (2022). doi.org/10.1016/j.amc.2021.126549
7. Johnson, C.-R., Shapiro, H.: Mathematical aspects of the relative gain array. SIAM J. Alg. Disc. Math. 7(4):627–644 (1986). doi.org/10.1137/0607069
8. Fuster,R., Gassó, M.-T., Giménez, I.: CMMSE algorithms for constructing doubly stochastic matrices with the relative gain array (combined matrix). J. Math. Chemistry 57:1700–1709 (2019). doi.org/10.1007/s10910-019-01032-1
9. Gassó, M.-T., Gil, I., Giménez, I., Santana, M., Segura, E.: Diagonal entries of the combined matrix of sign regular matrices of order three. Proyecciones. Journal of Mathematics 40(1):255–271 (2021). doi.org/10.22199/issn.0717-6279-2021-01-0016
10. Fiedler, M., Markham, T.-L.: An inequality for the Hadamard product of an M-matrix and inverse M-matrix. Linear Algebra and its Applications 101:1–8 (1988). doi.org/10.1016/0024-3795(88)90139-5
11. Horn, R.-A., Johnson, C.-R.: Topics in matrix Analysis. Cambridge University Press (1991).

Entropy estimation from horizontal visibility graphs

Òscar Garibo-i-Orts[1], Andrei Velichko[2], and J. Alberto Conejero[1]

[1] Instituto Univrsitario de Matemática Pura y Aplicada.
Universitat Politècnica de València. València. Spain
osgaor@upv.es,aconejero@upv.es,
[2] Institute of Physics and Technology, Petrozavodsk State University
185910 Petrozavodsk, Russia
velichko@petrsu.ru

Abstract. Artificial intelligence methods have been recently used for estimating the predictability and complexity of a time series through an entropy computation. This approach overcomes the parametric dependence of these methods. Time series can also be represented through the computation of horizontal visibility graphs. In this work, we show how a LogNNet neural network model also provides reliable results when fed with the sequence of degrees of the visibility graph instead of the time series itself.

Keywords: entropy; time series; neural networks; logistic equation; LogNNet; NNetEN; classification; natural visibility graphs; horizontal visibility graphs

1 Introduction

Being able to measure the complexity of time series helps to design and control nonlinear systems. There are several methods, both mathematical and statistical, which help to measure the degree of complexity of data represented as time series. Among other methods, we have the Kolmogorov complexity measure [14], the C1/C2 complexity measure [14] or the permutation entropy [15]. Recently, Velichko proposed the use of a LogNNet neural network for IoT applications with limited available memory [6, 10]. This is a feedforward neural network that uses filters based on logistic function and a reservoir inspired by recurrent neural networks. This enables to transform a signal into a high-dimensional space. Its efficiency was validated on the MNIST-10 data set [8]. The LogNNet is The extraction of features and the use of recurrent neural networks are very effective for processing physical properties of signals, as is the case of the combination of convolutional networks with Long Short Memory networks [3] or transformers [2] for analyzing anomalous diffusion. LogNNet is very competitive because the low level of memory requirements, since it consists of a weights multiplication layer and a dense layer (multilayer perceptron) to compute the output. Its efficiency was outperformed when replacing the the logistic map by the semi-linear a Henon

type discrete map [5]. As a matter of fact, they showed that the classification performance is proportional to the entropy of the time series and has a stronger correlation than the Lyapunov exponent of the time series used to feed the reservoir. Because of this property, Velichko and Heidari showed that LogNNet architecture permits estimating the entropy of a time series just by looking at the classification accuracy when we feed the reservoir with the time series values. Such an approach has the advantage over other methods: it does not depend on the choice of parameters.

Visibility graphs were introduced in [7] as a simple and fast computational way to represent a time series as a graph. Visibility graphs inherit the periodic and chaotic properties of the logistic equation [9]. For example, periodic series result in regular graphs, random series in random graphs, and fractal series in scale-free graphs. A *natural visibility graph* (NVG) is constructed as follows. Let $\{(t_i, y_i)\}_{i\in I}$ a time series indexed on I, where the series at time t_i takes the value y_i. We associate a node to each pair (t_i, y_i) to convert this series into a graph. Given two nodes (t_a, y_a) and (t_b, y_b), we consider that these two points have visibility and therefore, they are connected by an edge if any other pair (t_c, y_c) with $a < c < b$ satisfies

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}. \tag{1}$$

Horizontal Visibility Graphs (HVG) were introduced in [9] to simplify the previously described NVG. As with the NVGs, for computing the HVG, each value in the time series is assigned a node in the resulting graph. Two nodes (t_a, y_a) and (t_b, y_b) in this graph result in being connected if a horizontal line can be drawn connecting these two points without intersecting any intermediate value that is, if $y_a, y_b > y_c$ for all $a < c < b$, see the examples in Figure 1.

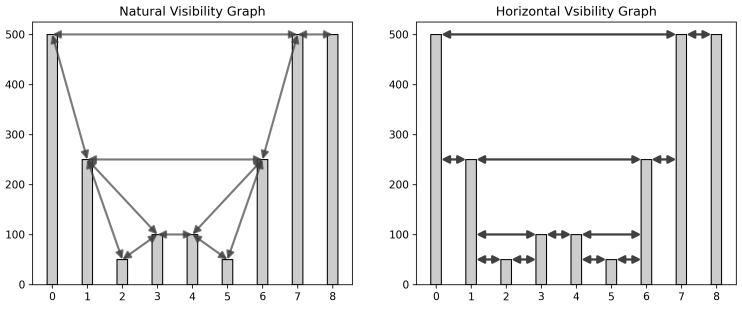


Fig. 1: Ilustrative example of the NVG representation for a time serie (left) and the HVG representation for the same time serie (right).

In this work, we reproduce the work in [12], but instead of feeding the LogNNet weight matrix with the logistic map time series, we will use the HVG de-

grees representation of a logistic time series, which implies that the resulting time series does not consist of real numbers, but only of integer numbers. We will analyze how accurate is to look at the classification accuracy to estimate the entropy of the HVG degree sequence, which should resemble the entropy of the given logistic time series. This paper is organized as follows: In Section 2, we introduce the methods we have used, while in Section 3 we explain the results we obtained, and finally, in Section 4 we state the conclusions, and we outline some ideas for future works.

2 Methodology

We generate trajectories from the logistic map $x_{n+1} = rx_n(1 - x_n)$ for $r \in [3.4, 4.0]$ with step size 0.01 and initial condition $x_0 = 0.3$ Then we compute the degrees of the HVG associated to each trajectory. The degree sequences will be the ones whose entropy will be analyzed using the LogNNET architecture.

As we have said, the LogNNET architecture permits us to estimate the entropy of a time series when we use it to fill the reservoir of the network just by looking at the classification accuracy on the MNIST-10 data set [12]. The MNIST-10 data set consists of images depicting hand-written digits from 0 to 9 and has consistently been used to benchmark image classification methods. In Figure 2, we show a sample of the images in MNIST-10.

Fig. 2: Sample images from the MNIST data set extracted from Wikipedia https://en.wikipedia.org/wiki/MNIST_database.

The LogNNet architecture has two main parts: the model reservoir and the output layer, see Figure 3. The model reservoir changes the space of representation of the input images, while the output layer learns the weights matrix W_2 to fit the true labels. In our case, the weights matrix W_1 in the reservoir is loaded with the HVG degrees of logistic time series, whose entropy we want to estimate.

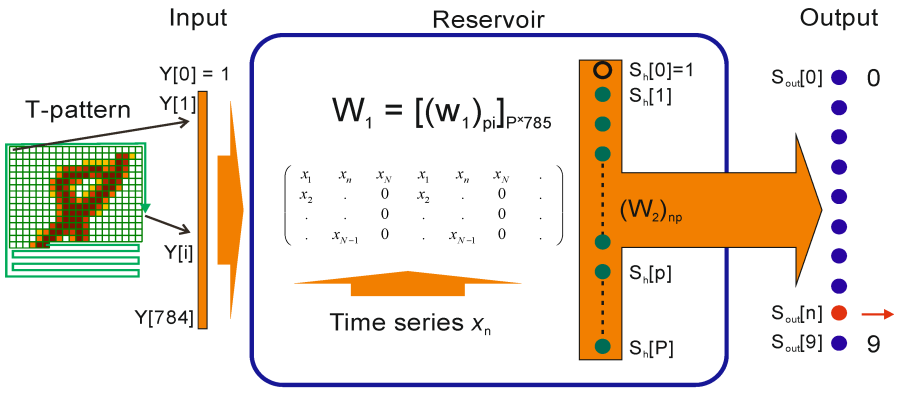


Fig. 3: LogNNet model arquitecture extracted from [12].

The input images are converted into the input vectors Y_n while the HVG time series are used to fill the W_1 weights matrix. In our case, the matrix is filled with HVG sequences. As described in [12], the usage of the LogNNet follows the algorithm shown in Algorithm 1:

Algorithm 1 LogNNet algorithm

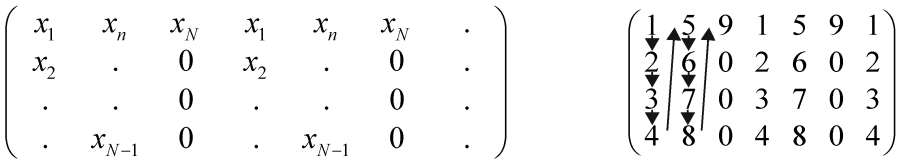
1. Load an HVG degree sequence and denote it as X_n.
2. Load MNIST images data set and transform the images into input vectors Y.
3. The output layer's weights vector W_2 is initialized to a constant value for all its elements of 0.5 for the sake of reproducibility when computing entropy.
4. The reservoir weights matrix W_1 is initialized using X_n.
5. Compute the coefficients for normalization.
6. Set the number of training epochs.
7. Train the W_2 weights by iterating over the input data for the selected number of epochs.
8. Evaluate the trained model with the MNIST-10 test data set.
9. Compute the model's accuracy.

The accuracy is considered the entropy measure and denoted by NNetEn, or entropy inferred by the LogNNet architecture. The matrix W_1 affects the resulting representation of the input vectors since the input vector representation is a matrix resulting from multiplying the input vector Y by the W_1 weights matrix. In this work, we consider two methods to fill the LogNNet reservoir, and we will evaluate how both methods perform.

- **Method 1**: The W_1 matrix is filled row by row with the HVG degrees time series,

- **Method 2**: The W_1 matrix is filled column by column with the HVG degrees time series as proposed in [12].

It is worth mentioning that for dealing with MNIST data set, the reservoir matrix W_1 has 25 rows and 785 columns. Consequently, 19,625 elements are needed in order to fill the W_1 matrix. Nevertheless, the length of a given time series selected to fill the W_1 matrix (becoming the filters of the neural network) may have more or fewer elements than this amount. In order to cope with this, in [12], it is recommended that given a time series of length N, to keep the last 19,625 elements in the time series if $N > 19,625$. If the time series length $N < 19,625$, Method 2 must be selected to fill the W_1 matrix (column-wise). The W_1 matrix is then fed column-wise until the number of elements in the time series is exhausted, then the column being filled is completed with zeros, and the time series is used again from the beginning to fill the next column. Should the time series be used up again before the W_1 matrix is complete, the same operation should be repeated until it is completely filled. In Figure 4, we depict an example for clarification.



(**a**) (**b**)

Fig. 4: Structure of the filling method used when the length of the time series $N < 16,625$ (a) and an toy example in a simple example with didactic goal in which W_1 has size (4x7) and the time series used to fill it has length 9. Figure extracted from [12].

In addition to entropy, we calculate the learning inertia (LI_{Ep_1/Ep_2}) as introduced in [12]. Learning inertia reflects how fast the entropy inferred by the LogNNet (NNetEN) converges to the plateau with respect to the number of training epochs and is calculated as follows:

$$\mathrm{LI}(Ep_1/Ep_2) = \frac{\mathrm{NNetEn}(\mathrm{epoch}Ep_2) - \mathrm{NNetEn}(\mathrm{epoch}Ep_1)}{\mathrm{NNetEn}(\mathrm{epoch}Ep_2)}, \tag{2}$$

being $Ep_1 < Ep_2$ two learning epochs used to infer the entropy.

3 Results

First, we use Method 1 to fill the reservoir matrix. We train the LogNNet network for 1 and 20 epochs and compute the resulting entropy. In Figure 5, we can

see that after 20 epochs the model's best performance remains below 60% of accuracy, or in other words, the entropy is below 0.60. As commented before, the best entropy results are achieved when the reservoir matrix W_1 is fed with HVG degrees time series from the logistic map's chaotic regions.

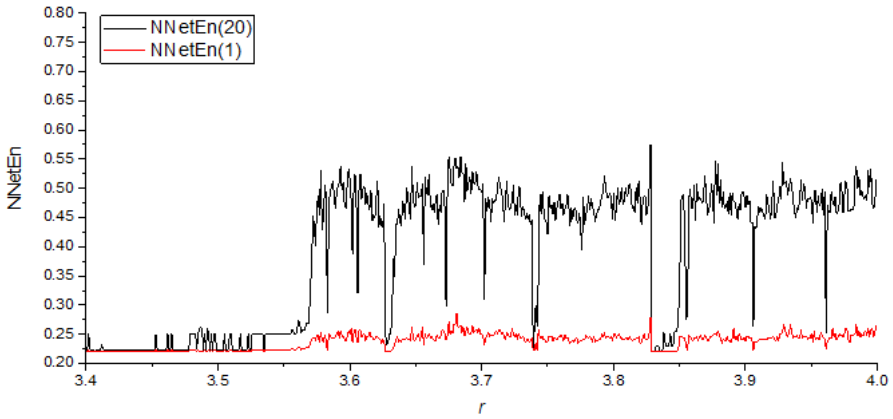


Fig. 5: NNetEn (entropy) inferred by the LogNNet model when the W_1 reservoir matrix is fed using Method 1 (row-wise).

Then, we used Method2 and consistently achieved better entropy values. LogNNet was trained for 1, 20, and 100 epochs in this case , as shown in Figure 6(a) and (c). Again, the best entropy values result from using HVG degrees time series from the chaotic regions of the logistic map. Since the resulting entropy at 20 epochs was better than the one obtained with the same number of epochs using Method 1 (slightly above 0.65 using Method 2 vs. below 0.60 using Method 1), we trained the LogNNet for 100 epochs to check if increasing the learning epochs increased the associated entropy. The learning inertia between both epochs is shown in Figure 6(b). We also see that the learning inertia decreases in the regions were NNetEn is low. We also see that abrupt changes in the accuracy result into sharp changes of the NNetEn.

We concluded that training for more epochs increased the accuracy (or entropy) to values above 0.70 but multiplying the training time by 5. Additionally, if compared to the results obtained for the logistic map in [12] for 100 epochs, we can not appreciate significative differences in the performance with respect to entropy estimation either the logistic map time series or the HVG degrees time series are used to feed the reservoir matrix.

4 Conclusions

In this work, we have shown that using the HVG degrees time series obtained from the logistic map to feed the W_1 matrix in the reservoir of the LogNNet to

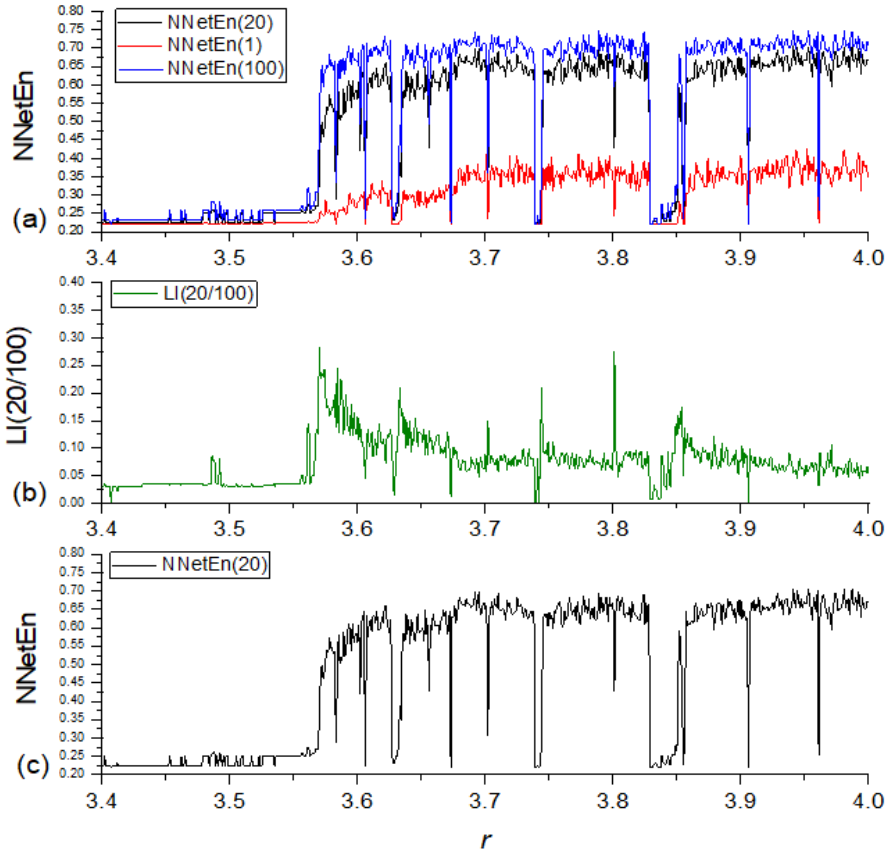


Fig. 6: (a) NNetEn (entropy) inferred by the LogNNet model when the W_1 reservoir matrix is fed with Method 2 (column-wise) after 1, 20, and 100 epochs, (b) the learning inertia $LI(Ep_1/Ep_2)$, and (c) NNetEn achieved by the LogNNet after 20 epochs.

infer the entropy and using the logistic map itself achieves a similar performance when inferring the entropy. Moreover, both time series (logistic map and HVG degrees from the logistic map) result in the best entropy in the range of values of the r parameter from the logistic map for the zone with chaotic behavior.

As future research lines, it would be interesting to study other structural features obtained from chaotic time series and in the case of short sequences, what is the best way to fill the reservoir matrix, beyond these two methods [4]. These methods are available on a Python package [11]. Besides, it will be interesting to explore the behavior of other chaotic maps, as it is the case of the fractional chaotic maps [13] and what machine learning can tell us about these fractional logistic maps [1].

References

1. J.A. Conejero, Ò. Garibo-i Orts, and C. Lizama. Inferring the fractional nature of wu baleanu trajectories. *Nonlinear Dyn.*, 111:12421–12431, 2023.
2. N. Firbas, Ò. Garibo-i Orts, M.A. Garcia-March, and J.A. Conejero. Characterization of anomalous diffusion through convolutional transformers. *J. Phys. A: Math. Theor.*, 56(1):014001, 2023.
3. Ò. Garibo-i Orts, A. Baeza-Bosca, M.A. Garcia-March, and J.A. Conejero. Efficient recurrent neural network methods for anomalously diffusing single particle short and noisy trajectories. *J. Phys. A: Math. Theor.*, 54(50):504002, 2021.
4. H. Heidari, A. Velichko, M. Murugappan, and M.E.H. Chowdhury. Novel techniques for improving nneten entropy calculation for short and noisy time series. *Nonlinear Dyn.*, 111(10):9305–9326, 2023.
5. H. Heidari and A.A. Velichko. An improved LogNNet classifier for IoT applications. In *J. Phys.: Conf. Ser.*, volume 2094, page 032015. IOP Publishing, 2021.
6. Y.A. Izotov, A.A. Velichko, A.A. Ivshin, and R.E. Novitskiy. Recognition of handwritten MNIST digits on low-memory 2 Kb RAM Arduino board using LogNNet reservoir neural network. In *IOP Conference Series: Materials Science and Engineering*, volume 1155, page 012056. IOP Publishing, 2021.
7. L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J.C. Nuno. From time series to complex networks: The visibility graph. *Proc. Natl. Acad. Sci. U.S.A.*, 105(13):4972–4975, 2008.
8. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
9. B. Luque, L. Lacasa, F. Ballesteros, and J. Luque. Horizontal visibility graphs: Exact results for random time series. *Phys. Rev. E*, 80(4):046103, 2009.
10. A. Velichko. Neural network for low-memory IoT devices and MNIST image recognition using kernels based on logistic map. *Electronics*, 9(9):1432, 2020.
11. A. Velichko, M. Belyaev, Y. Izotov, M. Murugappan, and H. Heidari. Neural network entropy (NNetEn): Entropy-based EEG signal and chaotic time series classification, Python package for NNetEn calculation. *Algorithms*, 16(5), 2023.
12. A. Velichko and H. Heidari. A method for estimating the entropy of time series using artificial neural networks. *Entropy*, 23(11):1432, 2021.
13. G.-C. Wu and D. Baleanu. Discrete fractional logistic map and its chaos. *Nonlinear Dyn.*, 75:283–287, 2014.
14. L. Yi and F. Yingle. Complexity measure applied to the analysis EEG signals. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 4610–4613. IEEE, 2006.
15. L. Zunino, F. Olivares, F. Scholkmann, and O.A. Rosso. Permutation entropy based time series analysis: Equalities in the input signal can lead to false conclusions. *Phys. lett. A*, 381(22):1883–1892, 2017.

Study of the semilocal convergence and dynamical behaviour for a modified Newton's method to solve nonlinear systems with singularities

Eva G. Villalba[1], M. A. Hernández-Veró[2], and Eulalia Martínez[1]

[1] Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València. València, Spain.
egarvil@posgrado.upv.es, eumarti@mat.upv.es
[2] Department of Mathematicas and Computation,
Universidad de La Rioja, Logroño, Spain.
mahernan@unirioja.es

Introduction

In numerical analysis it is common to solve systems of nonlinear equations of the form $H(w) = 0$, where $H(w) = (h_1(w), \ldots, h_m(w))$ is an operator defined on $D \subset \mathbb{R}^m$, $H : D \to \mathbb{R}^m$, where $h_i : D \subset \mathbb{R}^m \to \mathbb{R}$, with $1 \leq i \leq m$. In many cases, it is not easy to obtain the exact solution of this problems and, for this reason, we usually use iterative methods to approximate the solution.

One of the most used iterative schemes is Newton's method,

$$\begin{cases} w^{(0)} \in D, \\ w^{(n+1)} = w^{(n)} - H'(w^{(n)})^{-1} H(w^{(n)}), \, n \geq 0 \end{cases} \tag{1}$$

which, under certain conditions, is efficient, simple, has a good dynamical behaviour and converges quadratically in each step. However, the applications of Newton's method and its good qualities are restricted by some conditions such as: the need to iterate the method from an initial estimate close to the solution and the condition that the Jacobian matrix $H'(w^{(n)})$ needs to be nonsingular in an environment of solution. This last condition is necessary in order to the inverse $H'(w^{(n)})^{-1}$ exists and we are able to define Newton's method correctly.

For this reason, in papers such as [2]- [4] the authors propose some variants of Newton's method that converge quadratically to the solution despite the fact that the Jacobian matrix has come singularities.

In [2], the authors introduce a variant of Newton's method that works properly for a special case of systems with multiplicities, including a weight matrix in the iterative function.

In particular, in [3], the modified Newton's iterative scheme for systems, defined by the following iterative scheme, is proposed:

$$\begin{cases} w^{(0)} \in D, \\ w^{(n+1)} = w^{(n)} - [M(w^{(n)}) + H'(w^{(n)})]^{-1} H(w^{(n)}),\ n \geq 0, \end{cases} \tag{2}$$

where the components of the matrix $M(w^{(n)})$ are given by

$$(M(w^{(n)}))_{ij} = \begin{cases} \lambda_i^{(n)} h_i(w^{(n)}) \text{ if } i = j \\ 0 \qquad\quad \text{if } i \neq j \end{cases}$$

where $\lambda_i^{(n)} \in \mathbb{R}$, $0 < |\lambda_i^{(n)}| < +\infty$ with $i = 1, \ldots, m$ and $n = 0, 1, 2, \ldots$, is the i-th component of the variable vector $\lambda^{(n)}$, chosen so that the matrix $M(w^{(n)}) + H'(w^{(n)})$ becomes nonsingular in the domain D.

Our main objective in this work is to compare dynamically Newton's and Newton's modified methods when they are applied to systems whose Jacobians are singular in some points and, after that, we study the semilocal convergence of the method (2). We conclude the study with a numerical analysis in which we apply the modified Newton's method to solve different numerical examples where Newton's scheme is not applicable due to the presence of singularities.

The applicability of the method

We consider now the following example from the paper [5] in order to show the applicability of the studied iterative method for large nonlinear systems with singularities.

$$\begin{cases} w_i \sin(w_{i+1}) - 1 = 0,\ 1 \leq i \leq m-1, \\ w_m \sin(w_1) - 1 = 0. \end{cases} \tag{3}$$

For solving this nonlinear system, we use variable precision arithmetic with 100 digits and different initial guesses in order to reach the distance between consecutive iterations less than the tolerance 10^{-25}, that is we solve this example for some values of m comparing the behaviour of Newton's method and modified Newton's method (2). In this case we use λ as a constant vector of size m.

Firstly, we consider $m = 2$, the nonlinear system has infinite solutions as we can appreciate in Figure 1. If we concentrate in the solutions around the coordinates origin we have $s_1 = (1.1142\ldots, 1.1142\ldots)$ and the symmetric $s_2 = -s_1$.

We can check in Table 1 the numerical results starting by different initial estimates. Notice that, Newton's scheme does not converge in some cases because the jacobian is singular, but we find an adequate value of parameter λ for which modified Newton's method converges. In this Table, the second column represents the starting guess, the second one describes the parameter λ used as

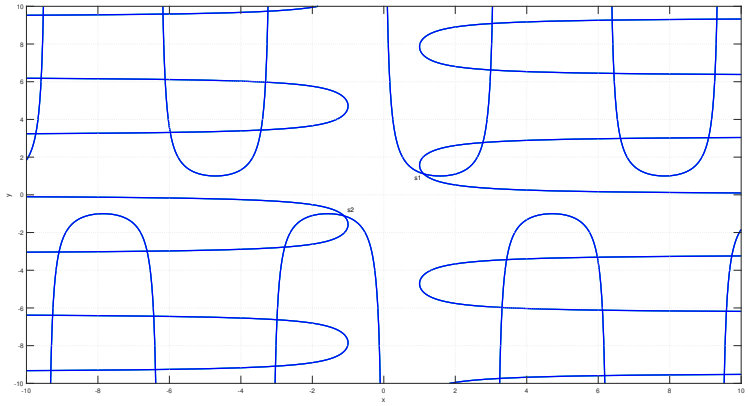


Fig. 1: Representation of nonlinear system (3) for $n = 2$.

.

a constant vector in the corresponding domain, the following is the number of iterations for reaching the tolerance, the following two columns are the distance between the last two consecutive iterates and the value of the nonlinear operator H at last iterate, the seventh is the approximated computational order of convergencence (defined in [1]) and, finally, the las one represents the approximated solution.

Table 1: Convergence results for Example 4 with $m = 2$.

Method	$w^{(0)}$	λ	iter	$\|w_m^{(k+1)} - w_m^{(k)}\|$	$\|H(w^{(k+1)})\|$	p	Sol
Newton (1)	$(0,0)$	–	–	–	–	–	–
Mod. Newton (2)	$(0,0)$	-1	6	$1.4542e-31$	$2.0195e-31$	2	s_1
Mod. Newton (2)	$(0,0)$	1	6	$1.4542e-31$	$2.0195e-31$	2	s_2
Newton (1)	$(0.9,0.2)$	–	12	$2.4496e-26$	$1.1518e-26$	1.9781	s_2
Mod. Newton (2)	$(0.9,0.2)$	-0.2	6	$2.4496e-26$	$7.2143e-28$	1.9992	s_1
Newton (1)	$(1,0)$	–	–	–	–	–	–
Mod. Newton (2)	$(1,0)$	-0.2	7	$3.2984e-32$	$3.0995e-32$	2	s_1

Now, we solve the nonlinear system consider for $m = 3$. We can see in Figure 2 the representation of the 3 surfaces having infinite solutions. The behaviour to the iterative methods in order to find solutions $s_1 = (1.1142\ldots, 1.1142\ldots, 1.1142\ldots)$ and the symmetric $s_2 = -s_1$ is analogous to the case $n = 2$, so we look for different solutions, for example, starting from $w^{(0)} = (-1, 1, 0.5)$, Newton's method does not converge while the modified Newton's method goes to the solution

$s_3 = 9.7165\ldots, 3.0385\ldots, -3.4770\ldots$. See Table 2 for contrasting numerical results.

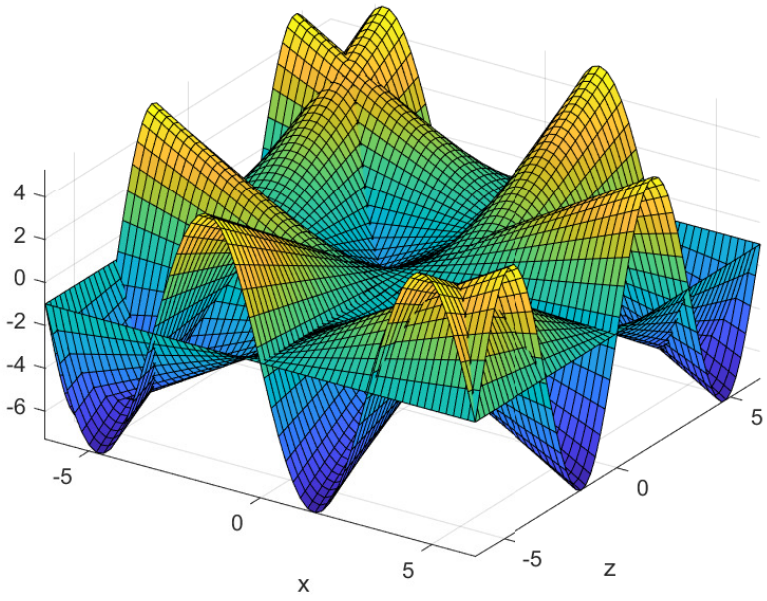


Fig. 2: Representation of nonlinear system (3) for $m = 3$.

.

We can check in Table 2 the numerical results starting by different initial guesses, again we have that Newton's method does not converge because the jacobian is singular while the modified version works properly.

Finally we solve this example for $m = 50$ with the aim of analyzing the behaviour of these methods for bigger size problems. The big system has also infinite solutions, being s_1 and s_2 similar to the previous cases but with the corresponding 50 components. We take as initial estimate the following vector of size 50: $w_a^{(0)} = [w_1^{(0)} w_2^{(0)}]$, where $w_1^{(0)}$ and $w_2^{(0)}$ are subvectors of sizes 20 and 30, respectively, defined as $w_1^{(0)} = ones(1, 20)$ and $w_2^{(0)} = 1/4 * ones(1, 30)$. In Table 2, we compare the numerical results as in the previous cases and, now, we also show the execution time in seconds in order to reach the approximated solution in the last column.

Table 2: Convergence results for Example 4 with $m = 3$.

Method	$w^{(0)}$	λ	iter	$\|w_m^{(k+1)} - w_m^{(k)}\|$	$\|H(w^{(k+1)})\|$	p	Sol
Newton (1)	$(-1, 1, 0.5)$	$-$	100	NC	NC	$-$	$-$
Mod. Newton (2)	$(-1, 1, 0.5)$	-1	9	$1.1912e{-26}$	$3.4423e{-26}$	2.0908	s_3
Mod. Newton (2)	$(-1, 1, 0.5)$	-1.2	11	$5.8898e{-49}$	$5.9379e{-49}$	1.9999	s_1
Newton (1)	$(9, 3, -3)$	$-$	7	$4.0701e{-48}$	$1.3556e{-47}$	1.9999	s_3
Mod. Newton (2)	$(9, 3, -3)$	0.1	6	$4.2756e{-26}$	$1.5069e{-25}$	2.0014	s_3
Mod. Newton (2)	$(9, 3, -3)$	0.01	7	$1.7024e{-48}$	$5.6709e{-48}$	1.9998	s_3
Mod. Newton (2)	$(9, 3, -3)$	1	7	$5.0351e{-46}$	$3.2056e{-45}$	1.9411	s_3

Table 3: Convergence results for Example 4 with $n = 50$.

Method	$w^{(0)}$	λ	iter	$\|w_m^{(k+1)} - w_m^{(k)}\|$	$\|H(w^{(k+1)})\|$	p	Sol	Time
Newton (1)	$w_a^{(0)}$	$-$	100	NC	NC	-	-	54.9641
Mod. Newton (2)	$w_a^{(0)}$	-1	8	$2.3557e{-32}$	$2.4208e{-32}$	2.0	s_1	4.3789
Mod. Newton (2)	$w_a^{(0)}$	-1.2	8	$3.4272e{-26}$	$3.3661e{-26}$	2	s_1	4.3982
Newton (1)	$w_a^{(0)}$	$-$	5	$2.1110e{-35}$	$2.9318e{-35}$	1.9999	s_1	3.4861
Mod. Newton (2)	$w_a^{(0)}$	0.1	5	$2.2719e{-30}$	$3.1553e{-30}$	2.0	s_1	3.1588
Mod. Newton (2)	$w_c^{(0)}$	-0.1	5	$6.6908e{-31}$	$9.2922e{-31}$	1.9999	s_1	2.8910

References

1. A. Cordero, J.R. Torregrosa, *Variants of Newton's method using fifth-order quadrature formulas*, Applied Mathematics and Computation, 190 (2007), 686-698.
2. José L. Hueso, Eulalia Martínez, Juan R. Torregrosa. *Modified Newton's method for systems of nonlinear equations with singular Jacobian*, Journal of Computational and Applied Mathematics, Volume 224, Issue 1, 2009, Pages 77-83.
3. Jisheng Kou, Yitian Li, Xiuhua Wang. *Efficient continuation Newton-like method for solving systems of nonlinear equations*, Applied Mathematics and Computation, Volume 174, Issue 2, 2006, Pages 846-853.
4. Xin-Yuan Wu. *Note on the improvement of Newton's method for system of nonlinear equations*, Applied Mathematics and Computation, Volume 189, Issue 2, 2007, Pages 1476-1479.
5. M.K.A. Ariyaratne, T.G.I. Fernando, S. Weerakoon. *Solving systems of nonlinear equations using a modified firefly algorithm (MODFA),* Swarm and Evolutionary Computation 48 (2019) 72–92.

Analysis of the local and semilocal convergence on Banach spaces of derivative free methods with memory

Eva G. Villalba[1], I. K. Argyros[2], M. A. Hernández-Veró[2], and Eulalia Martínez[1]

[1] Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València. València, Spain.
egarvil@posgrado.upv.es, eumarti@mat.upv.es
[2] Departament of Mathematics Sciences, Cameron University, Lawton, OK 73505, USA.
[3] Department of Mathematicas and Computation, Universidad de La Rioja, Logroño, Spain.
mahernan@unirioja.es

1 Introduction

In mathematical modelling, we often find the problem of solving the system $F(x) = 0$, where F is a nonlinear operator defined on some open set $\Omega \subset X$, $F : \Omega \to Y$ and X, Y are Banach spaces. Iterative methods are used to solve these problems, such as Newton's method, an optimal method of order two.

However, the implementation of this method requires calculating the inverse operator $F'(x_n)^{-1}$ per iteration, then, if the operator F is nondifferentiable, the method is not applicable. Therefore, it is common to replace this operators by a divided differences operator of order one. If this operator is of the form of $[x_n + F(x_n), x_n; F]$, we obtain Steffensen's method [3].

Due to the development of technology, it is necessary to design iterative processes with higher orders of convergence that are also efficient. Therefore, we are interested in studying the following family of derivative free iterative methods, constructed in [4],

$$\begin{cases} x_0 \text{ given in } D, \\ w_n = x_n + \gamma F(x_n), \\ y_n = x_n - [w_n, x_n; F]^{-1} F(x_n), \\ \lambda_n = I - [w_n, x_n; F]^{-1}[y_n, w_n; F], \\ z_n = y_n - P(\lambda_n)[y_n, x_n; F]^{-1} F(y_n), \\ \delta_n = I - [w_n, x_n; F]^{-1}[z_n, y_n; F]\, P(\lambda_n), \\ x_{n+1} = z_n - Q(\lambda_n, \delta_n)[z_n, y_n; F]^{-1} F(z_n), \quad n \geq 0, \end{cases} \tag{1}$$

where $\gamma \in \mathbb{R}$ and $P, Q : \mathcal{L}(E_1) \to \mathcal{L}(E_1)$ are linear weight operators, where $\mathcal{L}(E_1)$ is the domain of linear operators that are bounded, this is $\mathcal{L}(E_1) = \{\xi : E_1 \to E_1 \text{ linear operator} : \xi \text{ is bounded}\}$. This process is of convergence order 7 provided a number of constraints on the operators P and Q [5].

Moreover, taking the real parameter γ as operators of the form $\gamma_n : D \times D \to \mathcal{L}(E_2, E_1)$, the authors obtain the following processes with memory

$$\begin{cases} x_0 \text{ given in } D, , \\ w_n = x_n + \gamma_n F(x_n), \\ y_n = x_n - [w_n, x_n; F]^{-1} F(x_n), \\ \lambda_n = I - [w_n, x_n; F]^{-1}[y_n, w_n; F], \\ z_n = y_n - P(\lambda_n)[y_n, x_n; F]^{-1} F(y_n), \\ \delta_n = I - [w_n, x_n; F]^{-1}[z_n, y_n; F]\, P(\lambda_n), \\ x_{n+1} = z_n - Q(\lambda_n, \delta_n)[z_n, y_n; F]^{-1} F(z_n), \quad n \geq 0, \end{cases} \tag{2}$$

taking

$$\gamma_n = -[x_n, x_{n-1}; F]^{-1}, \tag{3}$$

$$\gamma_n = -[2x_n - x_{n-1}, x_{n-1}; F]^{-1}, \tag{4}$$

$$\gamma_n = -[x_n, y_{n-1}; F]^{-1}, \tag{5}$$

$$\gamma_n = -[2x_n - y_{n-1}, y_{n-1}; F]^{-1}. \tag{6}$$

Then, the convergence orders are $\frac{7+\sqrt{65}}{2}, 8, 4+\sqrt{17}$ and $\frac{9+\sqrt{89}}{2}$, respectively. That is, the convergence speed increases from 7 to 9.21699.

Motivation

The application of these important results presents the following problems, which limit the use of these processes:

(1) Although the processes (1) and (2) do not require the inversion of F', the proof of convergence is carried out by assuming that $F^{(4)}$ at least exists and is bounded.
(2) The results are shown on the space $\mathbb{R}^m$, with $m \in \mathbb{N}$. But, they can apply on equations defined on more general spaces such as Hilbert or Banach spaces.
(3) There are no results on the isolation of the solution x^* in a neighborhood containing it.
(4) It is not known a priori how many iterations must be run to obtain a predetermined accuracy.
(5) Local convergence results in [5] do not provide a criterion for choosing x_0 ensuring the convergence of these methods.

The problems $(1) - (5)$ constitute our motivation for writing this paper. Here is how to respond positively to these problems.

Novelty

(a_1) The convergence conditions depend only and exclusively on the operators which appear on the processes, i.e. the divided difference $[.,.;F]$ and the operator F.
(a_2) The results are valid on Banach spaces.
(a_3) Isolation of the solution results are given.
(a_4) A priori estimates on $||x_n - x^*||$ determine the required number of iterations such that $||x_n - x^*||$ is less than a certain accuracy.
(a_5) New local convergence as well as semilocal convergence results are developed using generalized conditions (ω-continuity).

2 Numerical experiments

Now, we want to apply the iterative family with memory analyzed in the Introduction section in order to approximate the solution of a nonlinear problem defined in generic Banach spaces.

Specifically, we consider the nonlinear integral equation

$$x(s) = g(s) + \lambda p(s) \int_a^b q(t)[\Phi(x)](t)\, dt, \tag{7}$$

we apply the iterative scheme (1) to solve the equation $F(x)(s) = 0$, where

$$[F(x)](s) = x(s) - g(s) - \lambda p(s) \int_a^b q(t)[\Phi(x)](t)dt, \tag{8}$$

with $F : \Omega \subseteq \mathcal{C}([a,b]) \to \mathcal{C}([a,b])$, considering the set $\mathcal{C}([a,b])$ with the max-norm as a Banach space, functions $x(s), g(s), p(s)$ and $q(s)$ are continuous functions defined in the interval $[a,b]$, and $[\Phi(x)](t)$ is a continuous function defined in Ω, known as Nemystkii's operator. In this case, $[\Phi(x)](t) = \phi(x(t))$, where $\phi : \mathbb{R} \to \mathbb{R}$ is a continuous function.

In order to obtain an approximation for a solution of (7) we deal with the approximation of the derivative operator $F'(x(s))$ by divided differences operator, see [1] and [2], where $[x, y; F] \in \mathcal{L}(\mathcal{C}([a,b]), \mathcal{C}([a,b]))$, that must verified, [3]:

$$[x, y; F](x - y) = F(x) - F(y), \tag{9}$$

with $\mathcal{L}(\mathcal{C}([a,b]), \mathcal{C}([a,b]))$ denotes the space of bounded linear operators in the defined Banach space.

We work with nondifferentiable problems, so, we can define for each $x, y \in \mathcal{C}([a,b])$ the following function:

$$\psi[x,y](t) = \begin{cases} \dfrac{\phi(x(t)) - \phi(y(t))}{x(t) - y(t)} & \text{if } t \in [a,b] \text{ with } x(t) \neq y(t), \\ \\ 0 & \text{if } t \in [a,b] \text{ with } x(t) = y(t). \end{cases}$$

Then, we have the following divided difference operator:

$$[x, y; F]u(s) = u(s) - \lambda p(s) \int_a^b q(t)\psi[x, y](t)u(t)\, dt = \omega(s), \tag{10}$$

so we can characterize the inverse operator as follows:

$$u(s) = [x, y; F]^{-1}\omega(s) = \omega(s) + \lambda p(s)\mathbb{E}(x, y, w) \tag{11}$$

where $\mathbb{E}(x, y, w) = \int_a^b q(t)\psi[x_n, y_n](t)u(t)\, dt$.

That is, to obtain $[x, y; F]^{-1}$ explicitly and independently of $u(t)$, we multiply (11) by $q(s)\psi[x, y](s)$, and integrate between a and b. So, we obtain:

$$\mathbb{E}(x, y, w) = \int_a^b q(s)\psi[x, y](s)\omega(s)\, ds + \lambda \int_a^b p(s)q(s)\psi[x, y](s)\, ds\ \ \mathbb{E}(x, y, w).$$

Next, if

$$C(x, y) = \int_a^b p(s)q(s)\psi[x, y](s)\, ds \ \text{ and } \ B(x, y, w) = \int_a^b q(s)\psi[x, y](s)\omega(s)\, ds, \tag{12}$$

then

$$\mathbb{E}(x, y, w) = \frac{B(x, y, w)}{1 - \lambda C(x, y)}.$$

Thus, we can define the action of $[x, y; F]^{-1}$, given by

$$[x, y; F]^{-1}\, \omega(s) = \omega(s) + \lambda p(s)\mathbb{E}(x, y, w).$$

So, the application of the family of iterative schemes (1) is given by the following algorithm.

Fixed $x_{-1}(s), x_0(s) \in \Omega \subseteq \mathcal{C}([a, b])$, for $n \geqslant 0$:

- **First step:** Calculate:

$$F(x_n)(s) = x_n(s) - g(s) - \lambda\, p(s) \int_a^b q(t)\Phi(x_n)(t)\, dt.$$

- **Second step:** Calculate: $\psi[x_{n-1}, x_n](s)$, by (12) we obtain $C(x_{n-1}, x_n)$ and $B(x_{n-1}, x_n, F(x_n))$ and then,

$$w_n(s) = x_n(s) - F(x_n)(s) - \lambda p(s)\mathbb{E}(x_{n-1}, x_n, F(x_n)).$$

- **Third step:** Calculate: $\psi[w_n, x_n](s)$, then obtain $C(w_n, x_n)$ and $B(w_n, x_n, F(x_n))$ and then,

$$y_n(s) = x_n(s) - F(x_n)(s) - \lambda p(s)\mathbb{E}(w_n, x_n, F(x_n)).$$

– **Fourth step:** Notice how the operator μ_n works, let I the identity matrix, then:

$$\begin{aligned}\mu_n(x(s)) &= (I-[w_n,x_n,F]^{-1}[y_n,w_n,F])(x(s))\\ &= x(s)-[w_n,x_n,F]^{-1}(x(s)-\lambda p(s)\int_a^b q(t)\psi[y_n,w_n](t)x(t)\,dt)\\ &= x(s)-(x(s)+\lambda p(s)\mathbb{E}(w_n,x_n,x(s))+[w_n,x_n,F]^{-1}(\widetilde{x}(s))\\ &= -\lambda p(s)\mathbb{E}(w_n,x_n,x(s))+\widetilde{x}(s)+\lambda p(s)\mathbb{E}(w_n,x_n,\widetilde{x}(s)),\end{aligned}$$

where $\widetilde{x}(s)=\lambda p(s)\int_a^b q(t)\psi[y_n,w_n](t)x(t)\,dt$.

– **Fifth step:** Calculate:

$$F(y_n)(s)=y_n(s)-g(s)-\lambda\, p(s)\int_a^b q(t)\Phi(y_n)(t)\,dt$$

and $\psi[y_n,x_n](s)$, then obtain $C(y_n,x_n)$ and $B(y_n,x_n,F(y_n))$ next,

$$z_n(s)=y_n(S)-A(\mu_n)(F(y_n)(s)+\lambda p(s)\mathbb{E}(y_n,x_n,F(y_n))$$

That is, for $A(\mu_n)=\mu^2+\mu_n+I$, if we denote $h_n(s)=(F(y_n)(s)+\lambda p(s)\mathbb{E}(y_n,x_n,F(y_n))$, we have:

$$\mu_n(h_n(s))=-\lambda p(s)\mathbb{E}(w_n,x_n,h_n(s))+\widetilde{h_n}(s)+\lambda p(s)\mathbb{E}(w_n,x_n,\widetilde{h_n}(s)),$$

where $\widetilde{h_n}(s)=\lambda p(s)\int_a^b q(t)\psi[y_n,w_n](t)h_n(t)\,dt$.
So, we obtain:

$$z_n(s)=y_n(s)-\mu_n^2(h_n(s))-\mu_n(h_n(s))-h_n(s).$$

– **Sixth step:** Notice how the operator δ_n works:

$$\begin{aligned}\delta_n(x(s)) &= (I-[w_n,x_n,F]^{-1}[z_n,y_n,F])A(\mu_n(x(s)))\\ &= A(\mu_n(x(s)))-[w_n,x_n,F]^{-1}(A(\mu_n(x(s)))-\lambda p(s)\int_a^b q(t)\psi[z_n,y_n](t)A(\mu_n(x(s)))\,dt)\\ &= -\lambda p(s)\mathbb{E}(w_n,x_n,A(\mu_n(x(s)))+[w_n,x_n,F]^{-1}(\widetilde{A}(\mu_n(x(s)))\\ &= -\lambda p(s)\mathbb{E}(w_n,x_n,A(\mu_n(x(s)))+\widetilde{A}(\mu_n(x(s)))+\lambda p(s)\mathbb{E}(w_n,x_n,\widetilde{A}(\mu_n(x(s))),\end{aligned}$$

where $\widetilde{A}(\mu_n(x(s)))=\lambda p(s)\int_a^b b(t)\psi[z_n,y_n](t)A(\mu_n(x(s)))\,dt$.

– **Seventh step:** Calculate:

$$F(z_n)(s)=z_n(s)-g(s)-\lambda\, p(s)\int_a^b q(t)\Phi(z_n)(t)\,dt$$

and $\psi[z_n,y_n](s)$, then obtain $C(z_n,y_n)$ and $B(z_n,y_n,F(z_n))$ so:

$$x_{n+1}(s)=z_n(s)-D(\mu_n,\delta_n)(F(z_n)(s)+\lambda p(s)\mathbb{E}(z_n,y_n,F(z_n)),$$

with $D(\mu_n,\delta_n)=I+\mu_n\circ\delta_n+13/6\mu_n\circ\delta_n^2$, so let be $f_n(s)=F(z_n)(s)+\lambda p(s)\mathbb{E}(z_n,y_n,F(z_n))$ then we have:

$$x_{n+1}(s)=z_n(s)-f_n(s)-\mu_n(\delta_n(f_n(s)))-13/6\mu_n(\delta_n^2(f_n(s))).$$

Remark: Notice that the algorithm has been described for case (3), but for obtaining the iterates in different cases, given by (4) − (6), we have to change in second step the parameters of function ψ, so we have:

I. Method	Equation	γ_n	Second step:
ALG_1	(3)	$-[x_n, x_{n-1}; F]^{-1}$	$\psi(x_n, x_{n-1})$
ALG_2	(4)	$-[2x_n - x_{n-1}, x_{n-1}; F]^{-1}$	$\psi(2x_n - x_{n-1}, x_{n-1})$
ALG_3	(5)	$-[x_n, y_{n-1}; F]^{-1}$	$\psi(x_n, y_{n-1})$
ALG_4	(6)	$-[2x_n - y_{n-1}, y_{n-1}; F]^{-1}$	$\psi(2x_n - y_{n-1}, y_{n-1})$

Table 1: Different algorithms with memory depending on the choice of γ_n.

Particular example

Now, in (7) we take $g(s) = (1 - 11/80\lambda)s - 1/2$, $\lambda = 1$, $p(s) = s$, $q(t) = t$ and $\Phi(x(t)) = x^3(t) + |x(t)|$, so we have the nonlinear integral equation:

$$[F(x)](s) = x(s) - ((1 - 11/80)s - 1/2) - s\int_a^b t(x^3(t) + |x(t)|)dt, \qquad (13)$$

which exact solution is $x^*(s) = s - 1/2$.

Then, by taking as starting estimates the functions $x_0(s) = s$ and $x_{-1}(s) = 1/3$, with $s \in [0, 1]$, we apply the iterative methods given in Table 1, following the steps one through seven described above, where all integrals have been approximated by Simpson's quadrature with 200 nodes. We work with MATLAB R 2019a with 5000 digits, by imposing the stopping criteria $\|x_{n+1}(s) - x_n(s)\| \leq 10^{-300}$. Notice that we have to work with variable precision arithmetic for running high order methods in order to reach the approximated computational order of convergence, see [**?**], that is shown in the numerical results of Table 2 by p. We observe in second column of Table 2 that $ALG2$ and $ALG4$ methods needs one less iteration for reaching the required tolerance, also we show in the Table the distance between the last two iterations and the value of the nonlinear operator F at the approximated solution.

I. Method	*iter*	$\|x_{n+1}(s) - x_n(s)\|$	$\|F(x_{n+1}(s))\|$	p
ALG_1	5	6.42499e-1950	1.29133e-5007	7.46643
ALG_2	4	7.56024e-341	1.02825e-2729	8.18241
ALG_3	5	6.78516e-896	4.11821e-5008	6.67125
ALG_4	4	2.40361e-316	3.06247e-2363	7.51117

Table 2: Numerical results for different values of γ_n.

References

1. S. Amat, S. Busquier, A. Grau and M. Grau-Sánchez, Maximum efficiency for a family of Newton-like methods with frozen derivatives and some applications. Applied Mathematics and Computation, 219 (2013), 7954–7963.
2. I. K. Argyros and S. George. Improved convergence analysis for the Kurchatov method. Nonlinear Functional Analysis and Applications, 22,1 (2017), 41–58
3. M. Balazs, and G. Goldner, On existence of divided differences in linear spaces. Rev. Anal. Numer. Theor. Approx., 2 (1973), 3–6.
4. Alicia Cordero, Neus Garrido, Juan Ramon Torregrosa, Paula Triguero-Navarro. *Design of iterative methods with memory for solving nonlinear systems.* Authorea. September 05, 2022.
5. A. Cordero, N. Garrido, J. R. Torregrosa and P. Triguero-Navarro, Design of iterative methods with memory for solving nonlinear systems. Math. Meth. Appl. Sci. (2023), 1–17. DOI 10.1002/mma.9182.

Efficient multidimensional family of iterative methods free of Jacobian matrices

Francisco I. Chicharro[1], Alicia Cordero[1], Neus Garrido[1], and Juan R. Torregrosa[1]

Instituto de Matemática Multidisciplinar, Universitat Politècnica de València
Camino de Vera, s/n, 46022 Valencia, Spain
neugarsa@mat.upv.es

Abstract. In this work, a numerical comparison of two families of iterative methods for solving nonlinear systems of equations is carried out. Specifically, from an initial three-step family of iterative schemes with a parameter and order of convergence 4, a biparametric family is designed approximating the Jacobian matrix by a divided difference operator. The new iterative family holds the order of convergence 4, but its iterative scheme is free of Jacobian matrices. To compare the efficiency of the two iterative classes, some of their members are selected by setting the parameters based on a previous multidimensional dynamical study, obtaining very efficient results for solving polynomial systems for the Jacobian-free methods.

Keywords: Nonlinear systems, Jacobian-free, Iterative family, Numerical analysis

1 Introduction

In many scientific applications, solving systems of nonlinear equations has become a common and recurring challenge. The complexity of analytically calculating precise solutions for such problems has led to the use of iterative fixed-point algorithms to approximate the solutions of these systems, allowing for practical use in real-world scenarios. In this sense, the problem of finding a real solution x^* of $F(x) = 0$, where $F : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a nonlinear multidimensional function, can be obtained as the fixed point of some function $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ by means of an iterative procedure of the form

$$x^{(k+1)} = G(x^{(k)}), \qquad k = 0, 1, 2, \ldots$$

being $x^{(0)}$ the initial estimation.

Over the last few decades, several multidimensional iterative methods have been proposed to enhance the speed of convergence and the computational efficiency [1–4]. In [5] the authors present the following family of iterative methods

with a real parameter $\beta \neq 0$:

$$y^{(k)} = x^{(k)} - \left[F'(x^{(k)})\right]^{-1} F(x^{(k)}),$$

$$z^{(k)} = y^{(k)} - \beta \left[F'(x^{(k)})\right]^{-1} F(y^{(k)}), \qquad k = 0, 1, 2 \ldots$$

$$x^{(k+1)} = z^{(k)} - \frac{1}{\beta} \left[F'(x^{(k)})\right]^{-1} \left(-(\beta-1)^2 F(y^{(k)}) + F(z^{(k)})\right).$$

It is shown in [5] that the previous class, called M4 family, has order of convergence 4 for any value of the parameter. In particular, when $\beta = 5$, then the order of the corresponding method is 5. The parametric class M4 is designed using Newton's composition with itself as the main technique, but with frozen Jacobian in order to reduce the number of different functional evaluations and then the computational and operational cost.

Since it is common that the nonlinear function to be solved does not have a known Jacobian matrix or it is a non differentiable multidimensional operator, in this work we propose an iterative family that does not require Jacobian matrices. Section 2 is devoted to the design of the new family of iterative schemes after approximating the Jacobian matrices present in M4. In Section 3 a numerical analysis of the two families of methods applied to different nonlinear functions is performed. Finally, the conclusions of this work are shown in Section 4.

2 Biparametric family of iterative methods

First of all, we consider the following approximation for the Jacobian matrix:

$$F'(x^{(k)}) \approx \left[x^{(k)}, w^{(k)}; F\right], \tag{1}$$

where $w^{(k)} = x^{(k)} + \gamma F(x^{(k)})$, $\gamma \in \mathbb{R}$, and $[\cdot, \cdot; F] : D \subseteq \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathcal{L}(\mathbb{R}^n)$ denotes the divided difference operator satisfying

$$[x, y; F](x - y) = F(x) - F(y), \quad x, y \in D.$$

When we replace the approximation (1) in family M4, the following Jacobian-free biparametric family, denoted by JF4, is obtained:

$$y^{(k)} = x^{(k)} - \left[x^{(k)}, w^{(k)}; F\right]^{-1} F(x^{(k)}),$$

$$z^{(k)} = y^{(k)} - \beta \left[x^{(k)}, w^{(k)}; F\right]^{-1} F(y^{(k)}), \qquad k = 0, 1, 2 \ldots$$

$$x^{(k+1)} = z^{(k)} - \frac{1}{\beta} \left[x^{(k)}, w^{(k)}; F\right]^{-1} \left(-(\beta-1)^2 F(y^{(k)}) + F(z^{(k)})\right),$$

where $w^{(k)} = x^{(k)} + \gamma F(x^{(k)})$, $\gamma \in \mathbb{R}$. It can be shown by using Taylor series developments that the family of iterative schemes JF4 has order of convergence 4 for any value of parameters γ and β.

In the next section, we are going to analyse numerically the performance of some members of the two iterative families to check their functionality in solving systems of nonlinear equations.

3 Numerical experiments

In [8] the stability of family M4 applied to quadratic polynomial systems is studied using fundamentals from real multidimensional dynamics. This study allows to determine the appropriate initial estimates to start the iterative process and converge to the solution of the problem. Furthermore, in a family of iterative methods with a parameter, this study allows to determine the most stable members in terms of stability. After the analysis performed in [8], we have selected parameters $\beta = -10$, $\beta = 1$ and $\beta = 5$, as these values provide iterative methods of family M4 with wide basins of attraction, visualized in their respective dynamical planes.

Now, we are going to use the selected iterative schemes to solve different nonlinear systems and also to compare the results with those obtained for the iterative family JF4 free of Jacobian matrices. As JF4 is a biparametric family, we have fixed $\gamma = 0.1$ in order to generate a sequence of points $\{w^{(k)}\}_{k\geq 0}$, where $w^{(k)} = x^{(k)} + \gamma F(x^{(k)})$, close to $\{x^{(k)}\}_{k\geq 0}$.

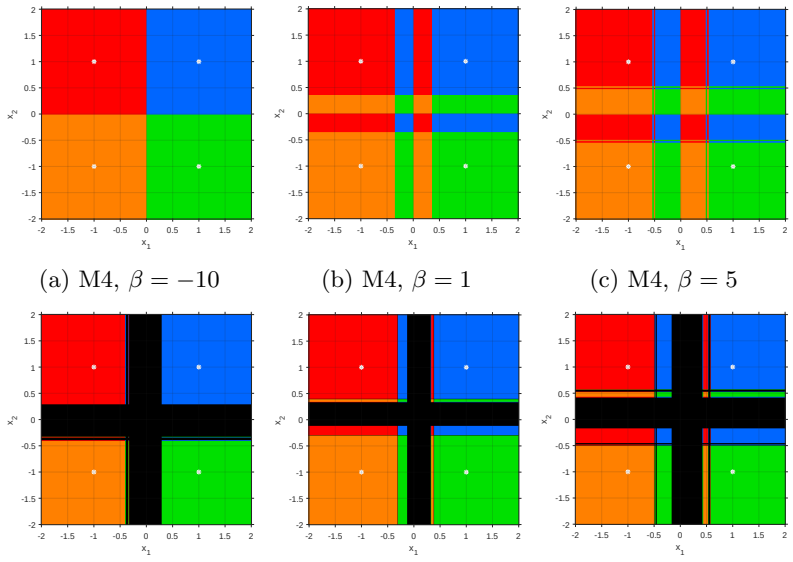
The nonlinear functions selected for the numerical tests are as follows:

- $F_1(x_1, x_2) = \begin{cases} x_1^2 - 1 = 0 \\ x_2^2 - 1 = 0 \end{cases}$, with $x^*_{1-4} = (\pm 1, \pm 1)$,
- $F_2(x_1, x_2) = \begin{cases} x_1^2 - x_2 - 19 = 0 \\ \frac{x_2^3}{6} - x_1^2 + x_2 - 17 = 0 \end{cases}$, with $x^*_1 = (-5, 6)$ and $x^*_2 = (5, 6)$,
- $F_3(x_1, x_2) = \begin{cases} x_1^2 + x_2^2 - 1 = 0 \\ x_1^2 - x_2^2 + 1/2 = 0 \end{cases}$, with $x^*_{1-4} = (\pm \frac{1}{2}, \pm \frac{\sqrt{3}}{2})$,

where x^* denotes the exact solutions in $\mathbb{R}^2$. We have select bidimensional nonlinear functions in order to represent the basins of attraction of each method in the corresponding dynamical planes. According to the implementation in [6, 7], the dynamical planes represent the set of initial estimations of a method that converge to the solution of the problem. As all the iterative schemes of M4 and JF4 require an initial point $x^{(0)}$, the dynamical planes will sow us the most convenient initial estimates to obtain approximations that converge to the solution of each problem.

Figures 1-3 represent the dynamical planes of F_1, F_2 and F_3, respectively, when the nonlinear functions are solved by using the methods of M4 and JF4 corresponding to $\beta = \{-10, 1, 5\}$ and $\gamma = 0.8$. Each point in the planes, taking

a mesh of 500×500 points (x_1, x_2), is considered as initial estimation of the method. The convergence is established when the difference between the point of the orbit of each initial estimate and any of the roots of the nonlinear function is less than 10^{-3} with a maximum of 50 iterations. When there is convergence, the point is represented in a colour different from black, which denotes divergence. In addition, the roots are represented with white stars.



(a) M4, $\beta = -10$ (b) M4, $\beta = 1$ (c) M4, $\beta = 5$

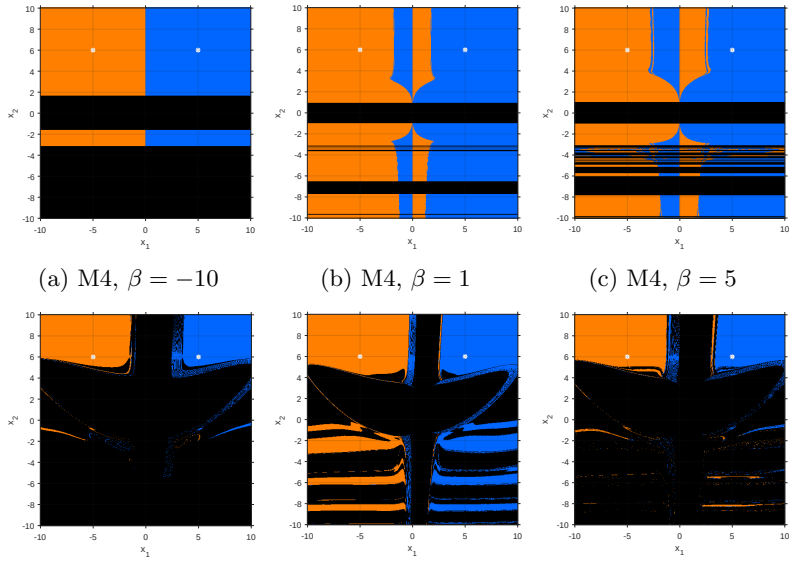
(d) JF4, $\beta = -10$ (e) JF4, $\beta = 1$ (f) JF4, $\beta = 5$

Fig. 1: Dynamical planes for solving $F_1(x) = 0$

The dynamical planes allow us to select the best initial estimations for each method in order to converge to the solution of the problem. We can observe in Figures 1-3 that methods from M4 family have wide basins of attraction than methods from JF4 family. Moreover, the schemes corresponding to $\beta = 1$ have the largest number of initial estimates that converge to some root.

To compare these results with the numerical performace, the methods of each family have been implemented using Matlab R2022b and variable precision arithmetics with 200 digits. Methods corresponding to $\beta = \{-10, 1, 5\}$ and $\gamma = 0.8$ have been used for solving the selected nonlinear functions and taking different initial estimations from the dynamical planes (from the colored regions). The convergence is set when $||x^{(k+1)} - x^{(k)}|| < 10^{-50}$ or $||F(x^{(k+1)})|| < 10^{-50}$, under 50 iterations. The results obtained are summarized in Table 1.

Table 1 shows the number of iterations required for each test, the difference between the two last iterations and the norm of the function in the last iterate. It can be observed that, although the basins of attraction of the methods of



(a) M4, $\beta = -10$ (b) M4, $\beta = 1$ (c) M4, $\beta = 5$

(d) JF4, $\beta = -10$ (e) JF4, $\beta = 1$ (f) JF4, $\beta = 5$

Fig. 2: Dynamical planes for solving $F_2(x) = 0$

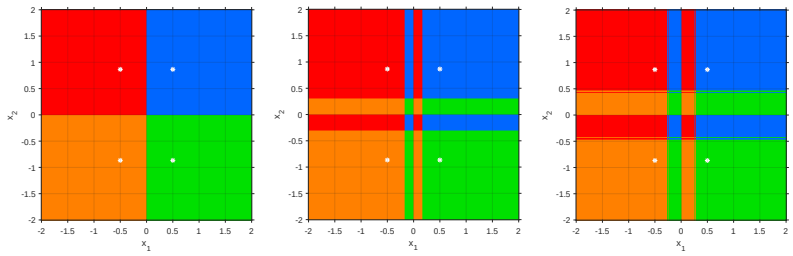
M4 family were larger than those of the Jacobian-free class, in the numerical implementation the Jacobian-free family is much more efficient for solving the considered nonlinear problems as it achieves convergence to solutions with few iterations and with high accuracy.

4 Conclusions

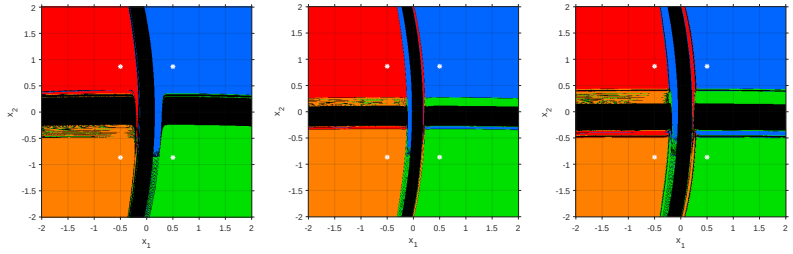
In this work, a comparison of the efficiency and stability of two families of iterative methods has been carried out. Starting from a family of iterative schemes with order of convergence 4, a biparametric family free of Jacobian matrices and with order 4 has been designed using a divided difference operator. After analyzing on different polynomial systems the basins of attraction of some of the members of the two families, the initial estimates have been selected to execute numerically these methods. The results show that the M4 family contains stable iterative schemes. However, the numerical implementation shows more efficient results for the JF4 family, achieving good approximations to the solutions of the proposed problems.

References

1. Kurchatov, V. A.: On a method of linear interpolation for the solution of functional equations, Dokl. Acad. Nauk. SSSR (Russian) vol. 198-3, pp. 524–526 (1971). Translation in Soviet Math. Dokl. vol. 12, pp. 835–838 (1971).



(a) M4, $\beta = -10$ (b) M4, $\beta = 1$ (c) M4, $\beta = 5$



(d) JF4, $\beta = -10$ (e) JF4, $\beta = 1$ (f) JF4, $\beta = 5$

Fig. 3: Dynamical planes for solving $F_3(x) = 0$

2. Grau-Sánchez, M., Grau, A., Noguera, M.: Ostrowski type methods for solving systems of nonlinear equations. Appl. Math. Comput. vol. 218, pp. 2377–2385 (2011).
3. Wang, W., Zhang, T., Qian, W., Teng, M.: Seventh-order derivative-free iterative method for solving nonlinear systems. Numer. Alg. vol. 70, pp. 545–558 (2015).
4. Chicharro, F.I., Cordero, A., Garrido, N., Torregrosa, J.R.: On the improvement of the order of convergence of iterative methods for solving nonlinear systems by means of memory. Appl. Math. Lett. vol. 104-106277 (2020).
5. Cordero, A., Gutiérrez, J.M., Magreñán, Á.A., Torregrosa, J.R.: Stability analysis of a parametric family of iterative methods for solving nonlinear models, Appl. Math. Comput. vol. 285, pp. 26—40 (2016).
6. Cordero, A., Soleymani, F., Torregrosa, J.R.: Dynamical analysis of iterative methods for nonlinear systems or how to deal with the dimension?. Appl. Math. Comput. vol. 244, pp. 398–412 (2014).
7. Chicharro, F.I., Cordero, A., Torregrosa, J.R.: Drawing dynamical and parameters planes of iterative families and methods. The Scientific World Journal. vol. 2013, Article ID 780153, pp. 1–11 (2013).
8. Chicharro, F.I., Cordero, A., Torregrosa, J.R.: Real stability of an efficient family of iterative methods for solving nonlinear systems. Proceedings of the XVI Congreso de Ecuaciones Diferenciales Y Aplicaciones, XV Concreso de Matemática Aplicada. pp 206–212. Cartagena (2017).

Function	β	$x^{(0)}$	**Method**	**iter**	$\|\|x^{(k+1)} - x^{(k)}\|\|$	$\|\|F(x^{(k+1)})\|\|$
F_1	-10	$\begin{pmatrix} -1.2 \\ 1.4 \end{pmatrix}$	M4	31	5.3575e–50	2.5588e–51
			JF4	4	9.8862e–18	2.6270e–68
	1	$\begin{pmatrix} -0.8 \\ -0.4 \end{pmatrix}$	M4	–	–	–
			JF4	4	2.3790e–49	0
	5	$\begin{pmatrix} 0.8 \\ 1.5 \end{pmatrix}$	M4	35	2.1065e–50	1.6204e–51
			JF4	4	3.7636e–25	2.0064e–98
F_2	-10	$\begin{pmatrix} -4 \\ 7 \end{pmatrix}$	M4	30	9.7407e–51	3.4069e–51
			JF4	7	2.3187e–48	2.5853e–191
	1	$\begin{pmatrix} 5.5 \\ -2.5 \end{pmatrix}$	M4	–	–	–
			JF4	13	6.4559e–32	0
	5	$\begin{pmatrix} 6 \\ 7 \end{pmatrix}$	M4	30	1.4816e–51	4.6232e–52
			JF4	5	3.8100e–17	6.8618e–67
F_3	-10	$\begin{pmatrix} 1 \\ 1.5 \end{pmatrix}$	M4	50	2.2010e–45	4.4467e–46
			JF4	5	1.0556e–25	1.9316e–99
	1	$\begin{pmatrix} -1.5 \\ -0.5 \end{pmatrix}$	M4	–	–	–
			JF4	4	1.5551e–17	4.8362e–135
	5	$\begin{pmatrix} -0.5 \\ -1.2 \end{pmatrix}$	M4	30	2.79339e–50	1.5078e–51
			JF4	4	7.8856e–33	7.2914e–129

Table 1: Numerical results for solving nonlinear problems

Impact of complex and real dynamical analysis on the performance of a new iterative family

Marlon Moscoso-Martínez[1,2], Alicia Cordero[1], Juan R. Torregrosa[1], and F. I. Chicharro[1]

[1] Institute for Multidisciplinary Mathematics, Universitat Politècnica de València, Camino de Vera s/n, 46022 València, Spain,
marmosma@doctor.upv.es,
WWW home page: https://damres.webs.upv.es/?page_id=7
[2] Faculty of Sciences, Escuela Superior Politécnica de Chimborazo (ESPOCH), Panamericana Sur km 1 1/2, 060106 Riobamba, Ecuador

Abstract. In this paper, we introduce a new three-step sixth-order uniparametric iterative family for solving nonlinear equations and its extension for systems. A complex dynamical analysis is carried out on scalar cases and a real dynamical study on vector cases. The purpose of this manuscript is to show the impact of these studies on the performance of the family, considering parameter spaces and dynamical planes as tools to determine the best and worst iterative schemes in terms of stability. Several numerical tests are performed with selected members to illustrate the errors and the number of iterations to converge to the solution.

Keywords: non-linear equations and systems; multistep iterative methods; convergence analysis; complex and real dynamics; chaos and stability

1 Introduction

The problem of solving non-linear equations and systems without algebraic solutions is a common challenge in scientific and engineering domains, especially when dealing with real-world physical phenomena. Many physical problems involve complex relationships that cannot be solved analytically, necessitating the use of numerical methods for their solution.

Numerical methods have made significant progress and are vital in scientific and engineering applications where analytical solutions are challenging or impossible to obtain. State-of-the-art numerical methods are typically iterative in nature and leverage dynamical analysis techniques to improve their convergence and stability.

Dynamical analysis involves studying the behavior of iterative schemes applied to dynamical systems, with a focus on understanding their accuracy, stability, and convergence properties. The impact of complex and real dynamical studies on the performance of numerical methods has been a subject of research in fields such as computational mathematics, physics, and engineering. The main

goal is to identify the limitations and potential errors of iterative schemes when applied to non-linear equations and systems.

To analyze the stability and reliability of numerical methods, researchers have found the dynamical behavior of the rational operator associated with these schemes applied to low-degree nonlinear polynomial equations or systems to be an effective tool. For example, refer to [1,3] and the references they contain, as they provide valuable insights into this aspect.

Thus, in this manuscript, we present a novel three-step sixth-order uniparametric iterative family designed for solving nonlinear equations, along with its extension for systems of equations. We conduct complex dynamical analysis on scalar cases and real dynamical analysis on vector cases. The primary objective of this study is to demonstrate how these analyses impact the performance of the iterative family, using parameter spaces and dynamical planes as tools to determine the most stable and reliable iterative schemes. We develop several numerical tests using selected members of the family to illustrate the errors and the number of iterations required for convergence to the solution.

The rest of the article is organized as follows: Section 2 presents the novel three-step uniparametric family of iterative methods, to solve equations and systems, and its order of convergence; Section 3 provides the complex and real dynamical analyses on scalar and vector cases, respectively; Section 4 presents the numerical tests for selected members of the family and the results obtained; and finally, Section 5 provides some relevant conclusions of this research.

2 Novel family of iterative schemes

The novel uniparametric family for non-linear equations, object of study in this manuscript and which we will call mctc(α), has the following iterative expression:

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f'(x_k)} \\ z_k = y_k - \dfrac{f(y_k)}{2f[x_k, y_k] - f'(x_k)} \\ x_{k+1} = z_k - (\alpha + (1+\alpha)u_k + (1-\alpha)v_k)\dfrac{f(z_k)}{f'(x_k)} \end{cases}, \tag{1}$$

where $u_k = 1 - \dfrac{f[x_k, y_k]}{f'(x_k)}$, $v_k = \dfrac{f'(x_k)}{f[x_k, y_k]}$, $k = 0, 1, 2, ...$, and α is an arbitrary parameter. The divided difference operator $f[\cdot, \cdot] : I \times I \subset \mathbb{R} \times \mathbb{R} \to \mathcal{L}(\mathbb{R})$, defined in [6], satisfies $f[x, y](x - y) = f(x) - f(y), \forall x, y \in I$.

Theorem 1 (scalar cases). *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a sufficiently differentiable function on an open interval I and $\xi \in I$ a simple root of the nonlinear equation $f(x) = 0$. Suppose that $f(x)$ is continuous and sufficiently differentiable in an environment of the simple root ξ, and x_0 is an initial estimate close enough to ξ. Then, the sequence $\{x_k\}_{k \geq 0}$ obtained by using the expression* (1) *converges to ξ with an order of convergence of six, being its error equation*

$$e_{k+1} = \left(6C_2^5 - 7C_2^3C_3 + C_2C_3^2\right) e_k^6 + \mathcal{O}\left(e_k^7\right),$$

where $e_k = x_k - \xi$, $C_q = \dfrac{1}{q!}\dfrac{f^{(q)}(\xi)}{f'(\xi)}$ *and* $q = 2, 3, ...$

Based on Theorem 1, whose proof can be found in [4], it is evident that the newly proposed uniparametric family of iterative methods for scalar cases achieves a remarkable order of convergence of six, regardless of the parameter α. This raises the intriguing question of whether this family can be extended to handle vector cases and, if so, how its order of convergence will be affected. To address this intriguing inquiry, we introduce the novel family tailored explicitly for systems and which we will call MCTC(α), as illustrated below:

$$\begin{cases} y^{(k)} = x^{(k)} - [F'(x^{(k)})]^{-1}F(x^{(k)}), \\ z^{(k)} = y^{(k)} - [2[x^{(k)}, y^{(k)}; F] - F'(x^{(k)})]^{-1}F(y^{(k)}), \\ x^{(k+1)} = z^{(k)} - (\alpha I + (1+\alpha)u^{(k)} + (1-\alpha)v^{(k)})[F'(x^{(k)})]^{-1}F(z^{(k)}), \end{cases} \tag{2}$$

where $u^{(k)} = I - [F'(x^{(k)})]^{-1}[x^{(k)}, y^{(k)}; F]$, $v^{(k)} = [x^{(k)}, y^{(k)}; F]^{-1}F'(x^{(k)})$, $k = 0, 1, 2, ...$, and α is an arbitrary parameter. The divided difference operator $[x, y; F]$ is the map $[\cdot, \cdot; F] : D \times D \subset \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^n)$, satisfying $[x, y; F](x - y) = F(x) - F(y), \forall x, y \in D$.

Theorem 2 (vector cases). *Let* $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ *be a sufficiently differentiable function in an open convex set* D *and* $\xi \in D$ *a solution of the non-linear system* $F(x) = 0$. *Let us suppose that* $F'(x)$ *is continuous and nonsingular at* ξ *and* $x^{(0)}$ *is an initial estimate close enough to* ξ. *Then, sequence* $\{x^{(k)}\}_{k\geq 0}$ *obtained by using expression* (2) *converges to* ξ *with order six, being its error equation*

$$e^{(k+1)} = \left(C_3^2C_2 - C_3C_2^3 + 6C_2^5 - 6C_2^2C_3C_2\right) {e^{(k)}}^6 + \mathcal{O}({e^{(k)}}^7),$$

where $e^{(k)} = x^{(k)} - \xi$, $C_q = \frac{1}{q!}[F'(\xi)]^{-1}F^{(q)}(\xi)$ *and* $q = 2, 3, ...$

According to Theorem 2, whose proof can be found in [5], it can be deduced that the uniparametric family for vector cases also exhibits an impressive order of convergence of six, for any value of α.

3 Complex and real dynamics for stability

The performance of the novel family is significantly influenced by complex and real dynamical analysis. This is evident as parameter spaces, parameter lines, and dynamical planes provide crucial insights into the stability of individual

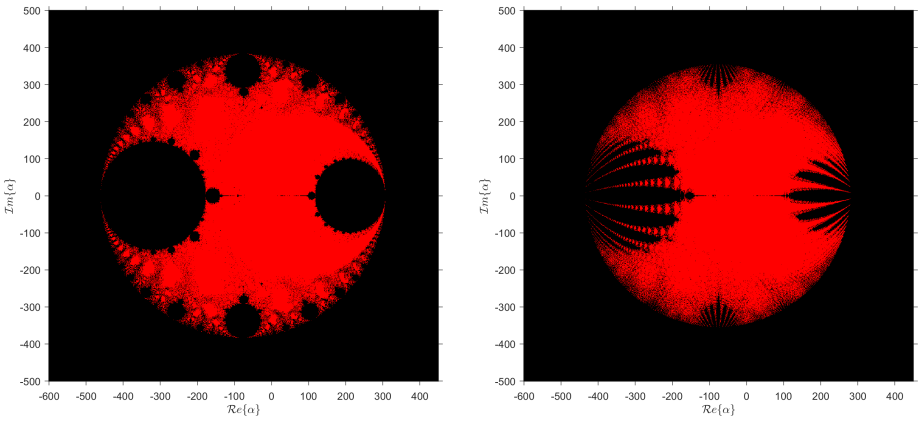
members within the family. For this reason, we analyze the dynamical behavior of the rational operators associated to iterative schemes and applied to low-degree non-linear polynomial equations (see Equation (3)) or systems (see Equation (4)), since the criterion of stability or instability of a method applied to these problems can be generalized to other unidimensional or multidimensional cases, respectively.

$$f(x) = (x-a)(x-b) = 0, \quad a, b \in \mathbb{R}. \tag{3}$$

$$F(x_1, x_2) = \left(x_1^2 - 1, x_2^2 - 1\right) = (0,0)\,. \tag{4}$$

By exploring the parameter spaces or parameter lines derived from critical points, we gain a comprehensive understanding of the performance of various members within the family, aiding us in selecting a suitable method. In the complex analysis of mctc(α) family, we can identify a maximum of nine free critical points. Of these, we obtained two different parameter spaces, P_1 and P_2, shown in Figure 1 with the following features:

- A mesh was built, from -500 to 500 in $Im\{\alpha\}$ and from -600 to 450 in $Re\{\alpha\}$, with a step equal to 0.01 (100000×105000 points).
- The point is red if the method linked to a value of α converges to any of the roots of the polynomial equation, and black if it diverges once all iterations are completed.
- The maximum number of iterations is 100 while the stopping criteria is a tolerance for the error estimation equal to 10^{-3}.

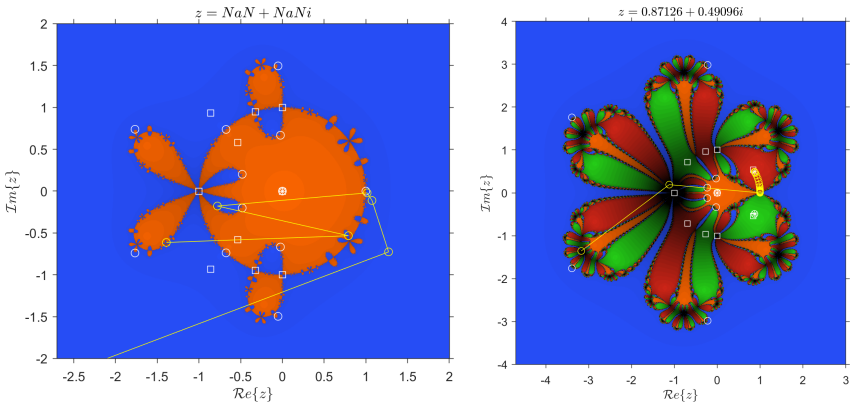


(a) P_1 (b) P_2

Fig. 1: Parameter spaces of free critical points for scalar cases

On one hand, when selecting a value of α within the stability regions (red regions) of the parameter spaces P_1 and P_2, such as $\alpha = 0$, the method linked to this parameter will exhibit favorable dynamical behavior concerning numerical stability. On the other hand, if a value of α is chosen outside the stability regions (black regions) of the same parameter spaces, such as $\alpha = 200$, the method associated with this parameter will exhibit poor dynamical behavior in terms of numerical stability. To corroborate this, we construct dynamical planes for methods associated with the mentioned values of α, as shown in Figure 2, with the following features:

- A mesh of 1000×1000 points was built.
- Every initial estimation is iterated 100 times (maximum) with stopping criteria as a tolerance equal to 10^{-3}.
- The points in the mesh are represented depending on the roots to which they converge: color is brighter when lesser are the iterations.
- If all the iterations are completed and not convergence to any roots is reached, then the point is represented in black.



(a) mctc(0) (b) mctc(200)

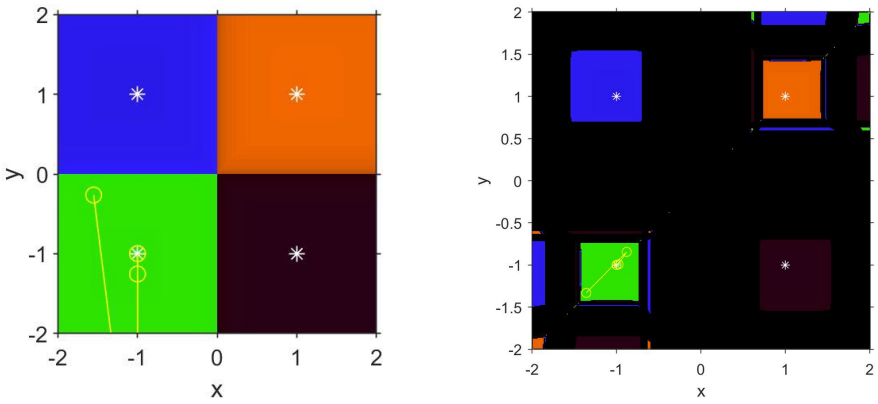
Fig. 2: Dynamical planes for scalar methods

As you can see in Figure 2, dynamical planes illustrate how specific methods behave in relation to the basins of attraction for their fixed points, periodic points, and other relevant characteristics. Examining the basins of attraction allows us to visually interpret the functioning of a method based on multiple initial estimates.

On one hand, we provide an example of a method within the stability region, specifically for $\alpha = 0$. Its dynamical plane, containing a convergence orbit highlighted in yellow, is depicted in Figure 2a. The method exhibits only two

basins of attraction associated with the roots, which are colored in orange and blue. Additionally, there are no black areas representing non-convergence or slow convergence to the solution. As a result, the method demonstrates favorable dynamical behavior and high stability. On the other hand, an example of a method outside the stability region is presented for $\alpha = 200$. Its corresponding dynamical plane, with a convergence orbit illustrated in yellow, is displayed in Figure 2b. Notably, the method exhibits more than two basins of attraction, indicating the presence of other attractors not related to the roots. The basins of the roots are colored in orange and blue, while the other basins are represented by red and green regions. Figure 2b demonstrates convergence to an attracting strange fixed point. Consequently, this method exhibits poor dynamical behavior and lack stability.

This same analysis is extended to vector cases. Now, it is necessary to work with the family of iterative methods designed to solve systems of nonlinear equations, MCTC(α). However, due to the difficulty that may arise in constructing parameter lines in the analysis of real dynamics, we will build on the results obtained from complex dynamics, attempting to extrapolate the stability and instability criteria from scalar cases to vector cases. Thus, we construct dynamical planes for the same values of α associated with stable and unstable methods in scalar cases, but this time for multidimensional cases, as shown in Figure 3.



(a) MCTC(0) (b) MCTC(200)

Fig. 3: Dynamical planes for vector methods

As observed in Figure 2 and Figure 3, the stability and instability results of the methods mctc(0) and mctc(200), respectively, extend to the methods MCTC(0) and MCTC(200). This finding is of great significance as it allows us to analyze the dynamics of iterative families designed for multidimensional cases, considering the analysis in unidimensional cases.

4 Numerical results

In this section, we conduct various numerical tests to validate the theoretical convergence and stability outcomes of the MCTC(α) family displayed in Equation (2). For this purpose, we employ both stable MCTC(0) and unstable MCTC(200) methods. These methods are then applied to five nonlinear test systems, each characterized by specific expressions and corresponding roots, shown in Table 1.

Table 1: Non-linear test systems and corresponding roots.

Non-linear test system	Roots
$F(x_1, x_2) = \left(x_1^2 - 1, x_2^2 - 1\right) = (0, 0)$	$\xi \approx (1, 1)^T$
$G(x_1, x_2) = \left(x_1^2 + x_2^2 - 1, x_1^2 - x_2^2 - \frac{1}{2}\right) = (0, 0)$	$\xi \approx \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)^T$
$M(x_1, x_2) = (e^{x_1} e^{x_2} + x_1 \cos(x_2), x_1 + x_2 - 1) = (0, 0)$	$\xi \approx (3.4706, -2.4706)^T$
$N(x_1, x_2) = \left(\ln(x_1^2) - 2\ln(\cos(x_2)), x_1 \tan(x_2)\right) = (0, 0)$	$\xi \approx (1, 0)^T$
$O(x_1, x_2) = (x_1 + e^{x_2} - \cos(x_2) + 0.5, 3x_1 - x_2 - \sin(x_2)) = (0, 0)$	$\xi \approx (-0.2535, -0.3851)^T$

The computations were performed using the Matlab R2020b programming package with variable precision arithmetic featuring 200 digits of mantissa. For each method, we analyze the number of iterations (iter) required to achieve convergence to the solution, subject to the stopping criteria $||x^{(k+1)} - x^{(k)}|| < 10^{-100}$ or $||F(x^{(k+1)})|| < 10^{-100}$. Here, $||x^{(k+1)} - x^{(k)}||$ represents the error estimation between two consecutive iterations, and $||F(x^{(k+1)})||$ is the residual error of the nonlinear test system.

To assess the theoretical order of convergence of the methods, we calculate the approximate computational order of convergence (ACOC) as defined in [7]. In the numerical results, if the ACOC values do not stabilize throughout the iterative process, they are denoted as '-'. Furthermore, if any of the methods employed fails to reach convergence within a maximum of 50 iterations, it is marked as 'nc'.

Thus, in Table 2 we show the numerical performance of MCTC(0) for initial estimates very close to the solution ($x^{(0)} \approx \xi$). We notice that MCTC(0) always converges to the solution. The theoretical convergence order is also verified by the ACOC, which is close to six. But, what about the dependence of MCTC(0) on initial estimations? To answer this question, we analyze this method for initial

estimates near to and far from the solution, that is, for $x^{(0)} \approx 3\xi$ and $x^{(0)} > 10\xi$, respectively. The results can be observed in Tables 3 and 4.

Table 2: Numerical performance of MCTC(0) on test problems for $x^{(0)} \approx \xi$.

System	$x^{(0)}$	$\|\|x^{(k+1)} - x^{(k)}\|\|$	$\|\|F(x^{(k+1)})\|\|$	iter	ACOC
$F(x_1, x_2)$	$(0.90, 0.90)^T$	4.1590e-41	1.0578e-162	3	6.0326
$G(x_1, x_2)$	$(0.80, 0.40)^T$	1.6140e-29	3.8389e-115	3	5.9785
$M(x_1, x_2)$	$(3.40, -2.40)^T$	1.1444e-49	1.3224e-131	3	5.5845
$N(x_1, x_2)$	$(0.90, 0.10)^T$	1.0160e-76	5.8602e-308	4	-
$O(x_1, x_2)$	$(-0.20, -0.30)^T$	1.2709e-37	6.3818e-107	3	5.9417

Table 3: Numerical performance of MCTC(0) on test problems for $x^{(0)} \approx 3\xi$.

System	$x^{(0)}$	$\|\|x^{(k+1)} - x^{(k)}\|\|$	$\|\|F(x^{(k+1)})\|\|$	iter	ACOC
$F(x_1, x_2)$	$(3.00, 3.00)^T$	2.9511e-49	2.6815e-195	4	5.8233
$G(x_1, x_2)$	$(2.60, 1.50)^T$	1.0829e-49	6.7038e-196	4	5.8689
$M(x_1, x_2)$	$(10.41, -7.41)^T$	2.3053e-41	1.0439e-113	4	5.9611
$N(x_1, x_2)$	$(3.00, 0.00)^T$	nc	nc	nc	nc
$O(x_1, x_2)$	$(-0.76, -1.16)^T$	7.0387e-71	7.9136e-174	5	-

The outcomes displayed in Tables 3 and 4 are promising as they reveal that MCTC(0) converges to the solution in four out of the five non-linear test systems, regardless of the initial estimates employed. Although the ACOC does not stabilize in some cases, it approaches a value of six when it does stabilize.

Now, we will proceed to evaluate the performance of the MCTC(200) method. The numerical results for initial estimates very close to the solution ($x^{(0)} \approx \xi$) and near the solution ($x^{(0)} \approx 3\xi$) can be found in Tables 5 and 6, respectively.

The findings from Tables 5 and 6 indicate that MCTC(200) encounters convergence issues. Even for initial estimates very close to the root ($x^{(0)} \approx \xi$), this method always converges to the solution. Moreover, for estimates near the roots ($x^{(0)} \approx 3\xi$), MCTC(200) does not converge to the solution in two out of the five cases, demonstrating a dependency on the initial estimate and the specific non-linear test system used. Additionally, the number of iterations increases for

Table 4: Numerical performance of MCTC(0) on test problems for $x^{(0)} > 10\xi$.

System	$x^{(0)}$	$\|\|x^{(k+1)} - x^{(k)}\|\|$	$\|\|F(x^{(k+1)})\|\|$	iter	ACOC
$F(x_1, x_2)$	$(11.00, 11.00)^T$	3.4914e-55	0	5	-
$G(x_1, x_2)$	$(9.53, 5.50)^T$	1.2350e-55	0	5	-
$M(x_1, x_2)$	$(38.18, -27.18)^T$	4.9654e-57	3.4199e-145	5	5.4814
$N(x_1, x_2)$	$(11.00, 0.00)^T$	nc	nc	nc	nc
$O(x_1, x_2)$	$(-2.79, -4.24)^T$	3.7780e-39	2.3868e-110	3	-

Table 5: Numerical performance of MCTC(200) on test problems for $x^{(0)} \approx \xi$.

System	$x^{(0)}$	$\|\|x^{(k+1)} - x^{(k)}\|\|$	$\|\|F(x^{(k+1)})\|\|$	iter	ACOC
$F(x_1, x_2)$	$(0.90, 0.90)^T$	1.9038e-29	4.6447e-116	3	6.0626
$G(x_1, x_2)$	$(0.80, 0.40)^T$	5.1091e-67	1.9467e-208	4	-
$M(x_1, x_2)$	$(3.40, -2.40)^T$	1.6761e-43	2.8365e-119	3	5.9400
$N(x_1, x_2)$	$(0.90, 0.10)^T$	2.0202e-48	7.7869e-208	4	-
$O(x_1, x_2)$	$(-0.20, -0.30)^T$	3.8365e-85	6.7625e-202	4	-

the systems in which the solution is reached, compared to the MCTC(0) method under the same conditions.

As a result, we conclude that the method for $\alpha = 0$ exhibits robustness, converging to the solution with few iterations regardless of the initial estimate and the non-linear test system utilized. On the other hand, the method for $\alpha = 200$ is unstable, tending not to converge to the solution based on the initial estimate and the non-linear test system employed. Furthermore, both methods converge to the solution with an order of 6, validating the theoretical results obtained in previous sections regarding the dynamical behavior and convergence analysis of the MCTC(α) family.

5 Conclusions

In this study, we have successfully developed a highly efficient family of iterative methods designed to tackle nonlinear equations and systems. Through extensive numerical experiments, the MCTC(α) family has demonstrated remarkable numerical performance, particularly when considering stable members as representatives. The results obtained from the experiments align with the theoretical analyses, and the order of convergence, as measured by the ACOC, closely approaches six.

Table 6: Numerical performance of MCTC(200) on test problems for $x^{(0)} \approx 3\xi$.

System	$x^{(0)}$	$\|\|x^{(k+1)} - x^{(k)}\|\|$	$\|\|F(x^{(k+1)})\|\|$	iter	ACOC
$F(x_1, x_2)$	$(3.00, 3.00)^T$	9.6219e-49	3.0304e-193	5	5.7669
$G(x_1, x_2)$	$(2.60, 1.50)^T$	3.4103e-49	7.5761e-194	5	5.8239
$M(x_1, x_2)$	$(10.41, -7.41)^T$	5.0005e-75	1.0922e-182	9	-
$N(x_1, x_2)$	$(3.00, 0.00)^T$	nc	nc	nc	nc
$O(x_1, x_2)$	$(-0.76, -1.16)^T$	nc	nc	nc	nc

Overall, the MCTC(α) family exhibits lower errors and requires fewer iterations to converge to the solution. Particularly noteworthy is the method for $\alpha = 0$, which has proven to be robust and stable, corroborating the results obtained from the complex and real dynamics analyses.

However, it is crucial to acknowledge that the method for $\alpha = 200$ presents a contrasting behavior. This particular method is found to be unstable and chaotic, leading to an inability to converge to the solution in accordance with the initial estimate and the nonlinear system used.

In summary, the MCTC(α) family of iterative methods has demonstrated its practical and theoretical value in solving nonlinear equations and systems. The stable members within the family showcase excellent performance, offering a viable and efficient solution for various scientific and engineering applications. On the other hand, cautious consideration is needed when employing the method for $\alpha = 200$, as it may lead to convergence issues and instability, depending on the specific problem at hand.

References

1. Kansal, M., Cordero, A., Bhalla, S., Torregrosa, J. R., New fourth- and sixth-order classes of iterative methods for solving systems of nonlinear equations and their stability analysis. *Numerical Algorithms*, 87:1017–1060, 2021.
2. Hueso, J. L., Martínez, E., Teruel, C., Convergence, effiency and dinamics of new fourth and sixth order families of iterative methods for nonlinear systems. *Comput. Appl. Math.*, 275:412–420, 2015.
3. Cordero, A., Soleymani, F., Torregrosa, J. R., Dynamical analysis of iterative methods for nonlinear systems or how to deal with the dimension?. *Applied Mathematics and Computation*, 244: 398–412, 2014.
4. Cordero, A., Moscoso-Martínez, M., Torregrosa, J. R., Chaos and Stability in a New Iterative Family for Solving Nonlinear Equations *Algorithms*, 14(4):1–24, 2021.
5. Moscoso-Martínez, M.; Chicharro, F.I., Cordero, A., Torregrosa, J.R., Performance of a New Sixth-Order Class of Iterative Schemes for Solving Non-Linear Systems of Equations *Mathematics*, 11(1374):1–15, 2023.
6. Ortega, J. M., Rheinboldt, W. C., Iterative Solution of Nonlinear Equations in Several Variables *Academic Press*, New York, 1970.
7. A. Cordero, Juan R. Torregrosa, Variants of Newton's Method using fifth-order quadrature formulas *Applied Mathematics and Computation*, 190(1):686–698, 2007.

Parametric family of derivative-free multi-step vectorial methods with weight function

Alicia Cordero[1], Eva G. Villalba[1] Neus Garrido[1], Juan R. Torregrosa[1], and Paula Triguero-Navarro[1]*

Universitat Politècnica de València, València, Spain,
* ptrinav@doctor.upv.es

Abstract. In this work, we show a derivative-free parametric family with matrix weight function for solving systems of nonlinear equations, where the elements of the family can have as many steps as we wish. The element of the family with m-steps has order of convergence $2m$ if the weight function verifies some certain conditions. In this work, the efficiency index is studied in order to see, depending on the size of the problem to be solved, which of the elements of the family is more recommendable.

Keywords: Iterative methods, Nonlinear Systems, Procedures with memory, Parametric Family, Derivative-free, Jacobian-free, Multistep procedure, Steffensen-type scheme

1 Design

There are many iterative methods that use Jacobian matrices in their iterative expressions in order to solve nonlinear systems, but when the problem is not differentiable or it is costly to calculate them, we must rely on Jacobian-free schemes.

In this paper, we present the following derivative-free parametric family with weight function, denoted as $M_{\beta,\delta}S_m$, defined and studied in [3] with the following iterative expression for m steps:

$$\begin{cases} z_1^{(k)} = x^{(k)} - [w^{(k)}, x^{(k)}; F]^{-1}F(x^{(k)}), \\ z_2^{(k)} = z_1^{(k)} - H(t^{(k)})[w^{(k)}, x^{(k)}; F]^{-1}F(z_1^{(k)}), \\ \quad \vdots \qquad\qquad\qquad \vdots \\ z_{m-1}^{(k)} = z_{m-2}^{(k)} - H(t^{(k)})[w^{(k)}, x^{(k)}; F]^{-1}F(z_{m-2}^{(k)}), \\ x^{(k+1)} = z_{m-1}^{(k)} - H(t^{(k)})[w^{(k)}, x^{(k)}; F]^{-1}F(z_{m-1}^{(k)}), \end{cases} \tag{1}$$

where $t^{(k)} = [w^{(k)}, x^{(k)}; F]^{-1}[z_1^{(k)}, v^{(k)}; F]$, being $w^{(k)} = x^{(k)} + \beta F(x^{(k)})$ and $v^{(k)} = z_1^{(k)} + \delta F(z_1^{(k)})$.

The semilocal convergence when the matrix weight function is $H(t^{(k)}) = {t^{(k)}}^{-1}$ for this family is studied in [5], where, in addition, several numerical experiments and dynamical planes are performed to see and study the behaviour of different elements of the family by changing the values of the parameters.

What we will show next is the efficiency index of the elements of the family when a system with size n is solved, in order to see if by increasing the number of steps the efficiency of the methods becomes worse, knowing that the family has order $2m$ as long as m is the number of steps that the method has and the weight function satisfies that $H(I) = I$ and $H'(I) = -I$, where I is the identity matrix.

2 Efficiency index

We will now look at the computational cost of comparing the efficiency of using more steps. We will use the concept of efficiency, which is usually measured by the efficiency index, defined in [7]:

$$I = p^{\frac{1}{d}},$$

where p is the order of convergence and d is the number of functional evaluations per iteration.

This comparison criterion is very useful as it establishes a relationship between the order of convergence of a method and the number of functional evaluations it performs per iteration.

For a system of size $n \times n$,

- n functional evaluations are required for a vector function F,
- n^2 functional evaluations for a Jacobian matrix J_F,
- and $n^2 - n$ functional evaluations for a first order divided-difference operator of the form $[x, y; F]$ (see [8]).

In this article, we make use of these concepts to calculate the efficiency index of the iterative family without memory (1) as a function of the number of steps m it performs.

- For $m = 1$, it performs an evaluation of F and calculate a divided difference operator, so the total number of functional evaluations is n^2. Therefore, the efficiency index is:

$$2^{\frac{1}{n^2}}.$$

– For $m = 2$, it performs two functional evaluations of F and two divided difference operators, so the total number of functional evaluations is $2n^2$. So, the efficiency index is:

$$4^{\frac{1}{2n^2}} = 2^{\frac{1}{n^2}}.$$

– For $m > 2$, it performs m functional evaluations of F and two divided difference operators, so the total number of functional evaluations is $2n^2 + (m-2)n$. The resulting efficiency index is

$$I_m = (2m)^{\frac{1}{2n^2+(m-2)n}}.$$

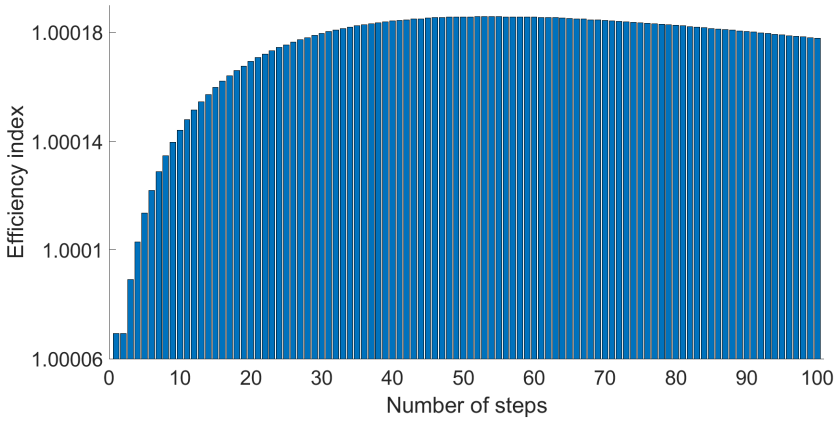


Fig. 1: Efficiency rates when the system size is 100.

The efficiency indices for solving a system of size 100×100 are shown in Figure 1. It can be seen in this figure that methods that perform the most or the least steps are not those that obtain the highest efficiency rates. From [3], we obtain that the number m that maximizes the efficiency index is $m* \approx 53.82$. Therefore, we must compare the values obtain for $m^- = 53$ and $m^+ = 54$. In this case, $I_{54} > I_{53}$, therefore, the number of steps that obtains higher efficiency index is 54.

3 Numerical experiments

In this section, first, we apply the multistep methods $M_{\beta,\delta}S_m$ to a nonlinear system to verify that the properties deduced theoretically in the analysis of the family are satisfied, both with and without memory.

We want to approximate the solution of the following nonlinear system with n equations and n unknowns

$$F_i(x) = x_i \sin(x_{i+1}) = 1, \qquad i \in \{1, \ldots, n-1\}$$
$$F_n(x) = x_n \sin(x_1) = 1.$$

The approximate solution of this system is $\alpha \approx (1.11416, \ldots, 1.11416, \ldots)^T$, which we try to approximate using the methods $M_{\beta,\delta}S_m$, for different values of m.

For the computational calculations we use MATLAB R2022a, using variable precision arithmetic with 5000 digits, iterating from an initial estimate $x^{(0)} = [1.3, \ldots, 1.3]^T$ until the following stopping criterion is satisfied:

$$|x^{(k+1)} - x^{(k)}\|_2 + \|F(x^{(k+1)})\|_2 < 10^{-300}$$

and the approximated computational order of convergence (ACOC), defined by Cordero and Torregrosa in [4].

Table 1 shows the results obtained by the above methods to solve the system, taking $n = 15$ and assuming that the weight function of the family of methods has the expression $H(t^{(k)}) = 3I_n - 3t^{(k)} + (t^{(k)})^2$.

The data we compare in Table 1 symbolize, from left to right, the multistep methods used for different steps $m = 1, 2, 3$, the distance between the last two iterations, the value of the function evaluated in the last iteration, the number of iterations needed to verify the stopping criterion, the approximate computational convergence order defined in [4] and the time it takes for each method to find an approximation to α, satisfying the required tolerance.

Method	$\|x^{(k+1)} - x^{(k)}\|$	$\|F(x^{(k+1)})\|$	Iteration	ACOC	Time
$M_{0.1,0.1}S_1$	6.32894×10^{-439}	6.95714×10^{-879}	9	1.99999	43.1406
$M_{0.1,0.1}S_2$	2.83143×10^{-508}	$1.74600 \times 10^{-2036}$	5	3.99999	55.0469
$M_{0.1,0.1}S_3$	7.33069×10^{-408}	$1.23838 \times 10^{-2452}$	4	5.99999	42.7188

Table 1: Numerical results for $M_{0.1,0.1}S_m$ with $m = 1, 2, 3$.

It is easy to check in Table 1 that, in this case, the method that performs three steps needs less iterations than the other to satisfy the required tolerance and the computational time is fewer or equal to the others. In all the case, the ACOC matches the theoretical order of convergence expected.

4 Conclusions

In this paper, we show the efficiency index of a derivative-free multi-step parametric family with matrix weight function, in order to illustrate that in certain occasions, although some iterative methods need more functional evaluations, they are more efficient if the size to be solved is large.

4.1 Financial support

This research was partially supported by Universitat Politècnica de València Contrato Predoctoral PAID-01-20-17 (UPV).

References

1. Chicharro, F., Cordero, A., Garrido, N. & Torregrosa, J. On the effect of the multidimensional weight functions on the stability of iterative processes. *Journal Of Computational And Applied Mathematics*. **405** pp. 113052 (2022)
2. Chicharro, F., Cordero, A., Garrido, N. & Torregrosa, J. On the improvement of the order of convergence of iterative methods for solving nonlinear systems by means of memory. *Applied Mathematics Letters*. **104** pp. 106277 (2020)
3. Cordero, A., G. Villalba, E., Torregrosa, J. & Triguero-Navarro, P. Introducing memory to a family of multi-step multidimensional iterative methods with weight function. *Expositiones Mathematicae*. **41**, 398-417 (2023)
4. Cordero, A. & Torregrosa, J. Variants of Newton's Method using fifth-order quadrature formulas. *Applied Mathematics And Computation*. **190**, 686-698 (2007)
5. G. Villalba, E., Martínez, E. & Triguero-Navarro, P. Semilocal Convergence of a Multi-Step Parametric Family of Iterative Methods. *Symmetry*. **15** (2023)
6. Ortega, J. M.; Rheinboldt, W. C., *Iterative solution of nonlinear equations in several variables*, Academic Press, (1970).
7. Traub, J., *Iterative Methods for the Solution of Equations*, Prentice-Hall, (1964).
8. Potra, F.; Pták, V., *Nondiscrete Induction and Iterarive Processes*, Pitman Publishing, (1984).

Part VI

Mathematical Models in Social Science and Financial Mathematics

Choquet integral for finite sets: new expression, computation, and applications (a ChatGPT-driven experience)

José Carlos R. Alcantud[1]

BORDA Research Group and IME,
University of Salamanca, Spain.
jcr@usal.es

Abstract

Abstract. The Choquet integral was developed in 1954 by Gustave Choquet [7]. It was later applied in decision making under uncertainty since the pioneering work of Schmeidler [12], who proved its most popular characterization. However it was not used for multi-criteria decision aid (MCDA) till the decade of the 1990s [10]. Nowadays many generalizations exist [8], and the topic is the subject of intense research and development in many areas [2, 11, 13].
In this chapter we narrate our experience with ChatGPT as an assistant to explore the utilization of the discrete Choquet with *Mathematica*, and to write a Beamer presentation with our findings. The goal of this exercise is to prepare the ground for the application of the discrete Choquet integral to N-soft sets [9].

Keywords: capacity; Choquet integral; N-soft set; ChatGPT.

1 Introduction

The focus of this chapter is the Choquet integral, particularly, its discrete version. This integral is defined with respect to a very important class of non-necessarily additive set functions called Choquet capacities, or fuzzy measures, after the work of Choquet [7]. In this chapter the sets that we explore are all finite although of course, a continuous version of the Choquet integral exists. In fact Schmeidler [12] produced a characterization of this monotonic functional, inspired by the foundations of Bayesian decision theory. Indeed, just like probability has its foundation in measure theory, degrees of belief are generalizations of probability that have their foundation in capacity theory. In formal terms, capacities provide a generalization of measure theory that dispenses with any additivity requirement.

Despite this motivation it is important to emphasize that originally, capacities were not meant to encapsulate a general type of uncertainty. Instead, in their inception their main applications were targeted to potential theory. However, capacities soon found applications in other fields like stochastic processes.

And they have been applied to multi-criteria decision aid (MCDA) since the decade of the 1990s [10].

To understand their semantics, two main interpretations of the figures given by a capacity are especially valuable. The first one conceives capacities as a representation of the importance of the elements of a set. It may be a set of properties in the framework of MCDA, or a set of agents with their own degree of expertise in the context of group decision making. Under the second interpretation, we are concerned with the uncertainty of events. In this case the value assigned to a set quantifies the uncertainty that this set contains the result of an event (for example, an experiment).

As for technicalities, the next section describes the components and definition of the discrete version of the Choquet integral.

2 The discrete Choquet integral

Let us fix $X = \{1, \ldots, n\}$. X may represent either a set of n properties (in MCDA) or experts (in group decision making), or the results of an event with n possible outcomes, depending on the intended application. We will not be concerned with the later interpretation here.

Definition 2. A discrete fuzzy measure (or **capacity**) is a mapping $\mu : 2^X \longrightarrow [0, 1]$ that is monotonic (this means $\mu(S) \leqslant \mu(T)$ whenever $S \subseteq T \subseteq X$), and it is such that $\mu(\varnothing) = 0$ and $\mu(X) = 1$.

The requirement $\mu(X) = 1$ means that the capacity is normalized. Some studies dispense with this property in the definition of capacity.

The capacity satisfies additivity if when $A, B \subseteq X$ are disjoint subsets, it must be the case that $\mu(A \cup B) = \mu(A) + \mu(B)$. Note that additive set functions are uniquely determined by n (instead of 2^n) values, namely, $\mu(1), ..., \mu(n)$. In addition, any additive discrete capacity is a probability measure. This observation crucially hinges on the fact that the capacity is normalized (i.e., that it satisfies $\mu(X) = 1$).

We say that there are **synergies** between A and B, disjoint subsets of X, with respect to μ when $\mu(A \cup B) > \mu(A) + \mu(B)$. There is **redundancy** between A and B when $\mu(A \cup B) < \mu(A) + \mu(B)$. And there is no interaction between A and B when $\mu(A \cup B) = \mu(A) + \mu(B)$.

Capacities for which there are synergies (resp., redundancies) between every pair of disjoint non-empty subsets of X are called *superadditive* (resp., *subadditive*).

If $\mu(A) = \mu(B)$ for every $A, B \subseteq N$ with the same cardinality, then the capacity μ is **symmetric**.

We are ready to define the main concept in this chapter:

Definition 2. The discrete Choquet integral with respect to μ, a discrete capacity on X, a set with n elements, is the function $C^\mu : \mathbb{R}^n \longrightarrow \mathbb{R}$ defined as $C^\mu(v_1, \ldots, v_n) = \sum_{i=1}^{n} \left[v_{(i)} - v_{(i-1)}\right] \mu(H_i)$, where the vector $\mathbf{v}_\nearrow = (v_{(1)}, \ldots, v_{(n)})$ is a non-decreasing permutation of $\mathbf{v} = (v_1, \ldots, v_n)$, with the convention $v_{(0)} = 0$.

In this formula, we let $H_i = \{(i), \ldots, (n)\}$ be the set of indices that correspond to the largest $n - i + 1$ components of $\mathbf{v}$.

In case that μ is a symmetric capacity, Definition 2 produces an OWA operator [14]. OWA holds for "ordered weighted average". To apply this operator, a fixed set of weights produces the weighted averages of the *ordered* values of the vector of values. By contrast, Definition 2 is more general because it produces a weighted sum of successive increments, with weights computed by the capacity (of the subsets of attributes that guarantee the respective increments in the non-decreasing vector of evaluations).

Figure 1 summarizes the computation of the value assigned by the Choquet integral to the vector of values $(2, 9, 6)$. It is expressed as a function of the values of a capacity on a set of 3 properties. The computations are

$$C^{\mu}(2, 9, 6) = 2 \cdot \mu(\{1, 2, 3\}) + 4 \cdot \mu(\{2, 3\}) + 3 \cdot \mu(\{2\}).$$

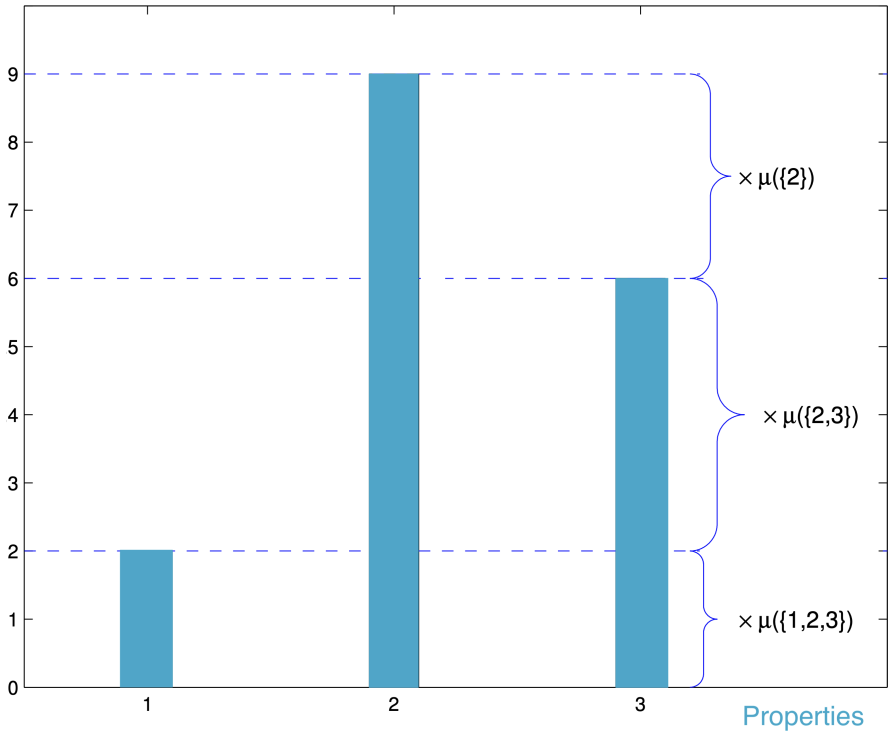


Fig. 1: A visual presentation of the computations leading to the evaluation of a three-dimensional vector $(2, 9, 6)$ by a Choquet integral.

3 Exercises assisted by ChatGPT: a failed attempt

If one needs to write a more complete example of the application of the Choquet integral, then a handicap is that the definition of the capacity is computationally costly. Note that one needs to define 2^n figures, n being the number of properties or experts listed in X. In addition, these figures must meet the requirements of Definition 2. The problem of producing capacities randomly has been the subject of interesting analysis such as Beliakov and Wu [6]. This section explores the possible utilization of ChatGPT to produce capacities on sets of small cardinalities. A brute-force algorithm can surely get it in little time, and this is probably what we expected from ChatGPT. Let us narrate what happened.

Figures 2 and 3 show selected pieces of a conversation with ChatGPT-3 attempting to solve this problem.

First we asked if ChatGPT can help us to solve problems involving the Choquet integral, The answer was affirmative. In fact, ChatGPT asked if more details could be provided about our specific problem. Emboldened by this enthusiastic answer, then we asked if ChatGPT could write *Mathematica* code for the generation of a fuzzy measure on a set with only 4 elements. Again, the answer was affirmative. But despite its acclaimed abilities for the production of high-quality code, the result was rather disappointing. Especially, because the explanation (and the code) that ChatGPT produced ignored the monotonicity requirement altogether. Figure 2 shows this piece of the conversation between the user and ChatGPT.

Figure 3 shows how this conversation unfolded as we tried to clarify the matter. First we asked whether the fuzzy measure produced with the *Mathematica* code was forcefully monotonic. Not only ChatGPT assured that this was the case, but also it added a complicated explanation involving the Sugeno integral as the germ of the construction of the capacity. We retorted that we had no hint that the Sugeno integral was used in the algorithm, and also that we had run the code and the result was not monotonic. ChatGPT apologized but insisted that it could still use the Sugeno integral to produce the code that we had requested.

In the end, we gave up. The only part of the code that we could make use of concerned the generation of the list of subsets of our set (one line of code).

Remark. *Mathematica* users should know that the Wolfram Function Repository hosts a "ChoquetIntegral" function. It can be downloaded from the URL

https://resources.wolframcloud.com/FunctionRepository/resources/ChoquetIntegral/

According to the description provided, "ChoquetIntegral" calculates numerically the Choquet integral of a function with respect to a probability distribution. Hence we cannot use the full force of this aggregation operator because we need to use additive (and normalized) capacities.

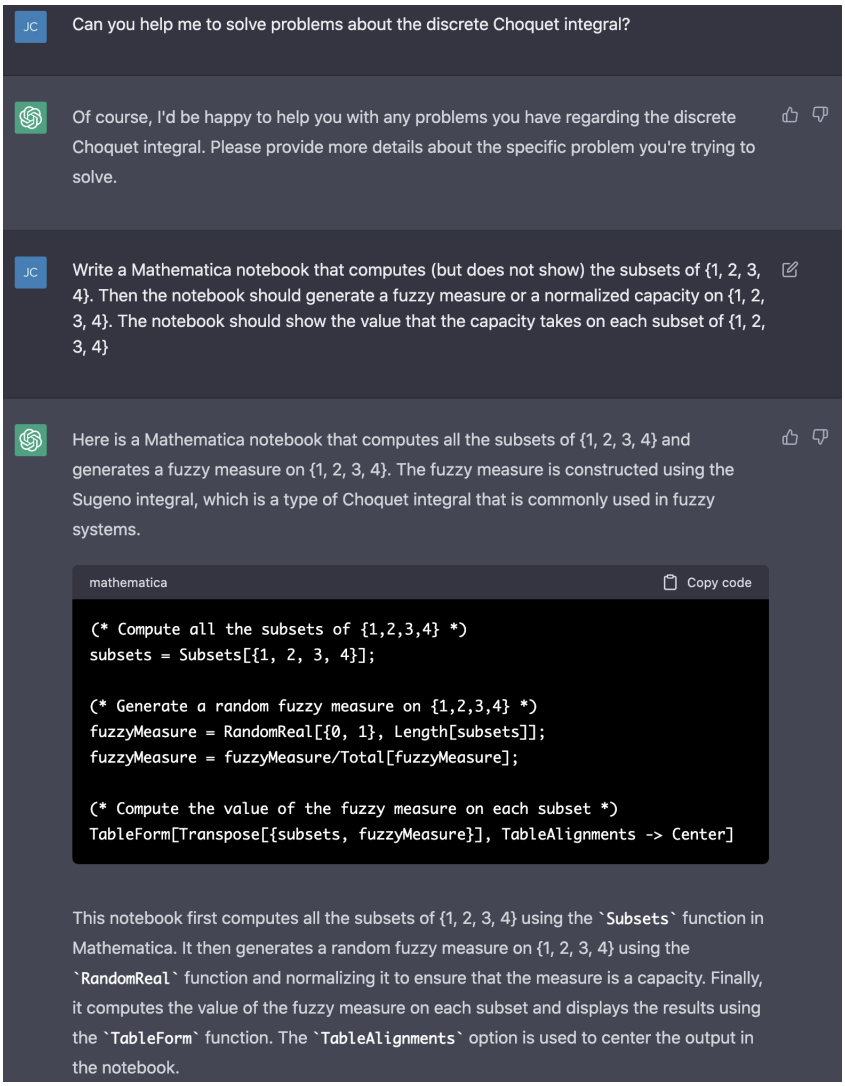


Fig. 2: First part of conversation with ChatGPT-3.

Fig. 3: Second part of conversation with ChatGPT-3.

4 Intended applications

Our interest in the exercises described above lies in the fact that the field of N-soft set theory appears to be a perfect fit for the application of the discrete Choquet integral in various ways.

Let us first explain that soft sets over a set extend its subsets with the help of several simultaneous characteristic functions. Under the most common interpretation, each characteristic function is associated with one of the several characteristics that explain the alternatives in our model. However many real situations can convince us that quite often, alternatives are better characterized in terms of their properties if there are several grades that quantify how satisfied we are for every feature. With this motivation, N-soft sets were defined by Fatimah *et al.* [9] in order to improve the informational ability of the soft set model. To understand its possibilities, Alcantud [1] has discussed the semantics of N-soft sets recently.

In practical terms, an N-soft set can be presented by Table 1 when both the set of options $U = \{o_1, \ldots, o_p\}$ and the set of properties $X = \{x_1, \ldots, x_n\}$ are finite. Table 1 contains exclusively 0's and 1's if we have a soft set. In an N-soft set, the v_{ij} values are numbers in the range $\{0, 1, 2, \ldots, N-1\}$. Real examples that adopt this form have been given in articles such as [1, 3–5] among others.

Table 1: A general N-soft set on alternatives $U = \{o_1, \ldots, o_p\}$ when the properties are $X = \{x_1, \ldots, x_n\}$.

	x_1	x_2	.....	x_n
o_1	v_{11}	v_{12}	.....	v_{1n}
o_2	v_{21}	v_{22}	.....	v_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_p	v_{p1}	v_{p2}	.....	v_{pn}

Within this framework, two direct applications of the discrete Choquet integral come to mind easily. Their practical implementation and respective comparative analyses are beyond the scope of this chapter.

First, the Choquet integral can be used to refine the evaluations of the alternatives o_i provided in articles such as [9], so that alternatives with higher evaluations are preferred over alternatives with smaller values. In fact this article defined an extended weighted choice value that uses a vector of weights for the properties. As explained above, OWAs and discrete Choquet integrals produce a richer environment for the evaluations, since capacities are able to rate not only individual attributes or experts, but also their combined influence in the final assessment. The utilization of the Choquet integral in this context

remains unexplored. It paves the way to considering attributes with synergies or redundancies in N-soft set decision making, because capacities are able to encapsulate these circumstances faithfully.

Secondly, Choquet integrals can be utilized to aggregate several N-soft sets with equal sets of alternatives and attributes. Suppose for example that for the evaluation of a decision situation, each N-soft set is submitted by an expert. The capacity should be now designed to assess the abilities of the experts, in such way that the capacity captures the possible existence of synergies or redundancies. As of now, simpler methodologies that address this aggregation problem include Alcantud *et al.* [4, 5].

5 Preparing the slides for our presentation

The slides presented at the conference were prepared manually. However, we asked ChatGPT if it was able to suggest a Beamer theme with a layout similar to the `metropolis` theme that we had used before for our presentations.

Figure 4 shows this piece of the conversation.

ChatGPT suggested not one, but five themes. Peculiarly, the first one was called "Fira". We knew the Fira typefaces but we were surprised to find out that a Beamer theme existed with the same name. After several unsuccessful searches with a powerful search engine, we asked ChatGPT what web page shows information about the Fira theme. Peculiarly, the answer pointed at https://ctan.org/pkg/fira Not surprisingly, this page of the CTAN repository contains the Fira fonts with LaTeXsupport.

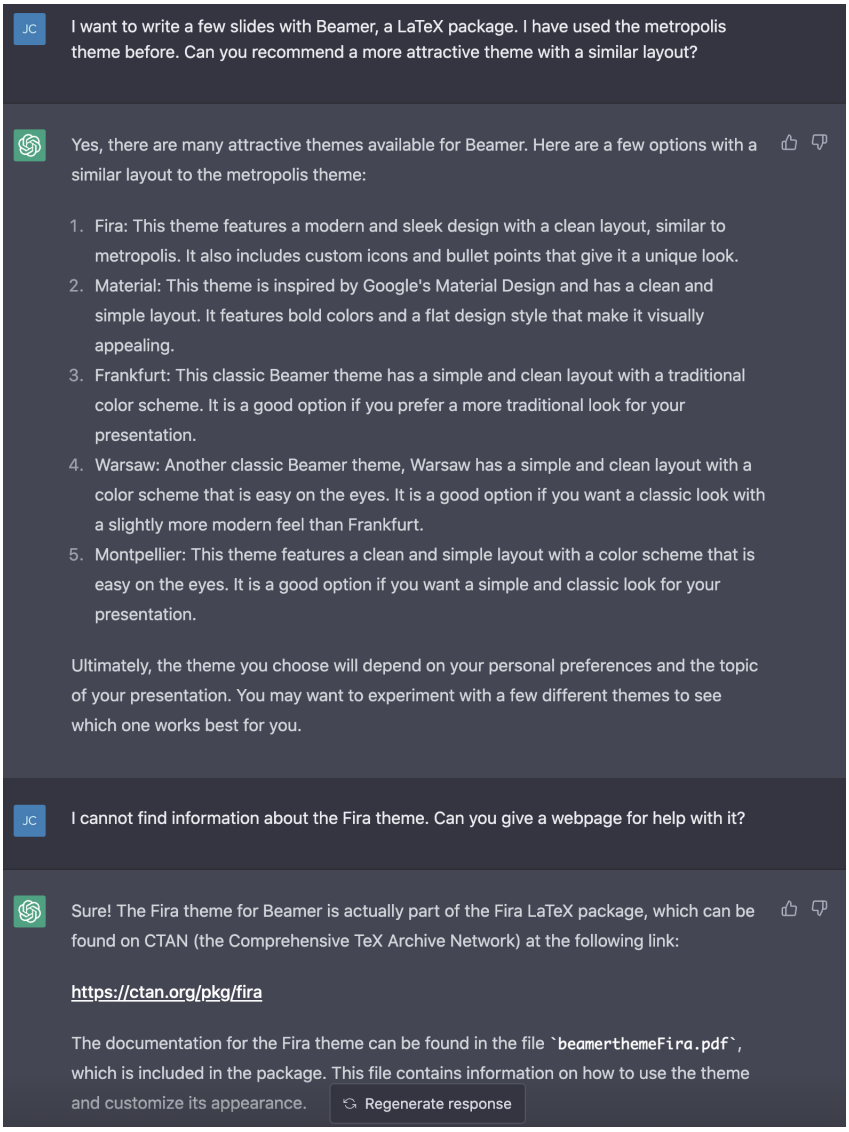


Fig. 4: Asking ChatGPT-3 what Beamer theme we should use for our presentation.

Acknowledgments

The author is grateful to the Junta de Castilla y León and the European Regional Development Fund (Grant CLU-2019-03) for the financial support to the Research Unit of Excellence "Economic Management for Sustainability" (GECOS).

References

1. Alcantud, J.C.R., The semantics of N-soft sets, their applications, and a coda about three-way decision. *Information Sciences* 606:837–852, 2022.
2. Alcantud, J.C.R., de Andrés, R., The problem of collective identity in a fuzzy environment. *Fuzzy Sets and Systems* 315:57–75, 2017.
3. Alcantud, J.C.R., Feng, F., Yager, R.R., An N-soft set approach to rough sets. *IEEE Transactions on Fuzzy Systems*, 28(11):2996–3007, 2020.
4. Alcantud, J.C.R., Santos-García, G., Akram, M., OWA aggregation operators and multi-agent decisions with N-soft sets. *Expert Systems with Applications* 203:117430, 2022.
5. Alcantud, J.C.R., Santos-García, G., Akram, M., A novel methodology for multi-agent decision-making based on N-soft sets. *Soft Computing*, 2023. Forthcoming.
6. Beliakov, G., J.-Z., Random generation of capacities and its application in comprehensive decision aiding. *Information Sciences*, 577:424–435, 2021.
7. Choquet, G., Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954.
8. Dimuro, G., et al., The state-of-art of the generalizations of the Choquet integral: From aggregation and pre-aggregation to ordered directionally monotone functions. *Information Fusion* 57:27–43, 2020.
9. Fatimah, F., Rosadi, D., Hakim, R. F., Alcantud, J. C. R. N-soft sets and their decision making algorithms. *Soft Computing*, 22:3829–3842, 2018.
10. Grabisch, M., The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89(3):445–456, 1996.
11. Meng, F., Chen, S.-M., Tang, J., Multicriteria decision making based on bi-direction Choquet integrals. *Information Sciences* 555:339–356, 2021.
12. Schmeidler, D., Integral representation without additivity. *Proceedings of the American Mathematical Society*, 97(2):255–261, 1986.
13. Torra, V., Narukawa, Y, The h-index and the number of citations: Two fuzzy integrals. *IEEE Transactions on Fuzzy Systems*, 16(3):795–797, 2008.
14. Yager, R.R., On ordered weighted averaging aggregation operators in multicriteria decision making, *IEEE Transactions on Systems, Man, and Cybernetics*, 18:183–190, 1988.

Successful romantic relationships explained by differential games

Jorge Herrera de la Cruz[1] and José-Manuel Rey[2]

[1] Department of Mathematics and Data Science. CEU San Pablo. Madrid.
[2] Department of Economic Analysis, Complutense University of Madrid.
j-man@ccee.ucm.es, ecomap.ucm.es/rey/

Abstract. In previous work, we modeled a romantic relationship as an optimal control problem, either one-person or two-person (differential game), either in a deterministic or a stochastic environment –see [12], [8] and [9]. In this note, we show how real data on happiness in romantic relationships can be accurately approximated by our computational models. We explain how to calibrate our model to fit the data. Relevance and implications are discussed.

Keywords: Romantic relationships. Marital satisfaction. Parameter Estimation, Differential Games.

1 Differential games as successful romantic relationships

Understanding the success of long-term romantic relationships is a substantial issue with huge implications for the well-being of individuals and society [3], [2]. The study of successful (long-lasting and happy) unions poses a challenging problem for the social sciences. In fact, despite many decades of scientific research, what is required to build a successful relationship is not well understood [5], [4]. Our research program aims to understand the underlying dynamics and time evolution of romantic relationships, particularly those that are successful, using mathematical models. It began over a decade ago with the design of a romantic relationship as an optimal control problem [12] –see also [13]. This love engineering approach was extended recently to account for a coupled two-person control problem, i.e. a differential game, both in a deterministic and in a stochastic environment –see [8] and [9], respectively.

The basics of our modeling of a romantic relationship as a stochastic differential game are as follows. The state of the relationship at time $t \geq 0$ is monitored by the (random) variable $x(t)$, called *feeling*, where $x : [0, \infty) \to X \subseteq \mathbb{R}$, X being the state space. So $x(t)$ gives a measure of relationship quality, or *marital satisfaction* as often called in the literature. –see e.g. [1]. Now, the evolution of the feeling variable obeys the following stochastic differential equation (SDE)

$$\mathrm{d}x\left(t\right) = \left[-rx(t) + a_1c_1(t) + a_2c_2(t)\right]\mathrm{d}t + \sigma\left(x\left(t\right)\right)\mathrm{d}w, \tag{1}$$

where $r, a_1, a_2 > 0$ are parameters, $c_i : [0, \infty) \to \mathbb{R}^+$, $i = 1, 2$, are (piecewise) continuous functions that represent the effort that each partner puts into

the relationship over time, and $w(t)$ is a Wiener process (see e.g. [10]). In the scheme of control theory, $c_i(t)$ are control variables that govern the evolution of the state variable $x(t)$, given r, a_1, a_2, σ. According to [5], equation (1) may be called the second law of thermodynamics for romantic relationships. In case that $\sigma(\cdot) = 0$, equation (1) is a deterministic differential equation. We will consider both the stochastic and deterministic versions of the second law for our analysis below.

Effort controls defined by $c_i = S_i(x)$, with $S_i : X \to \mathbb{R}^+$, are of particular interest here; they are called *feedback* controls (or strategies) . Feedback controls allow partners to decide how much effort to exert at any time t observing the state of the relationship $x(t)$. We assume that both partners independently seek to maximize their total (expected) happiness throughout the duration of the relationship, which is the aggregate difference between the benefits derived from the level of feeling –measured by certain functions $U_i(x)$- and the costs of exerting effort –given by suitable functions $D_i(c_i)$. Formally, the expected happiness of a partner i, $i = 1, 2$, is given by

$$\mathbb{E}\left[\int_0^\infty e^{-\rho_i t}\left(U_i\left(x\left(t\right)\right) - D_i\left(c_{i1}\left(t\right); c_i^*\right)\right) \mathrm{d}t | x(0) = x_0\right],$$

where ρ_i is a discount parameter accounting for impatience, and the parameter c_i^* indicates the level of effort that partner i is most happy to make. More specifically, c_i^* is the absolute minimum of the cost function D_i, which is assumed to be differentiable and strictly convex. Also, U_i is assumed to be differentiable, increasing, and strictly concave. These mathematical properties are derived from principles in human psychology (see [12] for the details). In the deterministic case, there is no need to include the operator $\mathbb{E}(\cdot)$, as the functional objectives are no longer probability distributions.

The relationship follows an equilibrium when there is a pair of optimal feedback controls $\left(S_1^\heartsuit(\cdot), S_2^\heartsuit(\cdot)\right)$, that is, $S_1^\heartsuit(x(t))$ solves

$$\max_{c_1(t)} \mathbb{E}\left(\int_0^\infty e^{-\rho_1 t}\left(U_1\left(x\left(t\right)\right) - D_1\left(c_1\left(t\right); c_1^*\right)\right) \mathrm{d}t | x(0) = y\right)$$

with $\mathrm{d}x(t) = \left[-rx(t) + a_1 c_1(t) + a_2 S_2^\heartsuit(x(t))\right] \mathrm{d}t + \sigma(x(t)) \mathrm{d}w$, and, also, $S_2^\heartsuit(x(t))$ solves

$$\max_{c_2(t)} \mathbb{E}\left(\int_0^\infty e^{-\rho_2 t}\left(U_2\left(x\left(t\right)\right) - D_2\left(c_2\left(t\right); c_2^*\right)\right) \mathrm{d}t | x(0) = y\right)$$

with $\mathrm{d}x(t) = \left[-rx(t) + a_1 S_1^\heartsuit(x(t)) + a_2 c_2(t)\right] \mathrm{d}t + \sigma(x(t)) \mathrm{d}w$, where $x(0) = y \in X$ is the initial feeling state and $c_i(t) \in \mathbb{R}^+$ for $t \geq 0$. Such a pair $\left(S_1^\heartsuit(\cdot), S_2^\heartsuit(\cdot)\right)$ is called a (stationary) *feedback* Nash Equilibrium for the relationship. Once the feedback controls of the problem are obtained, the optimal effort (random) paths are given by $c_i^\heartsuit(t) = S_i^\heartsuit(x(t))$, $i = 1, 2$, and the feeling

trajectory of the relationship $x^{\heartsuit}(t)$ is found by solving the SDE

$$\mathrm{d}x\left(t\right)=\left[-rx(t)+a_1S_1^{\heartsuit}\left(x\left(t\right)\right)+a_2S_2^{\heartsuit}\left(x\left(t\right)\right)\right]\mathrm{d}t+\sigma\left(x\left(t\right)\right)\mathrm{d}w,$$

for the initial state $x_0 \in X$. For a detailed presentation of the formulation above we refer to [8] –see also [12].

Finding feedback controls for a couple's relationship is far from trivial and, typically, only approximate solutions can be obtained from computational methods. We will use the RaBVItG algorithm, recently introduced by [7], to compute numerical feedback Nash equilibria for our problems.

2 Successful romantic relationships as differential games

It is shown in [8] and [9] that RaBVItG can efficiently solve computational versions of the differential game models described above. The synthetic trajectories of feeling and effort provided by the algorithm are useful to investigate qualitative properties of romantic relationships in the long term. The purpose of this note is to show that our computational models can indeed replicate actual trajectories of marital satisfaction. Specifically, we will show how the model can approximate a representative trajectory $x_{data}(t)$ obtained from the data set of a questionnaire recently published in [15]. As a by-product of the model approximation, our analysis produces effort control paths and happiness trajectories, which are relevant to understanding the dynamics of genuine successful relationships.

Notice that the scheme in the above section -either deterministic or stochastic- amounts to finding feedback controls and feeling trajectories for a set of hyperparameters. Assume that functions U_i, D_i, for $i = 1, 2$, are defined and are fixed for the sequel. Our computational scheme acts as a mapping

$$P := (r, a_1, a_2, \sigma, c_1^*, c_2^*) \mapsto (c_1^{\heartsuit}(t), c_2^{\heartsuit}(t), x^{\heartsuit}(t)).$$

Essentially, our goal is to solve an inverse problem, that is, to find the set of input parameters $\hat{P}$ that produces the best approximation $\hat{x}^{\heartsuit}(t)$ of the computational model –solved by RaBVItG– to the observed sequence $x_{data}(t)$. We consider both deterministic and stochastic versions of the differential game model. Next, we first describe the data set and the target trajectory $x_{data}(t)$, then we present the method for obtaining the best approximate solution of the model. In the third section, we discuss the result of our analysis.

2.1 Data

The target trajectory $x_{data}(t)$ represents the marital satisfaction of a successful union over 30 years. It is synthesized from raw data gathered in the recent study [15], which is a large cross-cultural study, covering five regions of the world, which took place over a short period (2012-2013). The questionnaire, known as

the Kansas Marital Satisfaction Scale (KMSS) (see [11, 14]), evaluated marital happiness among approximately 7000 couples from 33 different countries. The questionnaire consisted of three questions: "How satisfied are you with your marriage?"; "How satisfied are you with your wife/husband as a spouse?"; and "How satisfied are you with your relationship with your wife/husband?". Participants responded to these questions using a 7-point scale with higher values indicating greater marital satisfaction, from 1 (very dissatisfied) to 7 (very satisfied). For our study, we defined a measure of marital satisfaction (MMS) for each relationship as the sample average of the responses to the three questions above. So, MMS can be considered a proxy for the *feeling* variable in our model. The duration t of each relationship to date, measured in years, was considered to define the target trajectory at moment t, which is represented in red in Figure 1. To create this figure, we calculated the mean and standard error of the MMS in the sample of relationships each year. In our study, the target trajectory is $x_{data}(t)$, which is representative of the general trend of marital satisfaction around the world. This is represented in Figure 1 by the smoothed red line, which is obtained from the data points by local averaging.

Fig. 1: Marital satisfaction and duration of marriage –derived from the data set in [15]. The shaded area represents the 95% confidence interval for the population mean of the MMS obtained from the data. The target trajectory $x_{data}(t)$ is the smoothed red line obtained by locally averaging the MMS data points. The smoothing has been obtained with the R package *ggplot2*. The second graph shows the histogram of the duration of marriages (in years) in the sample. The first and third quartiles are plotted in blue, while the median is on the red line.

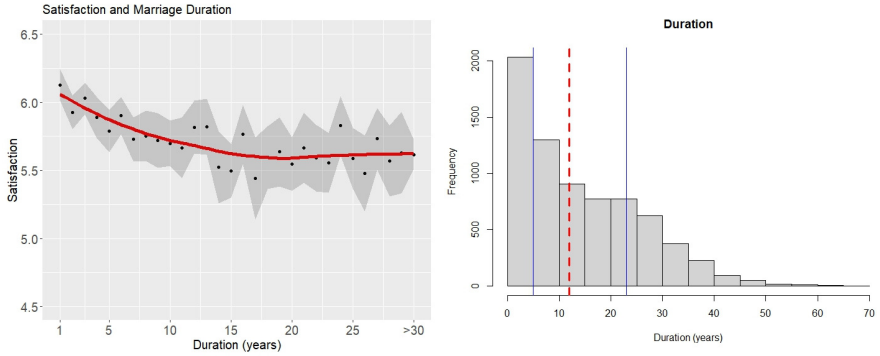


Figure 1 shows that 75% of relationships have a duration of less than 25 years. This fact influences the accuracy of measuring the central effect and therefore affects the variability shown in the vertical range of the shaded area. The evolution of marital satisfaction is described by the smoothed red trajectory: it starts at its maximum value 6.12 (in a range of 0–7) and decreases sharply during the

first ten years. It then continues to decline at a slower rate until it reaches a plateau. The shape of the red curve appears to be the typical pattern of marital satisfaction over time. It has been previously obtained with different samples and different time periods –see [1,16]. Furthermore, the shapes of the feeling trajectories generated by our computational models, both deterministic and stochastic, are qualitatively similar to that of the curve $x_{data}(t)$. This suggests that our model could be calibrated, by finding the appropriate set of input parameters, such that the computed feeling trajectory $\hat{x}(t)$ is a good approximation to the data set $x_{data}(t)$. Next, we explain how this approximation can be obtained.

2.2 Model calibration

As mentioned above, the benefit and cost functions U_i and D_i, $i = 1, 2$, are fixed and will be defined below. Our goal is thus to find a set of input parameters $\hat{P}$ such that the corresponding feeling trajectory $\hat{x}^{\heartsuit}(t)$ computed by the model (approximately) replicates the target trajectory $x_{data}(t)$.

Recall that, given the set $P = (r, a_1, a_2, \sigma, c_1^*, c_2^*)$ and $x(0)$, the algorithm RaBVItG finds a feedback Nash equilibrium $\left(S_1^{\heartsuit}(\cdot), S_2^{\heartsuit}(\cdot)\right)$ of the computational model –either deterministic or stochastic– that defines the feeling trajectory $x^{\heartsuit}(t)$ along with the effort control paths $c_i^{\heartsuit}(t) = S_i^{\heartsuit}\left(x^{\heartsuit}(t)\right)$, $i = 1, 2$.

Given P, let us write $x_d^{\heartsuit}(t) := d - \text{RaBVItG}[P]$ for the feeling trajectory of the deterministic model computed by the algorithm RaBVItG in [8], and $x_s^{\heartsuit}(t) := s - \text{RaBVItG}[P]$ for the feeling trajectory of the stochastic model computed by the version of RaBVItG in [9].

Now, the best approximation of the deterministic model is obtained by solving the optimization problem

$$\min_P \frac{1}{M} \sum_{k=1}^{M} [(x_{data}(k) - x_d^{\heartsuit}(k)]^2$$

with $x_d^{\heartsuit}(t) := d - \text{RaBVItG}[P]$, and $x_d(0) = x_{data}(0)$.

Similarly, the best approximation of the stochastic model to the target trajectory $x_{data}(t)$ is obtained by solving the optimization problem

$$\min_P \mathbb{E}(\frac{1}{M} \sum_{k=1}^{M} [x_{data}(k) - x_s^{\heartsuit}(k)]^2)$$

with $x_s^{\heartsuit}(t) := s - \text{RaBVItG}[P]$, and $x_s(0) = x_{data}(0)$.

Notice that the target trajectory $\{x_{data}(t) : t = 1, ..., M\}$ consists of a discrete sequence of M values, one for each year. In our study, it is $M = 30$. The best approximations are thus obtained by minimizing the mean squared error between the observed data and the trajectories computed by RaBVItG.

3 Results and discussion

Here we present the solutions for the minimization problems described in the previous section. We assume that $U_1 = U_2, D_1 = D_2, \rho_1 = \rho_2, a_1 = a_2$, and $c_1^* = c_2^*$. This type of relationship is called homogamous in the literature of marital psychology [17]; they are formed by individuals who are similar to each other. This is a natural assumption given that the data set [15] is obtained by collecting answers from one of the partners in each relationship.

We also assume that the benefit and cost functions are defined by

$$U_i(x) = 5\ln(x+1);\ D_i(c_i) = \frac{1}{2}(c_i - c_i^*)^2,\ \text{for } i = 1, 2.$$

These functions satisfy all the mathematical properties derived by psychological principles -see [12] for the details. This particular choice was first considered by [12] and it was also adopted in [8] and [9].

To find the minimizers $\hat{P}_d$ and $\hat{P}_s$ of the deterministic and stochastic approximation problems above we run an extensive routine to compute the mean squared error between the data $x_{data}(t)$ and the outputs $d-\text{RaBVItG}[P]$ and $s-\text{RaBVItG}[P]$ for each feasible set P. We restricted the search for $r \times a_i \times c_i^* \times \sigma$ within the parameter domains

$$r \in [0.01, 0.3],\ a_i \in [0.01, 0.07],\ \sigma(\cdot) \equiv \sigma \in [0, 1],\ c_i^* \in [0, 10].$$

These domains have been chosen after running several preliminary numerical tests, given that $x(t) \in X = [1, 7]$ for $t \in [0, 30]$

The results obtained for the approximation problems are given in Table 1. The rate of decay of the feeling r, the effort coefficients a_i, and the effort levels c_i^*, $i = 1, 2$, are similar for the deterministic and stochastic solutions. Also, the error is similar in both cases. As a measure of goodness of fit, the last column shows the linear correlation of the observed data $x_{data}(t)$ with the deterministic approximation $x_d^{\heartsuit}(t)$ and with the mean of the stochastic approximation $x_s^{\heartsuit}(t)$. The correlation values show a remarkable agreement between the data and the model approximations both deterministic and stochastic. The linear correlation between the deterministic and stochastic (mean) approximations is 0.9988. This good agreement can be observed in the first panel of graphs in Figure 2, where both data and approximation curves are displayed. Our procedure also provides the effort control policies, deterministic $c_{d,i}^{\heartsuit}(t)$ and stochastic $c_{s,i}^{\heartsuit}(t)$ which, our study, are the same for both partners in each case. They are represented as the extra effort relative to the preferred effort level c_i^* (in percentage), that is, $\frac{c_{d,i}^{\heartsuit}(t) - c_i^*}{c_i^*} \times 100$ y $\frac{c_{s,i}^{\heartsuit}(t) - c_i^*}{c_i^*} \times 100$. These (extra) effort curves are plotted below the feeling curves in Figure 2. In both cases, the effort levels increase over time with respect to the reference value c_i^* until reaching a plateau. This is a typical pattern of effort curves in our differential game models, already found in the numerical analysis, both deterministic [8] and stochastic [9].

Fig. 2: Successful romantic relationships as differential games. The first panel shows the actual data $x_{data}(t)$ of marital quality (the target trajectory) together with the best approximations –deterministic (left) and stochastic (right)– obtained by the differential game models. These feeling trajectories are the solutions of the model obtained by the RaBVItG algorithms for the set of input parameters in Table 1. The approximation is obtained by minimizing the mean square error. The second panel shows the corresponding effort policies of both partners –deterministic (left) and stochastic (right)– that control the dynamics of the feeling. They are represented as the extra effort relative to the estimated level c_i^* (in percentage). Extra effort is required to maintain a relationship successful in the long term.

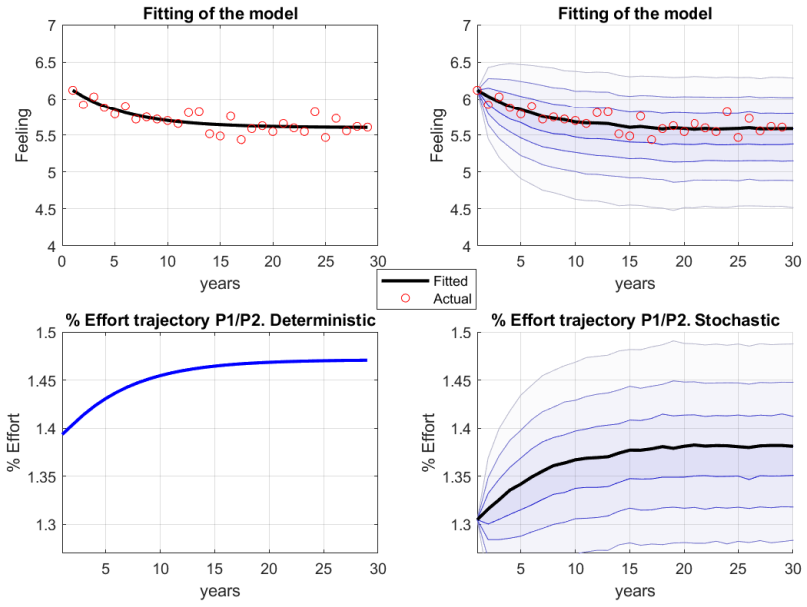


Table 1: Parametric estimates of the best model approximation for x_{data}, along with mean square error (MSE) and linear correlation (*corr*).

	r	$a_1 = a_2$	$c_1^* = c_2^*$	σ	MSE	*corr*
Deterministic	0.182	0.053	9.40	0	0.5129	0.8122
Stochastic	0.179	0.050	9.80	0.50	0.5135	0.8032

4 Conclusions

The computational differential game models introduced in [8] and [9] produce long-term feeling trajectories of successful romantic relationships. Since the models build on well-known psychological assumptions, these synthetic trajectories may serve as a proxy for observed variables like marital satisfaction or marital quality. The numerical analysis of these models is useful to explore significant questions of the dynamics of romantic relationships, if only qualitatively.

To address issues relevant to the proper functioning of real relationships, the models must be calibrated, that is, find a parametrization that replicates the observed data on marital satisfaction. In this note, we show that the models can be calibrated to replicate an actual trajectory of marital satisfaction data over 30 years. This target trajectory is extracted from a data set of a recent cross-cultural study, with 7,000 subjects from 33 countries [15]. The calibrated trajectories –either deterministic or stochastic– show a remarkable correlation with the target trajectory of data.

Once the model is adjusted to replicate real data, many quantitative explorations of the functioning and evolution of a real relationship can be carried out. For example, a main variable of the model, namely the effort required to maintain a long-lasting happy relationship can be estimated with our computational approach. The happiness that both partners derive from their romantic relationship can also be estimated and compared with other types of relationships. In addition, the study in this note on homogamous couples –formed by similar partners– can be used to analyze the effect of dissimilarity –heterogamy– on the relationship dynamics, for instance, the response in effort making or the happiness of different partners. The impact of an external shock on satisfaction or effort policies, as well as the risk of rupture and probability of recovery, can also be estimated by the computational model using real data. These and other intriguing questions stimulate our work in progress.

References

1. Amato, P.R., James, S.L.: Changes in spousal relationships over the marital life course. Alwin, Duane F.; Felmlee, Diane H.; and Kreager, Derek A. Editors (2018) Social Networks and the Life Course. Integrating the Development of Human Lives and Social Relational Networks pp. 139–158 (2018)
2. Carr, D.: The linked lives principle in life course studies: Classic approaches and contemporary advances. In Alwin, Duane F.; Felmlee, Diane H.; and Kreager, Derek A. Editors (2018) Social Networks and the Life Course. Integrating the Development of Human Lives and Social Relational Networks pp. 41–63 (2018)
3. Coontz, S.: Marriage, a history. New York, Viking (2005)
4. Finkel, E.J.: The All-or-Nothing Marriage. How the best marriages work. Penguin Random House (2017)
5. Gottman, J.M., Murray, J.D., Swanson, C.C., Tyson, R., Swanson, K.R.: The mathematics of marriage: Dynamic nonlinear models. MIT Press (2005)

6. Goudon, T., Lafitte, P.: The lovebirds problem: why solve hamilton-jacobi-bellman equations matters in love affairs. Acta Applicandae Mathematicae 136(1), 147–165 (2015)
7. Herrera, J., Ivorra, B., Ramos, Á.M.: An algorithm for solving a class of multi-player feedback-nash differential games. Mathematical Problems in Engineering 2019 (2019)
8. Herrera, J., Rey, J.M.: Controlling forever love. PloS one 16(12), e0260529 (2021)
9. Herrera, J., Rey, J.M.: A computational stochastic dynamic model to assess the risk of breakup in a romantic relationship. Mathematical Methods in the Applied Sciences pp. 1–18 (2023)
10. Kushner, H., Dupuis, P.G.: Numerical methods for stochastic control problems in continuous time, vol. 24. Springer Science & Business Media (2013)
11. Nichols, C., Schumm, W.R., Schectman, K., Grigsby, C.C.: Characteristics of responses to the kansas marital satisfaction scale by a sample of 84 married mothers. Psychological Reports 53(2), 567–572 (1983)
12. Rey, J.M.: A mathematical model of sentimental dynamics accounting for marital dissolution. PloS one 5(3), e9881 (2010)
13. Rey, J.M.: Sentimental equilibria with optimal control. Mathematical and Computer Modelling 57(7-8), 1965–1969 (2013)
14. Schumm, W., Nichols, C., Schectman, K., Grigsby, C., Schumm, W., Paff-Bergen, L., Hatch, R., Obiorah, F., Copeland, J.: Kansas marital satisfaction scale (kms). Journal of Family Therapy 29, 307–324 (2000)
15. Sorokowski, P., et al.: Marital satisfaction, sex, age, marriage duration, religion, number of children, economic status, education, and collectivistic values: Data from 33 countries. Frontiers in Psychology 8 (2017), https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01199
16. Van Laningham, J., Johnson, D.R., Amato, P.: Marital happiness, marital duration, and the u-shaped curve: Evidence from a five wave panel study. Social Forces 79(4), 1313–1341 (2001)
17. Whyte, M.K.: Dating, mating, and marriage. Routledge (2018)

Prediction of Violence Risk Levels: Simulated Statistical Model

Leal-Enríquez E.[1] and Gutierréz-Antúnez A.R.[2]

[1] Insituto Politécnico Nacional. Av. Luis Enrique Erro S/N, Unidad Profesional Adolfo López Mateos, Zacatenco, Alcaldía Gustavo A. Madero, C.P. 07738, Ciudad de México, México,
eleale@ipn.mx

[2] Universidad del Valle de México, Calz. de Tlalpan 3016/3058, Coapa, Ex-Hacienda Coapa, Ciudad de México, México,
aime.gutierrezan@uvmnet.edu

Abstract. In this work, statistics of the maximum and minimum limits that a victim of domestic violence can suffer in a cycle of violence at the hands of his perpetrator are estimated. The estimates are obtained from a mathematical model proposed in the literature that takes a violence risk questionnaire. The results are shown in Tables of simple descriptive statistics, as well as in storyboards where the probable scenarios of violence that the victim could experience over twelve months are graphically shown. These results could serve as monitoring tables (traffic lights) to researchers and experts in the area of violence in order to establish probable protocols and decision-making to help victims of domestic violence.

Keywords: domestic violence, risk levels, statistics, questionnaire

1 Intimate Partner Violence

Violence ("violence is the intentional use of threatened or actual physical force or power, against oneself, another person, a community or a group, that may result in a high likelihood of psychological harm, death, injury, deprivation or maldevelopment" [1]) is a major public health problem that afflicts all societies worldwide [2]. Specifically, violence against women is of significant importance to international organizations and leading human rights groups (violence against women is defined by the United Nations as "any act of gender-based abuse that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life" [3]) [4]. Violence against women does not respect social class, race, age or religious beliefs [5]. Statistics show that, in 75% of the cases, a man is identified as the perpetrator of such violent acts and a woman as the victim. In most countries, programs to monitor cases of violence and to assist the victim have been established. These programs are estimated to last between one and two years [4].

The monitoring programs begin with the application of a violence risk questionnaire. This questionnaire is applied by the violence experts (social workers

from the domestic violence care center), in which the initial violence risk level of the victim taking into account four dimensions: psychological violence, physical violence, severe physical violence and sexual violence [6]. The way to obtain this measurement is through a linear equation (see for example [7] y [8].); however, only a small number of victims seek help, advice, emotional support and information to find a solution to the cycle of violence they are experiencing, for example only the 21% [9])) of adolescents immersed in domestic violence seek help.

With these premises and in order to help experts in this field make decisions based on predictive data on risk levels, this article provides simple statistical tables of risk levels by dimension of violence (psychological / physical / sexual), as well as storyboards with which they are intended to provide quantitative and qualitative tools in cycles of domestic violence [10]).

For this, in this article the work is divided into the following sections:

- Questionnaire: an example of a risk questionnaire is provided with the four dimensions of violence with their severity weights for each question and obtaining their initial conditions of violence by dimension.
- Statistical simulation model: simple statistical risk prediction tables are generated for two cycles of violence, taking as premises the initial conditions of the risk questionnaire.
- Discussion and Conclusions: violence prediction storyboards based on statistical simulation models and a discussion of the use of violence prediction tables are shown in this article.

2 Questionnaire

The first part to obtain the level of risk of the victim in a cycle of violence is the application of a questionnaire that evaluated the dimensions of domestic violence (for this article a questionnaire of 18 questions selected from the instruments of "Spouse Abuse (ISA) "and "Severity of Violence Against Women Scale (SVAWS)" [11], [12]), psychological violence (Dimension I), physical violence (Dimension II), severe physical violence(Dimension III) y sexual violence (Dimension IV).This questionnaire is shown in Table 1, which takes as parameters: the severity weight ω_n that is assigned to each question (violent actions that the perpetrator may have whose method of assigning these weights is by expert judgment [6] [7]) and the frequency of violent acts carried out by the perpetrator towards his victim f_n (the frequency values are discrete assigned as follows: 0=never, 1= sometimes, 2=several times and 3=many times).

2.1 Measurement of the initial condition of violence

After that, both frequency and weight must be assigned to each one of the items f_n and ω_n, consequently, the level of violence Ω_n can be calculated by using the following equation [13] [14]:

Table 1: Violence Risk Questionnaire

n	Question	Weight			$\Omega_n(0)$	
		ω_n	f_n 0	f_n 1	f_n 2	f_n 3
1	Has he ever told you that you are not attractive or that you are ugly?	4.5	0	4.5	9	13.5
2	Has he ever displayed jealousy towards you or become suspicious of your friends?	4	0	4	8	12
3	Has he ever rejected you?	5	0	5	10	15
4	Has he ever offended you?	4	0	4	8	12
5	Has he ever made you feel worthless in front of other people?	5.5	0	5.5	11	16.5
	Indicator of psychological violence	$\Omega^I(0)$	0	23	46	69
6	Has he ever kicked you?	8	0	8	16	24
7	Has he ever pushed you intentionally?	5	0	5	10	15
8	Has he ever beaten you or slapped you on your face?	7	0	7	14	21
9	Has he ever twisted your arm?	6.5	0	6.5	13	19.5
10	Has he ever pulled you forcefully?	5	0	5	10	15
	Indicator of physical violence	$\Omega^{II}(0)$	0	31.5	63	94.5
11	Has he extinguished a cigarette on your body or burned you with any other item or substance?	6	0	6	12	18
12	Has he ever threatened you with a gun or any other type of firearm?	6.5	0	6.5	13	19.5
13	Has he ever shot at you with a gun or any other type of firearm?	9.5	0	9.5	19	28.5
14	Has he ever threatened you with a knife?	7	0	7	14	21
15	Has he ever tried to drown you or suffocate you?	9.5	0	9.5	19	28.5
	Indicator of severe physical violence	$\Omega^{III}(0)$	0	38.5	77	115.5
16	Has he ever forced you to engage in sexual intercourse?	6	0	6	12	18
17	Has he ever used physical force to have sex?	9	0	9	18	27
18	Has he ever threatened you with leaving you for other women if you do not agree to engage in sexual intercourse?	4	0	4	8	12
	Indicator of sexual violence	$\Omega^{IV}(0)$	0	19	38	57
	Initial condition of global violence	$\Omega(0)$	0	112	224	336

.

$$\Omega_n\left(0\right) = \omega_n\left(0\right) \times f_n(0), \tag{1}$$

where $f_n(0) = 0, 1, 2, 3$ and $\omega_n(0)$ are the frequency of violent acts and the weights of severity associated with each question in the questionnaire see Table 1. For example, in the event that a perpetrator has offended his victim many

times ($f_n = 3$)(see question $n = 4$), their associated severity weight is $\omega_4 = 4$), therefore from (1) we obtain that the initial condition of violence is [6]:

$$\Omega_4^I(0) = \omega_4(0) \times f_4(0) = 4 \times 3 = 12. \tag{2}$$

Note, in equation (2) the dimension of violence is set as a superscript. Therefore, the calculation of the level of initial violence of the victim $\Omega(0)$, considering factors I, II, III and IV (where the corresponding dimension has already been assigned to every single question a factor analysis to identify the grouping of the variables that best explain each of the dimensions is usually conducted, obtaining the relevant factors for each type of violence and also assigning a dimension to each question [6].), is calculated by using the equation [6] [13]:

$$\begin{aligned}
\Omega^I(0) &= \sum_1^p \omega_n(0) f_n(0) \\
\Omega^{II}(0) &= \sum_1^q \omega_n(0) f_n(0) \\
\Omega^{III}(0) &= \sum_1^r \omega_n(0) f_n(0) \\
\Omega^{IV}(0) &= \sum_1^n \omega_n(0) f_n(0).
\end{aligned} \tag{3}$$

In equation (3) each of the four factors contained in the questionnaire to be given to the victims is separated. Limits p, q, r and n are the questions to be included to calculate the initial condition of violence by dimension (see Table 1). As an example, the following is noticed in in the questionnaire displayed in Table 1 by dimension of violence and frequencies of violent acts.

2.2 Mathematical Model for Predicting Violence

After the violence questionnaire has been applied a prediction model can be applied for the victim during the following twelve months by cycles of violence. The initial condition ($\Omega(0)$ (see Table 1) is the starting point (see equations (1)-(3)).

For this to be achieved, it is advisable to use the model proposed by Leal-Enríquez E. [13]. This model is composed as follows:

$$\Omega(k) = \Omega(0)\alpha(k) = \sum_1^n \omega_n(k) f_n(k) + \Theta(\omega_p, f_p) \tag{4}$$

where

$$\Theta(\omega_p, f_p) = \sum_{p=n+1}^{m} \omega_p(k) f_p(k) \tag{5}$$

$$\alpha(k) = \sum_{1}^{k} \beta(k), \tag{6}$$

$$\beta(k) = \xi(k) \times \sigma_-(k), \tag{7}$$

$$\left[\sigma_+(k)\ \sigma_-(k)\right] = \left[\sigma_+(0)\ \sigma_-(0)\right] \begin{bmatrix} (1-\lambda) & \lambda \\ \mu & (1-\mu) \end{bmatrix}^k. \tag{8}$$

and

$$\Omega(0) \cong \sum_{1}^{n} \omega_n(0) f_n(0) \tag{9}$$

$\Omega(0)$ represents the initial condition of violence that the victim experiences, which may be approximately (9) applying a risk questionnaire including n questions (refer to Table 1 where $n = 18$). $\alpha(k)$ expresses the accumulated level of violence displayed by the perpetrator. $\beta(k)$ constitutes the probable proportion of violence which may cause injuries to the victim. $\xi(k)$ represents the amount of loss of control that the perpetrator displays in a given cycle of violence. $\sigma_-(k)$ y $\sigma_+(k)$ accounts for the respective loss of control of the aggressor and his self-control level. $\Theta(\omega_p, f_p)$ are the p number of questions which were not included in the initial questionnaire applied to the victim when she arrives at the care center for the first time, in order to assess the risk of violence. ω_n and f_n represent the weight of severity and the frequency which is correlated to every single question included in the risk questionnaire (refer to Table 1). 1-λ and 1-μ comprise the respective prevalence shown by the perpetrator when preserving a state of self-control or loss of control.

3 Simulated Statistical Model

From model analysis (4) it can be deduced that the maximum limits of domestic violence by dimension for the violence risk questionnaire (see Table 1) are when the man is violent $\sigma_-(k) = 1$ (see (8)), this is true when the prevalence of out-of-control state violence $(1 - \mu) = 1$ and that the percentage of loss of control of the perpetrator that results in violent acts $\xi(k) = 1$ (see (7)) for all months k, substituting this into the equation (6) we have that the accumulation $\alpha(k)$ of

Table 2: Maximum Statistical Results: violence

Dimension	Initial Condition of Violence, $\Omega(0)$	Frequency, $f_n = 3$	Violence Risk Prediction, $\Omega(12)$
Psychological Violence I	69	3	828
Physical Violence II	94.5	3	1134
Severe Physical Violence III	115.5	3	1386
Sexual Violence IV	57	3	684

violence of the perpetrator for the last month is twelve. In the table 2 shows the maximum statistical results in the event that a victim arrives at the care center and has suffered a violent act from the three-frequency risk questionnaire (see Table 1).

3.1 Simulation: tension-outburst-honeymoon

The following values of $\xi(k)$ are values published in the literature [13] for the tension-explosion-honeymoon cycle for percentages of loss of control of the perpetrator that are reflected in violent acts [10] [13]):

$$\begin{aligned}\xi(k) = [&0.3135 \quad 0.0763 \quad 0.2003 \quad 0.6556\\ &0.9272 \quad 0.8406 \quad 0.6358 \quad 0.3424\\ &0.8803 \quad 0.0450 \quad 0.0619 \quad 0.0794].\end{aligned} \tag{10}$$

In Table 3 are shown the maximum statistical results for the frequency of violent acts 1, 2 y 3.

3.2 Simulation: outburst-honeymoon-tension

The following are the values which are considered to determine the proportion of loss of control by the perpetrator $\xi(k)$. These can result in injuries or violent acts towards the victim during a period of twelve months (data distribution for $\xi(k)$ is determined by considering a cycle of violence de outburst-honeymoon-tension, occurring between a perpetrator of violent acts and the victim [10] [13]):

$$\begin{aligned}\xi(k) = [&1 \quad 0.6 \quad 0.4 \quad 0.7 \quad 0.8 \quad 0.35\\ &0 \quad 0.01 \quad 0.02 \quad 0.2 \quad 0.3 \quad 0.1]\end{aligned} \tag{11}$$

In Table 4 are shown the maximum statistical results for the frequency of violent acts 1, 2 y 3.

Table 3: Results for tension-outburst-honeymoon

Dimension	Initial Condition of Violence, $\Omega(0)$	Frecuency, f_n	Violence Risk Prediction, $\Omega(12)$
Psychological Violence I	23	1	116.3409
Physical Violence II	31.5	1	159.3365
Severe Physical Violence III	38.5	1	194.745
Sexual Violence IV	19	1	96.1077
Psychological Violence I	46	2	232.682
Physical Violence II	63	2	318.673
Severe Physical Violence III	77	2	389.489
Sexual Violence IV	38	2	192.215
Psychological Violence I	69	3	349.023
Physical Violence II	94.5	3	478.009
Severe Physical Violence III	115.5	3	584.234
Sexual Violence IV	57	3	288.323

4 Discussion and Conclusions

Statistical simulations (see Tables 2-4) clearly show that when a man is violent $\sigma_{-}(0) = 1$ (see (8)) and a proportion of that loss of control manifests itself in some specific violent act (see 7 and Table I), all this is reflected in violence towards the victim, whose value is estimated with a risk questionnaire that indicates the initial condition of violence of the victim (9).

This initial condition of violence (see Table I) when substituting it in the mathematical model given in the equations (4)-(9) generates tables of values whose descriptive statistics (see Tables 2-4), which can be interpreted by experts on the subject of violence in victim care centers.

As an example of this, we have Tables II-IV for two cycles of violence that the victim could experience: outburst-explosion-honeymoon and explosion-honeymoon and outburst. From Tables III-IV, it can be clearly seen that the

Table 4: Results for outburst-honeymoon-tension

Dimension	Initial Condition of Violence, $\Omega(0)$	Frecuency, f_n	Violence Risk Prediction, $\Omega(12)$
Psychological Violence I	23	1	103.04
Physical Violence II	31.5	1	141.12
Severe Physical Violence III	38.5	1	172.48
Sexual Violence IV	19	1	85.12
Psychological Violence I	46	2	206.08
Physical Violence II	63	2	282.24
Severe Physical Violence III	77	2	344.96
Sexual Violence IV	38	2	170.24
Psychological Violence I	69	3	309.12
Physical Violence II	94.5	3	423.36
Severe Physical Violence III	115.5	3	517.44
Sexual Violence IV	57	3	255.36

outburst-explosion-honeymoon cycle, for all dimensions of violence, has higher values than the cycle of violence that begins with explosion-honeymoon-tension. This can help experts in violence that if they detect that a woman arrives at the care center in that cycle of violence in the coming months requires help so that she does not reach the maximum level of violence given in Table I, this can be translated mathematically into the following equation:

$$\Omega_{TOH}(k) \geq \Omega_{OHT}(k), \tag{12}$$

where $\Omega_{TOH}(k)$ is the risk of violence for the outburst-explosion-honeymoon cycle and $\Omega_{OHT}(k)$ is the risk of violence for the explosion-honeymoon-tension cycle for $k = 1, 2, \ldots, 12$ months.

Also from Table I, we have that the maximum levels of violence really occur when $\xi(k) = 1$ throughout the cycle of violence, which means that for the victim

not to suffer that maximum level of violence, the perpetrator should also receive help to control his loss of control.

The maximum level of violence that a victim can suffer for a risk questionnaire that is applied is given by (taking into account that the maximum frequency must be taken $f_n = 3$):

$$\Omega(k) = \Omega_{max}(0) \times 12, \tag{13}$$

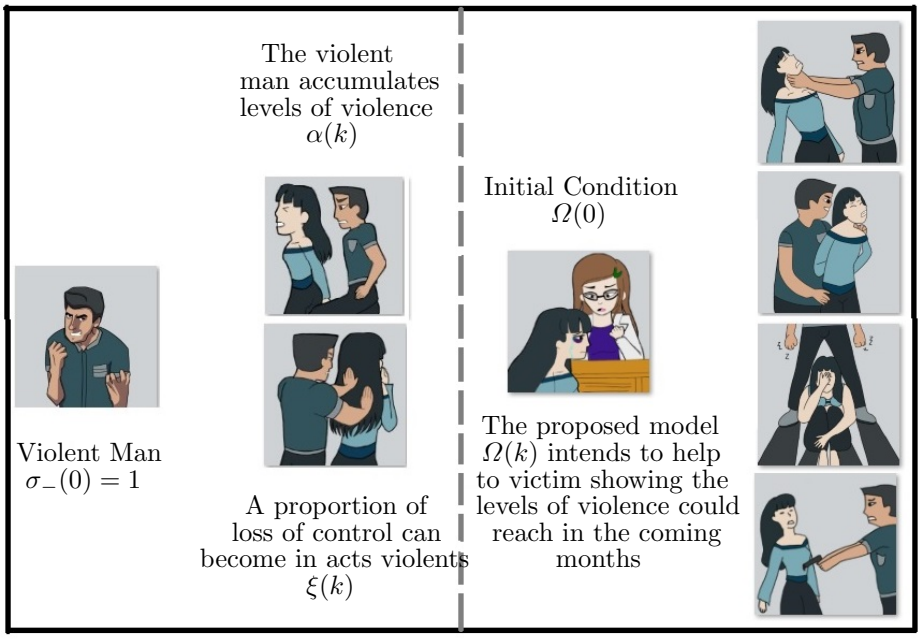
where

$$\Omega_{max}(0) = \sum_{j=I}^{IV} \Omega^j(0) \tag{14}$$

The equations (13) and (14) are valid for the other frequencies, but it must be taken into account that the maximum that would be reached would be for dimension I, II, III and IV (see Table I).

In Figure 1, shows a storyboard where what is presented in this article is graphically observed.

Fig. 1: Storyboard: victim of violence



References

1. World Health Organization (WHO). http://www.who.int/en/
2. Hijar-Medina, M., López-López, M. V., & Blanco-Muñoz, J. La violencia y sus repercusiones en la salud: Reflexiones teóricas y magnitud del problema en México. salud pública de México, 1997; 39, 565-572.
3. UN Women, http://www.unwomen.org/en, 2015
4. Fenlason, T. Change in intimate partner violence: The domestic couple's perspective on perpetrator change (Doctoral dissertation, Antioch University), 2009.
5. Women, D. V. A. Putting Women First. World Health Organization, 2001.
6. Valdez-Santiago, R., Híjar-Medina, M. C., Salgado de Snyder, V. N., Rivera-Rivera, L., Avila-Burgos, L., & Rojas, R. Escala de violencia e índice de severidad: una propuesta metodológica para medir la violencia de pareja en mujeres mexicanas. Salud pública de México, 2006; 48, s221-s231.
7. Zamudio Sánchez, F. J., Andrade Barrera, M. A., Arana Ovalle, R. I., & Alvarado Segura, A. A. (2017). Violencia de género sobre estudiantes universitarios (as). Convergencia, 24(75), 133-157.
8. Papadakaki, M., Tzamalouka, G. S., Chatzifotiou, S., & Chliaoutakis, J. (2009). Seeking for risk factors of intimate partner violence (IPV) in a Greek national sample: The role of self-esteem. Journal of Interpersonal Violence, 24(5), 732-750.
9. Ashley, O. S., & Foshee, V. A. Adolescent help-seeking for dating violence: Prevalence, sociodemographic correlates, and sources of help. Journal of Adolescent Health, 2005; 36(1), 25-31.
10. Gunter, J. Intimate partner violence. Obstetrics and gynecology clinics of North America, 2007; 34(3), 367-388.
11. Hudson, W. W., & McIntosh, S. R. The assessment of spouse abuse: Two quantifiable dimensions. Journal of Marriage and the Family, 1981; 873-888.
12. Marshall, L. . Development of the severity of violence against women scales. Journal of family violence, 1992; 7(2), 103-121.
13. Leal-Enríquez E., & AR, G. A. (2022). Mathematical Model of Predictive Indicators of Violence: Limit of Fatality. Engineering Letters, 30(4).
14. Valdez-Santiago, R., Híjar-Medina, M. C., Salgado de Snyder, V. N., Rivera-Rivera, L., Avila-Burgos, L., & Rojas, R. Escala de violencia e índice de severidad: una propuesta metodológica para medir la violencia de pareja en mujeres mexicanas. Salud Pública de México, 2006; 48, s221-s231.

Models for Hospital Bed Management in an EU University Hospital

Mario Picans[1], Maria Isabel Borrajo[1], Mercedes Conde-Amboage[1] and Francisco Reyes-Santias[2,3,4,5]

[1] University of Santiago de Compostela (USC), Santiago de Compostela, Spain.
[2] University of Vigo, Vigo (Pontevedra) Spain.
[3] Institute for Health Research Foundation (FIDIS), Santiago de Compostela, Spain.
[4] Governance And Economics Research Network (GEN), University of Vigo, Vigo (Pontevedra), Spain.
[5] Center for Biomedical Research Network. Cardiovascular Diseases (CIBERCV), Santiago de Compostela, Spain.

Abstract. Service of Admission and Documentation Clinic del Complex Univer-sity Hospital of Santiago de Compostela, is the unit in charge of the patients managament, whose main purpose it is to organize the patient flow through the healthcare facility, and to process the clinic documentation they generate. With the aim of providing tools for managing the occupancy rate in the hospital units, in this work we will apply different regression models to the data provided by the Fundaci on Instituto de Investigaci on Sanitaria (FIDIS), which contain infor-mation of every admission during the period that goes from January 2016 to No-vember 2021. The point of using the regression models we will adjust during this work is, among others, to get the nature of some variables, such as daily admis-sions or patient's length of stay. The objectives of this dissertation are to conduct a literature review of the methods that are commonly used for modeling this type of response variables, and the set up of different regression models that will al-low us to predict the occupancy rate of a hospital unit, or an admission's length of stay.

Keywords: GAM, Internal Medicine, Inpatiens, Discharges.

1 Introduction

Patients' management is a non-clinical but essential area in any hospital. In this one the Admitting and Medical Record Department oversees this area. Bed management is a crucial task for this department, involving the constant monitoring of admissions and discharges, and the flow of patients within the hospital. Developing an effective tool to predict different demands will improve hospital response capacity deriving in increasing patient wellbeing.

2 Objective

The objectives of this study are to conduct a literature review of the methods that are commonly used for modeling this type of response variables, and the

set up of different regression models that will allow us to predict the occupancy rate of a hospital unit, or an admission's length of stay.

3 Materials and Methods

The data set we have used in this manuscript consist of a census of 188135 admissions in the period 2016 to 2020. For each department of interest, we have built a data base in which, for every day of the time period 2016 to 2020, we obtain the following variables: Day of the week; Month of the year; Number of inpatients; Number of discharges; Number of occupied beds the day before; Median length of stay on patients the day before; Number of inpatients the day before; Number of discharges the day before; Difference between the number of admissions the same weekday the week before and the admissions the day before; Difference between the number of discharges the same weekday the week before and the discharges the day before. To be able to predict bed availability we have divided the problem in three different ones: modelling admissions, and modelling discharges using regression methods, and get those results together to model bed occupancy. Both, admissions, and discharges, are count variables, i.e., their possible values are only non-negative integers. This characteristic prevents them to be used in any linear regression model, since at least two of its basic assumptions (linearity and normal distribution of the error) would never be fulfilled. Therefore, to analyse this type of response variables, more complex regression models need to be used.

To be able to predict bed availability we have divided the problem in three different ones: modelling admissions, and modelling discharges using regression methods, and get those results together to model bed occupancy. Both, admissions, and discharges, are count variables, i.e., their possible values are only non-negative integers. This characteristic prevents them to be used in any linear regression model, since at least two of its basic assumptions (linearity and normal distribution of the error) would never be fulfilled. Therefore, to analyse this type of response variables, more complex regression models need to be used.

Avoiding parametric assumptions for the effect of explanatory variables, we have focussed on the Generalized Additive Models (GAMs), presented by [2]. GAMs blend the properties of GLMs and Additive Models (AMs), which were first introduced by [1]. AMs assume that the relationship between the response variable, Y, and a collection of explanatory variables, $X_1, \ldots, X_p$, can be written as follows:

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{1}$$

where the functions f_j represents the partial effect of the explanatory variable, X_j, on the response variable, Y. These models, despite being more flexible, are still interpretable, because we can graphically represent the relationship between each explanatory variable and the response one. Additionally, categorical

explanatory variables may be included in the model through parametric effects. In that scenario, we can complete (2) by considering:

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + Z\rho + \varepsilon, \tag{2}$$

where Z denotes the design matrix for categorical variables, and ρ the vector of parameters associated with them.

GAMs are an extension of additive models just as GLMs are an extension of linear models. The relationship between the response and the explanatory variables is of the form:

$$g(E(Y|X_1, \ldots, X_p)) = \beta_0 + \sum_{j=1}^{p} f_j(X_j). \tag{3}$$

GAMs will be our choice to model hospital admissions and discharges in each of the departments of interest. In our scenario we suffer from overdispersion, so we will define a GAM with Negative Binomial response for the daily number of admissions, and another one for the daily number of discharges. Once we have these estimates, the number of occupied beds, C, at a specific day, t, can be computed by:

$$\hat{C}_t = C_{t-1} + \hat{A}_t - \hat{D}_t, \tag{4}$$

where C_{t-1} states for the real number of occupied beds on day $t-1$, $\hat{A}_t$ denotes the estimated number of admissions, and $\hat{D}_t$ denotes the estimated number of discharges on day t, obtained from the GAMs. Recall that to estimate the number of occupied beds at time t, we need admissions and discharges estimates at time t. Hence, we will first adjust models for admissions and discharges in each department of interest, and then we will apply (3) to obtain the estimation of the number of occupied beds.

To evaluate the predictions obtained with the different regression models, we define the following error measures all of them based on the prediction errors, $\hat{\varepsilon}_i = \hat{Y}_i - Y_i$ for $i = 1, \ldots, n$: *Mean Absolute Error (MAE); Root-Mean-Square Error (RMSE); Relative Absolute Error (RAE); Relative Squared Error (RSE); Mean Absolute Percentage Error (MAPE).*

4 Results

The procedure presented in Material and Methods can be applied to any department in the hospital. Due to space constrains and not to result overwhelming, as explained before we have chosen the Department of Internal Medicine.

Department of Internal Medicine

Department of Internal Medicine has the highest number of patients, at around 17% of hospital patients are theirs. The GAM fitted for the number of admissions

and discharges in this unit is detailed in Table 1. The expressions result after grouping the non-significant elements of categorical variables, and removing the non-significant effects of continuous variables at a 5% significant level.

First half of Table 1 shows the result for admissions, including the estimated coefficients for the different groups of the categorical variables. So does the second half, but for discharges.

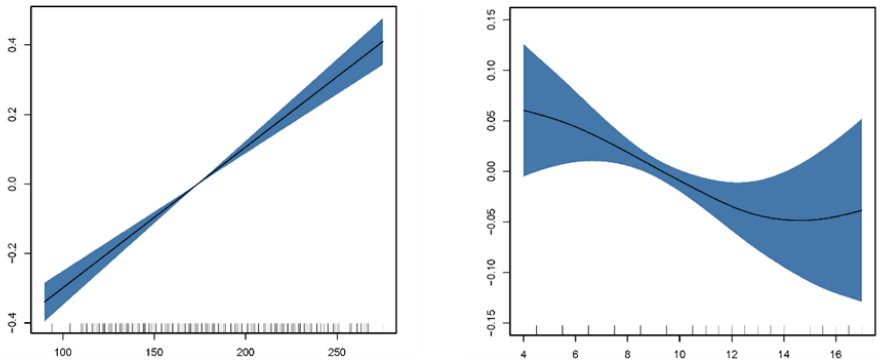
Table 1: Department of Internal Medicine model summary, detailing the adjustment and coefficient estimates for admissions and discharges. In admissions, the reference category for `Day` is (Monday/Wednesday/Friday/Saturday) and for `Month` is (January/February/December); whereas for discharges the reference category for `Day` is (Monday).

Admissions												
$\beta_0 + \boldsymbol{\beta_1}\text{Day} + \boldsymbol{\beta_2}\text{Month} + f_{BDB}(\text{BedsDayBefore}) + f_{ML}(\text{MedianLoS})$												
$\widehat{\beta_0}$ (Std. error)	$\widehat{\beta_{1,Tue}}$ (Std. error)	$\widehat{\beta_{1,Thu}}$ (Std. error)	$\widehat{\beta_{1,Sun}}$ (Std. error)	$\widehat{\beta_{2,Mar}}$ (Std. error)	$\widehat{\beta_{2,Apr}}$ (Std. error)	$\widehat{\beta_{2,May}}$ (Std. error)	$\widehat{\beta_{2,Jun}}$ (Std. error)	$\widehat{\beta_{2,Jul}}$ (Std. error)	$\widehat{\beta_{2,Aug}}$ (Std. error)	$\widehat{\beta_{2,Sep}}$ (Std. error)	$\widehat{\beta_{2,Oct}}$ (Std. error)	$\widehat{\beta_{2,Nov}}$ (Std. error)
2.71 (0.02)	0.12 (0.02)	0.04 (0.02)	-0.22 (0.02)	-0.07 (0.03)	-0.07 (0.03)	-0.11 (0.03)	-0.09 (0.03)	-0.08 (0.03)	-0.12 (0.03)	-0.08 (0.03)	-0.06 (0.03)	-0.07 (0.03)

Discharges						
$\beta_0 + \boldsymbol{\beta_1}\text{Day} + f_{BDB}(\text{BedsDayBefore}) + f_{ML}(\text{MedianLoS}) + f_{DDB}(\text{DischargesDayBefore}) + f_{DG}(\text{DischargesGradient})$						
$\widehat{\beta_0}$ (Std. error)	$\widehat{\beta_{1,Tue}}$ (Std. error)	$\widehat{\beta_{1,Wed}}$ (Std. error)	$\widehat{\beta_{1,Thu}}$ (Std. error)	$\widehat{\beta_{1,Fri}}$ (Std. error)	$\widehat{\beta_{1,Sat}}$ (Std. error)	$\widehat{\beta_{1,Sun}}$ (Std. error)
2.61 (0.04)	0.16 (0.05)	0.18 (0.06)	0.21 (0.06)	0.63 (0.07)	-0.85 (0.07)	-1.49 (0.05)

We can see in Table 1 (admissions part) that the estimate of the intercept is 2.71 (exponential 15.07), which means that for a Monday in January (or any other day and month in the reference categories) with no occupied beds the previous day, and with a median length of stay of 0 days, 15.07 admissions are expected in the Department of Internal Medicine.

Partial effects associated with the continuous significant variables BedsDayBefore and MedianLoS are shown in Figure 1. The effect of BedsDayBefore is negative when the number of occupied beds in the department the day before is low. This can imply that low occupancy in the department, relates to fewer expected admissions, while when the department is saturated, a greater number of admissions is expected. The effect associated to MedianLoS presents a different behaviour; it is positive for shorter LoS, which we can be understood as shorter stays in the previous days relate to higher expected admissions; and it decreases below zero as LoS increases, that is, fewer admissions are expected when longer stays take place.



(a) Beds day before. (b) Median LoS.

Fig. 1: Estimated partial effect of the continuous significant variables (black line) with 95% confidence bands (blue shading) in the admissions model for the Department of Internal Medicine.

5 Conclusions

With regard to the esteem of the efficiency of the different categories of the qualitative explanatory variables, the great influence of the inpatient of urgent character in the harshness of the stay in the Internal Medicine Service.

On the other hand, is highlighted, the estimated key associated with women's and negative income in the service of the study case, implying that slightly but short stays are expected for them than for men.

An intense seasonal component was appreciated in the number of discharges than in inpatients,

References

1. Friedman, J. H. & Stuetzle, W. Projection Pursuit Regression. *Journal of the American Statistical Association*, **1981** 76(376), 817-823. DOI: 10.1080/01621459.1981.10477729
2. Hastie, T. J., & Tibshirani, R. J. Generalized Additive Models. *Chapman & Hall.* **1990**. DOI:10.1201/9780203753781-6
3. Nelder, J. A., & Wedderburn, R. W. Generalized Linear Models. *Journal of the Royal StatisticalSociety: Series A*. **1972**, 135(3), 370-384. DOI:10.2307/2344614

Time series analysis for the COMEX platinum spot price foretelling by using models based on SVM, MARS, MLP, VARMA and ARIMA: A case study

Luis Alfonso Menéndez-García, Paulino José García-Nieto, Esperanza García-Gonzalo and Fernando Sánchez Lasheras

Department of Mathematics, Faculty of Sciences, University of Oviedo, C/ Leopoldo Calvo Sotelo 18, 33007 Oviedo, Spain
sanchezfernando@uniovi.es

Abstract. This article examines the forecasting of platinum prices using time series and machine learning models in conjunction with 12 other commodity prices. Platinum, a rare and remarkable chemical element, has significant industrial and artistic value. This research contributes to econometrics by demonstrating the applicability of advanced modelling techniques in predicting precious metal prices, and provides valuable implications for the field. The price of platinum is volatile, but it is considered an important indicator of the global economy. Changes in the price of platinum indicate higher global growth or an impending recession. This paper investigates the forecasting of platinum spot prices from the New York Commodity Exchange using several time series machine learning (MARS, SVM and MLP) and classical techniques (ARIMA and VARMA). Among the models considered, the Artificial Neural Network (MLP) shows the highest predictive accuracy with an RMSE of 9.24. The ARIMA time series model performs the worst with an RMSE of 74.94. The superior performance of the MLP method indicates its ability to capture complex relationships between platinum and other commodities. The study demonstrates the potential of machine learning techniques, particularly MLP, for accurate platinum price forecasting, benefiting investors, industry professionals and policy makers.

Keywords: Time series analysis; Multivariate adaptive regression splines (MARS); Support vector machines (SVMs); Artificial neural networks (ANNs); Vector autoregressive moving-average (VARMA); Autoregressive integrated moving-average (ARIMA); Platinum price forecasting

1 Introduction

Platinum, symbol Pt and atomic number 78, is a rare and remarkable chemical element similar to gold as a transition metal and belongs to group 10 of the periodic table [1]. It is a dense, lustrous, silvery-white metal with exceptional physical properties such as high malleability and ductility, and resistance to

corrosion and tarnishing [1]. Platinum occurs mainly in the form of small grains or nuggets, often in association with other precious metals such as palladium, rhodium and iridium [2]. Most platinum production is a by-product of nickel and copper mining. South Africa is the largest producer of platinum, followed by Russia and Zimbabwe [3]. Platinum's rarity and beauty have made it a highly prized metal throughout human history and plays a vital role in many industrial applications [4]. Demand for platinum comes from a variety of sectors: about 60% is used in automotive catalytic converters, 25% in jewellery, oil refining and the production of electronic components [5]. In addition, platinum's resistance to high temperatures, corrosion resistant, and chemical reactions make it an integral part of laboratory equipment, electrical contacts and thermocouples [6]. In the medical field, platinum's biocompatibility is used in a number of applications, such as platinum-based drugs, which have proven to be highly effective in treating certain types of cancer [7].

Platinum is also one of the most important metal commodities actively traded on major physical futures exchanges such as the London Metal Exchange (LME), the New York Commodity Exchange (COMEX) and the Shanghai Futures Exchange (SHFE). Its value depends on a delicate balance of supply, demand and market dynamics and is considered a key indicator of the global economy [5].

The analysis and prediction of commodity prices has been studied extensively in previous research. For example, Mingming and Jinliang [8] used a Multiple Wavelet Recurrent Neural Network (MWRNN) model to forecast the price of crude oil, taking into account the price of gold. Similarly, many studies have focused on precious metals, particularly gold. In a recent paper, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Linear Regression (LR) methods were used to predict whether the price of gold would rise or fall in the near future [9]. Other studies used classical time series methods such as Autoregressive Integrated Moving Average (ARIMA) to predict gold prices [10]. Another study attempted to predict the price of platinum, along with other metals such as gold, silver, copper and palladium using various methods such as LR [11].

The aim of this study is to evaluate and compare the predictive accuracy of different methods for forecasting the monthly spot price of platinum in the COMEX series. Specifically, two approaches are used: (a) traditional time series techniques such as Autoregressive Integrated Moving Average (ARIMA) and Vector Autoregressive Moving Average (VARMA), and (b) machine learning models including Multivariate Adaptive Regression Spline (MARS), Support Vector Machine (SVM), and Multilayer Perceptron Neural Network (MLP-NN). The forecasting horizon is up to 18 months, taking into account the price dynamics of 12 variables from the COMEX series, such as energy commodities, food and metals. The objective is to obtain accurate predictions for platinum prices based on the interaction of these various variables over the specified time frame.

2 Materials and methods

2.1 Experimental dataset

The data is taken from the World Bank's Commodity Price Data, specifically the Pink Sheet CMO Historical Data Monthly, covering a total of 71 commodities. The data set covers the period from 1960 to today. The training dataset used includes monthly spot closing price data from January 1990 to August 2021, while the monthly predicted prices cover the period from September 2021 to February 2023. There are no missing values and the training and forecast datasets are complete for the variables and time periods indicated.

The analysis incorporates 12 predictor variables, categorised into four groups: energy commodities - Brent crude oil, US natural gas and European natural gas; food commodities - Arabica coffee, tea, orange, beef and sugar; raw materials - cotton; and metals - aluminium, nickel and silver. These predictor variables are used in the modelling process to derive accurate forecasts for platinum prices over the specified forecast horizon.

2.2 Autoregressive integrated moving average (ARIMA) model

Autoregressive models (AR) are based on the idea that the current values of series, x_t, can be explained as a function of certain number p of previous values x_{t-p}. The autoregressive model of order p, called $AR(p)$, can be expressed as $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + w_t$, where x_t is stationary, $\phi_1, \phi_2, \ldots, \phi_p$ are constants with values different from zero and w_t is a Gaussian white noise with ${\sigma_w}^2 = 1$. The autoregressive model $AR(p)$ is usually denoted as $\left(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p\right) x_t = w_t$ [12] [13].

The Moving average model of order q, called $MA(q)$, assumes that the white noise wt is combined linearly to form the data. The MA(q) model is defined to be $x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}$, where there are q lags in the moving average and $\theta_1, \theta_2, \ldots, \theta_q$ with $\theta_q \neq 0$. The moving average $MA(q)$ model is usually denoted as: $x_t = \theta(B) w_t$ where $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$ [12, 13].

A time series can be considered as $ARMA(p, q)$ if it is stationary $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}$. The parameters p and q are called the autoregressive and moving average orders, respectively [12]. To fix ideas, the $ARIMA$ model is a broadening of the class of ARMA models to include differencing [13] [14]. Hence, a process x_t is said to be $ARIMA(p,\ d,\ q)$ if $\nabla^d x_t = (1 - B)^d x_t$.

2.3 Vector autoregressive moving–average (VARMA) model

VARMA is an extension for the multivariate case of ARMA as it is considered to be a very flexible model for describing relationships among variables. The VARMA(p, q) model is specified according to the following expression [15]: $y_t = c + \delta t + \sum_{i=1}^{p} \Phi_i y_{t-i} + \sum_{j=1}^{q} \Theta_j \varepsilon_{t-j} + \varepsilon_t$, where y_t is the vector of response time

series variables at time t of length n and c is a constant vector of offsets, with n elements. Φ_i are the $n \times n$ autoregressive matrices. There are p autoregressive matrices, and some can be entirely composed of zeros. ε_t is a vector of serially uncorrelated innovations, vectors of length n. The ε_t are multivariate normal random vectors with a covariance matrix Σ. Θ_j are the $n \times n$ moving average matrices. Finally, δ is a constant vector of linear time trend coefficients with n elements. Please note that it is possible to build VARMA models with either $p = 0 or q = 0$.

2.4 Support vector regression (SVR) for time series analysis

Given a set of time series data, a training set consisting of a continuous dependent variable $y_i \in \mathcal{R}$, $\forall i = 1, 2, \ldots, m$ and covariates $x_i \in \mathcal{R}^p, \forall i = 1, 2, \ldots, m$ can be constructed by taking p lags of y_i. The method $\varepsilon - SVR$ constructs a function $f(x) = w^T x + b$, $w \in \mathcal{R}^n$, $b \in \mathcal{R}$ that has at most a deviation of ε from y_i for all training instances x_i, and at the same time is as flat as possible [16]. The radial basis function (RBF) kernel $K(x_i, x_j) = e^{\|-\sigma x_i - x_j\|^2}$ is chosen due to its superior performance [17].

2.5 Multilayer perceptron neural network (MLP NN) model

The MLP utilizes the function $f : X \subset \mathcal{R}^d \longrightarrow Y \subset \mathcal{R}^c$, which can be described by [18], [19] $f(x) = \phi(\psi(x)) = (\phi \circ \psi)(x)$, $\phi : X \subset \mathcal{R}^d \longrightarrow U \subset \mathcal{R}^h, \psi : U \subset \mathcal{R}^h \longrightarrow Y \subset \mathcal{R}^c$, where U is the hidden variables' space, referred to as the characteristics space. Taking into account the MLP architecture, it can be established that: $\psi_j(x) = \psi\left({w_j}^T x + w_{j_0}\right)$, where ψ is the activation function of the hidden layer's neurons, $w_j \in \mathcal{R}^d$ is the parameter vector of the distinct neurons and $w_{j_0} \in \mathcal{R}$ is a limit value known as the threshold value. The two usual activation functions ψ are both sigmoid functions: hyperbolic tangent and logistic function. $\phi_j(u) = \phi\left({c_j}^T u + c_{j_0}\right)$, where ϕ is the activation function of the output layer's neurons, $c_j \in \mathcal{R}^h$ is the weight vector of the neurons and $c_{j_0} \in \mathcal{R}$ is a limit value known as the threshold value. In practice, ϕ is normally the identity transformation, Heaviside step function or a dichotomy function.

Finally, the function used by the MLP NN is formulated as follows $f(x) = \sum_{j=1}^{h} c_j \psi\left({w_j}^T x + w_{j_0}\right) + c_0$.

2.6 Multivariate adaptive regression splines method (MARS)

Multivariate adaptive regression splines (MARS) is a multivariate nonparametric classification and regression technique [18]. MARS defines the functional relationship between the dependent and the independent variables by means of a group of coefficients and piecewise-defined polynomials, also called splines, of degree q (basis functions) that are entirely "driven" from the regression data [18], [20]. The MARS regression model is constructed by fitting basis functions to different intervals of the independent variables. Generally, splines have

pieces smoothly connected together to describe the behavior of the dependent variable. The degree q of the splines is usually selected by achieving a compromise between performance and complexity of the model.

In general, any MARS model makes use of the following model: $f(x) = \sum_{i=1}^{k} c_i \cdot B_i(x)$, where $B_i(x) = \begin{cases} x \ \ if \ x \geq 0 \\ 0 \text{ otherwise} \end{cases}$ represents the basis functions of the model and c_i are constant coefficients.

In order to prune and obtain the definitive MARS model, a two steps process is performed. Firstly, a progressive selection of basis functions leads to a very complex and overfitted model. Such model, although is able to fit the data, has poor predictive ability for new objects. To improve prediction, redundant basis functions are removed one at a time using a regression procedure. To determine which basis functions will be included in the model, the generalized cross validation (GCV) methodology is employed. In this methodology, the root mean squared residual error is divided by a penalty parameter, which depends on the complexity of the model.

The GCV equation is $GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_M(x_i))^2}{\left(1-\frac{M+1+d\cdot M}{n}\right)^2}$, where M is the number of basis functions, and the parameter d is a penalty for each base function included in the model [21]. Once the MARS model is constructed, the relevance of the explanatory variables can be expressed as their contribution to the goodness of fit of the model. To determine variable importance scores, MARS calculates how much the goodness of fit is reduced when eliminating each variable

2.7 The goodness-of-fit of this approach

The goodness of fit of the regression models was evaluated to assess the accuracy of the model, taking into account the discrepancies between observed and predicted values. The root mean square error (RMSE) was the primary criterion used to evaluate the models [22], [23]. This statistic is used frequently to evaluate the forecasting capability of a mathematical model. Indeed, if the observed values are y_i, the corresponding modelled values are $\hat{y}_i$, n the number of samples (or different predictions), then the RMSE is given by the formula $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$ [22], [23].

If RMSE has a value of zero, it means that there is no difference between the predicted and observed data.

3 Results and discussion

The dataset used in this study covers the period from January 1990 to February 2023, and the forecast of monthly platinum prices covers the period from September 2020 to February 2023, with data from January 1990 to August 2020 used for model training.

Feature selection is a critical step in the modelling process, primarily aimed at identifying and removing non-informative or redundant predictors from the model. The variables selected for inclusion in the multivariate models are Brent crude oil, US natural gas, European natural gas, Arabica coffee, tea, orange, beef, sugar, cotton, aluminum, nickel and silver.

To improve model generalization and reduce the risk of over-fitting, a $k-$fold cross-validation approach with $k = 10$ was used during model development. This cross-validation technique ensures that the performance of the models is evaluated on multiple subsets of the data, promoting a more robust assessment of their predictive capabilities and reducing the influence of data peculiarities.

The artificial neural network model provided the most accurate approximation with a root mean square error (RMSE) of 9.24, followed by the e-SVM model with an RMSE of 20.05. In comparison, the VARMA multivariate time series model outperformed the MARS model with RMSE values of 40.78 and 58.48 respectively. However, the ARIMA time series model, the only univariate technique used, showed the worst performance with an RMSE of 74.94.

Table 1: Summary of performance.

	ARIMA	VARMA	SVM	MARS	MLP NN
RMSE	74.945	40.781	20.050	58.481	9.236

Table 1 presents the detailed results obtained for each of the models, while Fig. 1 shows the predictions for the 18-month period, contrasting the values obtained by each method with the actual platinum prices.

The evolution of platinum prices over the study period from 1990 to 2023 has been shaped by various factors that have influenced its market behaviour, including economic crises, political and socio-economic developments, energy factors and trade relations. Economic crises, such as the global financial crisis of 2008, have had an impact on platinum prices. In times of economic uncertainty, investors tend to seek safe-haven assets, including precious metals such as platinum [5].

During the 18-month period used to forecast and compare the results of various time series and machine learning models, the global economy experienced stagnation due to the COVID-19 pandemic, resulting in the automotive sector being hampered by chip shortages [24]. The platinum price was particularly sensitive to this situation, as its use in catalytic converters is closely linked to the automotive industry.

In addition, during this period, significant political and socio-economic events in major platinum producing countries such as South Africa and Russia could have an impact on platinum supply and production.

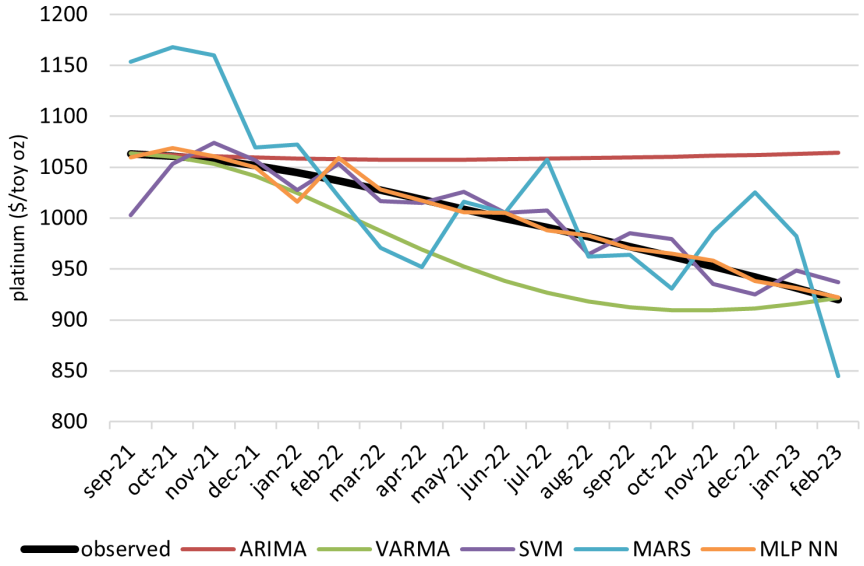


Fig. 1: Actual platinum prices and forecasts from ARIMA, VARMA, SVM, MARS, MLP NN models for the 18-month study period from September 2021 to February 2023.

Based on the observed results, it is clear that the machine learning models outperform the time series models, achieving better approximation and predictive performance in forecasting platinum prices. These results highlight the importance of applying advanced machine learning techniques in the field of econometrics to improve forecasting accuracy and decision making in financial and commodity markets.

The interplay of these multiple factors underlines the complexity of platinum price dynamics and underlines its sensitivity to both macroeconomic and sector-specific developments. This comprehensive understanding is critical to the proper assessment and forecasting of platinum prices, given their strong links to global economic and political conditions and their specific applications in key industries.

4 Conclusions

The purpose of this study is to determine the method that best predicts the platinum price over an 18-month horizon, using the prices of 12 other commodities. Several time series and machine learning models were used, including ARIMA, VARMA, SVM, MARS and MLP.

After rigorous evaluation, the MLP method was found to outperform the other models, producing the lowest root mean square error (RMSE) for platinum price predictions. This indicates that the MLP has superior predictive

accuracy in capturing the complex relationships between platinum and selected commodities.

The incorporation of both time series and machine learning methods allowed for a comprehensive analysis of platinum price dynamics, taking into account both short and long term trends. The inclusion of 12 predictor variables improved the robustness of the models and provided valuable insights into the interconnected nature of commodity markets.

The results highlight the potential of machine learning techniques, particularly MLP, to accurately predict platinum prices relative to other commodities. These predictive capabilities are important for investors, industry professionals and policy makers as they enable better informed decisions and risk management strategies.

This research contributes to the field of econometrics by demonstrating the applicability of advanced modelling techniques in predicting precious metals prices, and provides valuable implications.

Finally, we believe that it would be interesting for future work to combine models and extend the studies by looking at other economic and social factors related to the price of platinum.

References

1. Soriano, C.M.: Platinum. IEEE Potentials 16:29–32. (1998)
2. Bossi, T., Gediga, J.: The Environmental Profile of Platinum Group Metals. Johnson Matthey Technology Review 61:111–121 (2017).
3. Glaister, B.J., Mudd, G.M.: The environmental costs of platinum–PGM mining and sustainability: Is the glass half-full or half-empty? Minerals Engineering 23:438–450 (2010).
4. Griffith, W.P.: The group VIII platinum-group metals and the Periodic Table. Found Chem 12:17–25 (2010).
5. Yang, C-J.: An impending platinum crisis and its implications for the future of the automobile. Energy Policy 37:1805–1808 (2009).
6. Devendra, B.K., Praveen, B.M., Tripathi, V.S., et al.: Highly corrosion resistant platinum-rhodium alloy coating and its photocatalytic activity. Inorganic Chemistry Communications 134:109065 (2021).
7. Dilruba, S., Kalayda, G.V.: Platinum-based drugs: past, present and future. Cancer Chemother Pharmacol 77:1103–1124 (2016).
8. Mingming, T., Jinliang, Z.: A multiple adaptive wavelet recurrent neural network model to analyze crude oil prices. Journal of Economics and Business 64:275–286 (2012).
9. Al-Dhuraibi, W.A., Ali, J.: Using classification techniques to predict gold price movement. In: 2018 4th International Conference on Computer and Technology Applications (ICCTA). IEEE, Istanbul, pp 127–130 (2018)
10. Guha, B., Bandyopadhyay, G.: Gold Price Forecasting Using ARIMA Model. JOAMS 117–121 (2016).
11. Cohen, G.: Algorithmic Strategies for Precious Metals Price Forecasting. Mathematics 10:1134 (2022).
12. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and Its Applications: With R Examples. Springer International Publishing, Cham (2017).

13. Taneja, K., Ahmad, S., Ahmad, K., Attri, S.D.: Time series analysis of aerosol optical depth over New Delhi using Box–Jenkins ARIMA modeling approach. Atmospheric Pollution Research 7:585–596 (2016).
14. Sánchez, A.B., Ordóñez, C., Lasheras, F.S., et al.: Forecasting SO2 Pollution Incidents by means of Elman Artificial Neural Networks and ARIMA Models. Abstract and Applied Analysis 2013:1–6 (2013).
15. Tsay, R.S.: Multivariate time series analysis: with R and financial applications. John Wiley & Sons, Hoboken, New Jersey (2014).
16. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning with Applications in R. Springer New York, New York, NY (2013).
17. Schölkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond, Reprint. MIT Press, Cambridge, Mass. (2002).
18. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer New York, New York, NY (2009).
19. Bishop, C.M.: Pattern recognition and machine learning. Springer, New York (2006).
20. Busto Serrano, N., Suárez Sánchez, A., Sánchez Lasheras F., et al.: Identification of gender differences in the factors influencing shoulders, neck and upper limb MSD by means of multivariate adaptive regression splines (MARS). Applied Ergonomics 82:102981 (2020).
21. Kartal Koc, E., Bozdogan, H.: Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function. Mach Learn 101:35–58 (2015).
22. Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. Springer New York, New York, NY (2004).
23. Krzemień, A., Riesgo Fernández, P., Suárez Sánchez, A., Sánchez Lasheras, F.: Forecasting European thermal coal spot prices. Journal of Sustainable Mining 14:203–210 (2015).
24. Ramani, V., Ghosh, D., Sodhi, M.S.: Understanding systemic disruption from the Covid-19-induced semiconductor shortage for the auto industry. Omega 113:102720 (2022).

Part VII

Complex Networks, Graphs, and Applications

Higher order networks and hypergraphs: A different approach for the detection of communities

Gonzalo Contreras-Aso[1,2], Regino Criado[1,2,3], Guillermo Vera de Salas[1] and Jinling Yang[1,4]

[1] MACIMTE Dept., Rey Juan Carlos University, C/ Tulipán s/n, 28933-Madrid (Spain),
[2] Laboratory of Mathematical Computation on Complex Networks and their Applications, Universidad Rey Juan Carlos,
C/ Tulipán s/n, Móstoles, 28933 Madrid, Spain
{gonzalo.contreras,regino.criado,guillermo.vera,jinling.yang}@urjc.es,
WWW home page: https://lacomarca-lab.github.io/
[3] Data, Complex Networks and Cybersecurity Sciences Technological Institute, Universidad Rey Juan Carlos,
Plaza Manuel Becerra 14, 28028 Madrid, Spain
[4] Unmanned Systems Research Institute, Northwestern Polytechnical University, Xi'an 710072, Shaanxi, PR China

Abstract. The communities of nodes of a hypergraph or higher order network are formed by groups of nodes that share many hyperedges, so that the number of hyperedges they share with the rest of the nodes is significantly smaller, so that these communities can be considered as independent compartments (or superclusters) of the hypergraph. An approach is presented that relies on the so-called derivative graph of a hypergraph, which allows the detection of communities of a higher order graph without a high computational cost, showing important computational advantages of the proposed methods over other existing methods.

Keywords: Hypergraph, Derivative of a hypergraph, Higher-order network, Communities in a hypergraph

1 Introduction and previous concepts

In the real world, there are many examples related to the usefulness of the study of communities in the context of complex network science: families, virtual communities (Facebook, Twitter,...), groups of proteins with similar functions within the cell, companies and customers with the same profiles and many others [1–5]. Thus, different algorithmic methods for the detection of communities have been appearing in the many disciplines in which this tool has applications [6–9]. Community detection in the context of higher-order networks has also received much attention from the network science community [10–13].

We are going to present a new approach for community detection in higher-order networks based on the concept of a derived graph associated with a hypergraph [14, 15], which, in addition to being naturally adapted to hypergraphs and higher-order networks, presents certain computational advantages over other approaches and the associated methods usually used on this type of structures.

A hypergraph (or higher order network) is a pair of sets $\mathcal{H} = (X, \varepsilon)$ in which $X = \{1, ..., N\}$ is a finite set of nodes and $\varepsilon = \{h_1, h_2, \ldots, h_n\}$ is a collection of subsets of X such that $h_i \neq \emptyset \quad (i = 1, 2, \ldots, n)$ and $X = \bigcup_{i=1}^{n} h_i$. The elements of ε are called hyperedges. Thus, hypergraphs appeared as the natural extensions of graphs to describe group interactions between sets of nodes.

It is usual to resort to some matricial representation of the hypergraph. In this context, it is usual to employ the incidence matrix $I(\mathcal{H}) \equiv (I_{ih}) \in \mathbb{R}^{N \times |\varepsilon|}$. This matrix is defined as

$$(I_{ih}) = \begin{cases} 1 & \text{if} \quad i \in h, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

It is not difficult to check that

$$I(\mathcal{H}) \cdot I(\mathcal{H})^t = A(\mathcal{H}) = (a_{ij}) \in \mathbb{R}^{N \times N},$$

where

$$a_{ij} = \begin{cases} |\{h \in \varepsilon \mid \ i \in h\}| & \text{if} \quad i = j, \\ |\{h \in \varepsilon \mid \ i, j \in h\}| & \text{if} \quad i \neq j. \end{cases} \tag{2}$$

To define a criterion for how different communities are established, we need a measure to establish the degree of similarity between sets of nodes. The basic Jaccard index to compare the degree of coincidence or similarity between two sets A and B can be obtained from the formula

$$\mathcal{J}(\mathcal{A}, \mathcal{B}) = \frac{|A \cap B|}{|A \cup B|} \qquad 0 \leq \mathcal{J}(\mathcal{A}, \mathcal{B}) \leq 1.$$

It is important to note that since the establishment of this similarity measure, several generalizations and refinements of this mean (including the overlapping index) have been appearing in the scientific literature. In particular, one of the most widely used is

$$\mathcal{J}_n(A, B) = \frac{|A \cap B|^n}{|A \cup B|}, \ \ \mathcal{I}(A, B) = \frac{|A \cap B|}{min\{|A|, |B|\}}$$

and $\mathcal{C}(A, B) = \mathcal{J}(\mathcal{A}, \mathcal{B}) \cdot \mathcal{I}(\mathcal{A}, \mathcal{B})$. Observe that $0 \leq \mathcal{C}(\mathcal{A}, \mathcal{B}) \leq 1$. In our case, to define the similarity criterion between nodes we will use the concept of derivative graph of a hypergraph [14, 15]:

Definition 1. *Given a hypergraph $\mathcal{H} = (X, \varepsilon)$, with $A(\mathcal{H}) = (a_{ij}) \in \mathbb{R}^{N \times N}$, the derivative hypergraph of $\mathcal{H}$ with respect to the pair of nodes $i, j \in X$ is the numerical value*

$$\frac{\partial \mathcal{H}}{\partial \{i, j\}} = \frac{a_{ii} - a_{ij} + a_{jj} - a_{ij}}{a_{ij}} = \frac{a_{ii} - 2a_{ij} + a_{jj}}{a_{ij}}. \tag{3}$$

2 Two new methods to detect communities in hypergraphs

2.1 Hierarchical clustering and linking function

Within hierarchical clustering, there are two procedures: agglomerative and divisive. Agglomerative clustering, which is the focus of this work, merges the pair of closest clusters in each step until there is one final node left, which comprises the entire data set.

The agglomerative hierarchical clustering method is particularly useful for partitioning data sets for which merely two pairwise distance functions are defined: to measure distances between nodes and to measure distances between clusters linkage function. In our study we will focus on the so called average link (UPGMA) (Unweighted Pair Group Method with Arithmetic Mean) [17] using the derivative between nodes as a "semidistance".

The **linking** function considered is the basis of the corresponding agglomerative method of hierarchical clustering. Thus, given a "proximity" matrix (i.e.: adjacency matrix from a derivative graph):

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix} \rightarrow \begin{pmatrix} 0 & \frac{\partial \mathcal{H}}{\partial\{1,2\}} & \cdots & \frac{\partial \mathcal{H}}{\partial\{1,n\}} \\ \frac{\partial \mathcal{H}}{\partial\{1,2\}} & 0 & \cdots & \frac{\partial \mathcal{H}}{\partial\{2,n\}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{H}}{\partial\{1,n\}} & \frac{\partial \mathcal{H}}{\partial\{2,n\}} & \cdots & 0 \end{pmatrix}$$

the **linkage** is merging two clusters X, Y if $D(X, Y) \in \mathbb{R}$ is the minimum than any other $D(U, V)$ for any clusters U, V.

Thus, the linking function is merging two clusters C_i, C_j if $D(C_i, C_j) \in \mathbb{R}$ is the minimum than any other $D(C_m, C_n)$ for any clusters C_m, C_n. In our case (UPGMA)

$$D(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$$

where, in this case, $d(x, y) = \frac{\partial \mathcal{H}}{\partial\{x,y\}}$. From here, the method consists of calculating the derivatives of the hypergraph with respect to each pair of its nodes, and calculating the linkage function between the clusters obtained at each step, recording the largest gap between two steps, obtaining the corresponding partition in communities by cutting the dendrogram by the segment corresponding to the largest gap.

It is important to emphasize that the use of the type of similarity related to the derivative graph represents a certain conceptual leap, as comparing the similarity of two nodes in both a hypergraph and a hyperstructure can be understood in the context of the study of mesoscale [14, 16].

2.2 Maximizing modularity

Modularity of a pairwise network is a concept used in traditional network analysis to measure the strength of this community structure. It is a scalar value that ranges from -1 to 1, with higher values indicating a stronger modular structure.

$$Q = \frac{1}{2m} \sum_{ij} \left[a_{ij} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j), \tag{4}$$

where $A = (a_{ij})$ is the adjacency matrix of the network, m is the number of edges, k_i and k_j are the degrees of nodes i and j.

The Iteratively Reweighted Modularity Maximization (IRMM) algorithm by Kumar et [18] aims to discover a partitioning that maximizes the community structure in the hypergraph. This type of algorithms can be applied iteratively to each of the obtained communities. The idea of the IRMM algorithm is to apply on the clique reduced graph of $\mathcal{H}$ to get a partition that maximizes the modularity. By iteratively reweighting the hyperedges maximizing the modularity, the IRMM algorithm aims to discover a partitioning that maximizes the community structure in the hypergraph. The iterative process helps refine the clustering by gradually adapting the hyperedge weights and node assignments. In order to compare this algorithm with our method, it is necessary to use the same adjacency matrix to make such a comparison. Our second method consists of calculating the derivatives of the hypergraph with respect to each pair of its vertices and, from there, consider the reduced graph (with the possible appearance of clusters with more than one node in the first step, if the value of the derivative is zero) and the linkage function over the clusters in the dendogram until the number of communities that maximizes modularity is obtained. This second method can be applied iteratively.

Figure.1 shows a Toy-Model, in which the considered hypergraph has 13 nodes (numbered from 0 to 12) and 16 hyperedges, serves to illustrate both the described method and the other two methods considered, our second method, baseda on the concept of modularity and the method of Kumar et al. [18]. As it can be seen, when applying the three methods in this Toy-Model, 3 communities appear in the three cases: [[0, 1, 2, 3, 3, 4], [5, 6, 7, 8], [9, 10, 11, 12]] with a modularity of 0.473064735174287.

3 Real world examples and comparative results

In Figure.2 each node is an author who has collaborated with Prof. Boccaletti, with each hyperedge being a scientific publication. The source of the data is Scopus, and it amounts to a total of 338 publications with 413 co-authors. We included another set with all publications (not including Prof. Boccaletti), of each of the co-authors. The hypergraph is thus enlarged, containing now a total of 15237 hyperedges. The hypergraph is filtered based on the following criteria: we only keep authors with 5 or more publications in common with Prof. Boccaletti

(i.e. we are considering *frequent* co-authors). This filtered hypergraph contains 67 authors with 1685 publications among them and/or Prof. Boccaletti.

It is important to highlight that while our algorithm can work with the hypergraph as-is, we have found that the IRMM algorithm does not converge in reasonable time (more than 24 hours in a dedicated server with 4.0GHz Intel Xeon Gold 5220R) when applied to it. In fact, when applying the four methods (derivative graph highest gap cut, maximum modularity, iterated maximum modularity, IRMM), to the Stefano Boccaletti's coauthors hypergraph we obtain the following quantitative results

Method	**N. of com.**	**Mod.**	**Av. time**
Height-based cut	16	0.642	0.002s
Max. modularity	9	0.678	0.457s
Max. modularity, iterated	24	0.564	0.719s
IRMM	9	0.714	146.604s

One of the main issues one faces when applying community detection to a real network is whether the partition obtained "makes sense". In order to validate the proposed methods, and apply them to a real dataset where there is "universal agreement" on the communities obtained, based on the information in the dataset, we have analyzed a classically labeled dataset [19,20]. A quantitative comparison between both methods applied to this dataset can be found in the table:

Method	**N. of com.**	**Mod.**	**Av. time**
Height-based cut	8	0.428	0.008s
Max. modularity	6	0.435	8.256s

As it can be seen the highest-based cut method performs the community detection task very efficiently and both methods give the expected answers that match perfectly with what is expected. The maximum modularity method suffers from the fact that the construction of the clique-reduction, so it is computationally more expensive.

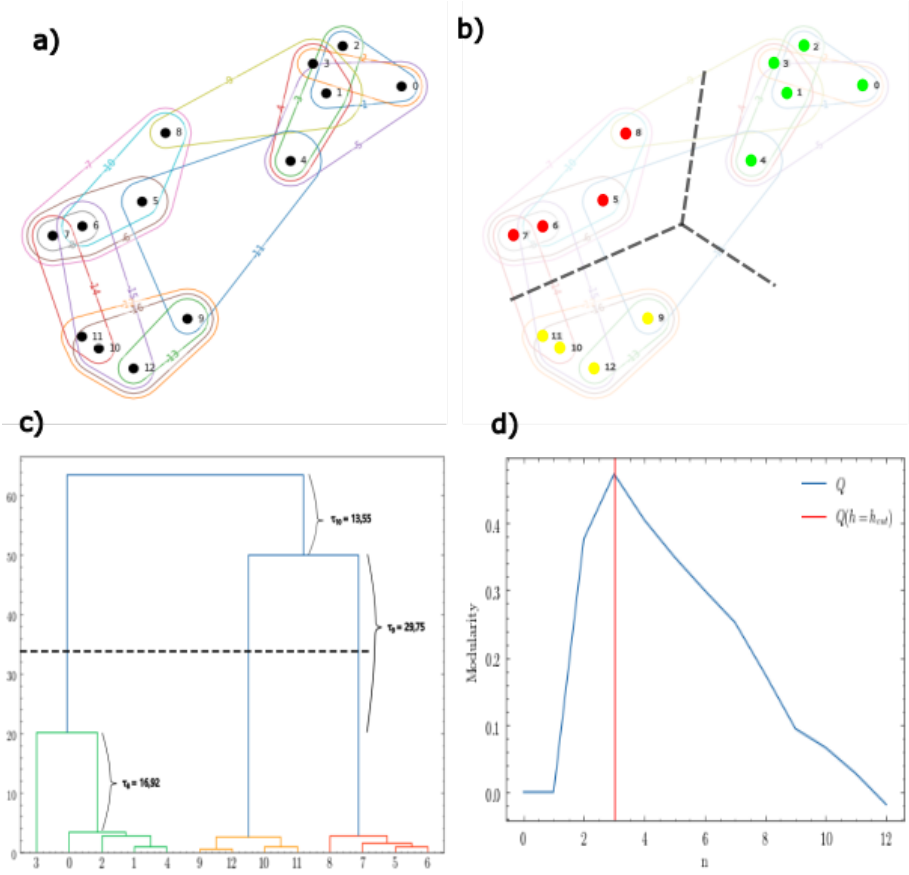


Fig. 1: **a)** Toy example hypergraph. **b)** Partition into communities using either of the two methods based on the derivative graph. **c)** Dendrogram corresponding to the average (UPGMA) clustering via the derivative graph, as discussed in the main text. **d)** Modularity at each partition of the toy hypergraph given by its dendrogram, where n is the number of communities. The modularity given by the partition at the highest gap is also explicitly shown.

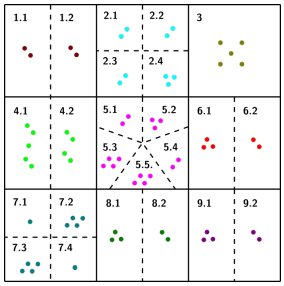


Fig. 2: The image represents the communities of Professor Stefano Boccaletti's network of co-authors as described criteria classified by our modularity maximization method. The image also shows the subcommunities obtained by iterating over each of the partitions.
1.1: Li D., Havlin S.; **1.2:** Barzel B., Zhang X.;
2.1: Bragard J., Mendoza C.; **2.2:** Kurths J., Zhou C.S.; **2.3:** Mancini H., Maza D.; **2.4:** Meucci R., Allaria E., Arecchi F.T.;
3: Bortolozzo U., Ramazza P.L., Pampaloni E., Residori S., Giaquinta A.;
4.1: Jusup M., Wang Z., Li X., Dai X., Perc M.; **4.2:** Shi L., Guo H., Jia D., Shen C.;
5.1: Sousa P.A.C., Menasalvas E.; **5.2:** Papo D., Buldú J.M., Zanin M.; **5.3:** del-Pozo F., Gutiérrez R., Maestú F., Bajo R.; **5.4:** Jaimes-Reátegui R., Sevilla-Escoboza R.; **5.5:** Navas A., Sendiña-Nadal I., Leyva I., Almendral J.A.;
6.1: Hramov A.E., Koronovskii A.A., Moskalenko O.I.; **6.2:** Maksimenko V.A., Makarov V.V.;
7.1: Raigorodskii A.M.; **7.2:** Frasca M., Moreno Y., Latora V., Gómez-Gardeñes J.; **7.3:** del Genio C.I., Alfaro-Bittner K., Criado R., Romance M.; **7.4:** Musatov D.;
8.1: Guan S., Liu Z., Zou Y.; **8.2:** Qiu T., Bonamassa I.;
9.1: Chavez M., Amann A., Hwang D.-U.; **9.2:** Valladares D.L., Pecora L.M.

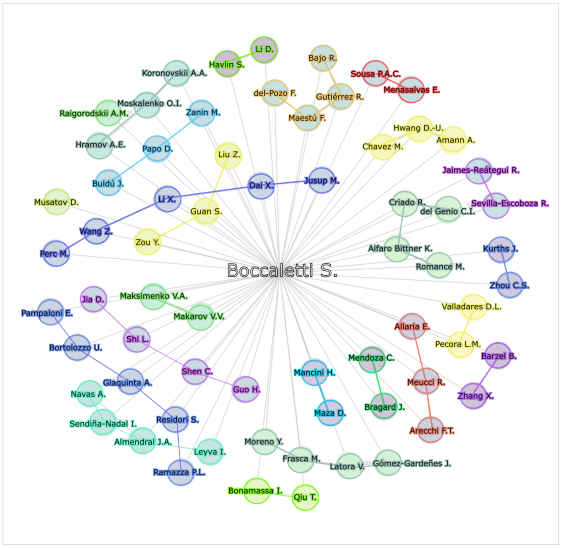


Fig. 3: Representation of Prof. Stefano Boccaletti's co-author network communities (source: Scopus. Updated as of June 30, 2023).

4 Conclusions

We present two methods for the detection of high-order network communities (hypergraphs) without high computational cost and that relies on the so-called derivative graph of a hypergraph. The second method that maximizes modularity is computationally more expensive than the first one. And last, but not least, through several simulations it is shown that this second method achieves a very high modularity value close to that of other methods (IRMM) which are much more computationally expensive and, moreover, in one of the examples shown, fails to complete the required computation.

References

1. Fortunato, S. Community detection in graphs. Physics reports 486, 75-174 (2010).
2. Fortunato, S., Newman, M.E.J., 20 years of network community detection. Nature Physics 18, 848–850. (2022).
3. Newman, M.E.J., Modularity and community structure in networks. Proceedings of the National Academy of Science of the USA, Vol 103, No. 23, pp 8577–8582 (2006).
4. Evans, T.S., Lambiotte, R., Line graphs, link partitions, and overlapping communities, *Phys.Rev. E* **80**, 016105 (2009).
5. Newman, M. E. J. & Peixoto, T. P. Generalized communities in networks. Phys. Rev. Lett. 115, 088701 (2015)
6. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U., Complex networks: Structure and dynamics,*Phys.Rep.* 424 75–308 (2006).
7. Estrada, E., Networks Science, Springer, New York (2010).
8. Wasserman, S., Faust, K., Social Network Analysis, Cambridge University Press, Cambridge, (1994).
9. Costa, L.d.F., Oliveira, O.N., Travieso, G., Rodrigues, F.A., Villas Boas, P.R., Antiqueira, L., Viana, M.P., Correa Rocha, L.E., Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Advances in Physics 60, 3, 329–412 (2011).
10. Carletti, T., Fanelli, D., Lambiotte, R (2021). Random walks and community detection in hypergraphs, J. Phys. Complex. 2, 015011 (2021).
11. Chien, I., Lin, C., Wang, I.. (2018). Community Detection in Hypergraphs: Optimal Statistical Limit and Efficient Algorithms. Proceedings of Machine Learning Research, 84:871–879 (2018). Available from https://proceedings.mlr.press/v84/chien18a.html.
12. Contisciani M., Battiston F., and De Bacco C. (2022). Inference of hyperedges and overlapping communities in hypergraphs, Nature Communications, 13:7229, (2022).
13. Zhang, Y., Lucas, M. & Battiston, F. Higher-order interactions shape collective dynamics differently in hypergraphs and simplicial complexes. Nat Commun 14, 1605 (2023).
14. Criado-Alonso, A., Aleja, D., Romance, M., Criado, R., Derivative of a hypergraph as a tool for linguistic pattern analysis, Chaos, Solitons & Fractals 163, 112604 (2022).
15. Criado-Alonso, A., Aleja, D., Romance, M., Criado, R.: A new insight into linguistic pattern analysis based on multilayer hypergraphs for the automatic extraction of text summaries. Math. Meth. Appl. Sci.. 2023;1–18 (2023).

16. Criado R., Romance M., Vela-Pérez M., Hyperstructures, a new approach to complex systems, I.J.B.C., 20 (3) (2010), pp. 877–883
17. Cormack, R. M., A review of classification (with discussion). Journal of the Royal Statistical Society. Series A. General, 134 (3), 321–367 (1971)
18. Kumar, T., Vaidyanathan, S., Ananthapadmanabhan, H. et al. Hypergraph clustering by iteratively reweighted modularity maximization. Appl Netw Sci 5, 52 (2020).
19. Benson, A. R., Abebe, R., Schaub, M.T., Jadbabaie, A., Kleinberg, J., Simplicial closure and higher-order link prediction. Proceedings of the National Academy of Science 115 (48): E11221-E11230, (2018).
20. Stehlé, J., Voirin, N., Barrat, A., Cattuto C., Isella, L., Pinton, J-F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., Vanhems, P., High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. PLoS ONE 6 (8) (2011).

Searching communities' border in badly conditioned graphs through fuzzy convolution techniques on linearized graphs

J.M. Montañana[1], A. Hervás[2], S. Morillas[3], and J. Pellicer[4]

[1] inpeek GmbH, Darmstadt, Germany,
[2] IUMM, Universitat Politècnica de València, Valencia, Spain,
[3] IUMPA, Universitat Politècnica de València, Valencia, Spain,
[4] ETSINF, Universitat Politècnica de València, Valencia, Spain.
ahervas@imm.upv.es

Abstract. Many complex systems can be modeled by graphs and networks.A series of factors must be considered when studying the properties of graphs. It is primarily interesting the study graph communities. Such communities are subgraph structures that have more connections within the same subgraph than with the rest of the graph. In some problems, the study of communities allows quantitative and qualitative approaches and obtaining some knowledge about the structure of the graph and what it represents [2, 3, 7].
There is extensive literature on the study of communities, mostly focused on non-directed graphs [2, 3, 7]. In our case, we focus our work on the study of communities in directed graphs, weakly connected, with weights on the edges. In this paper, we present briefly two alternatives to obtain the communities in a directed graph using convolution techniques and fuzzy tools applied on linearized graphs that allow us to prune non-significant edges and study the behavior for the different procedures presented.

Keywords: Complex Networks, Community Detection, Convolution, Fuzzy filters.

1 Introduction

In a complex network, a community of vertices is defined as a subset of network vertices that are highly related to each other and less to the rest. Nodes grouped in the same community have common characteristics that make them play a certain role within the network.

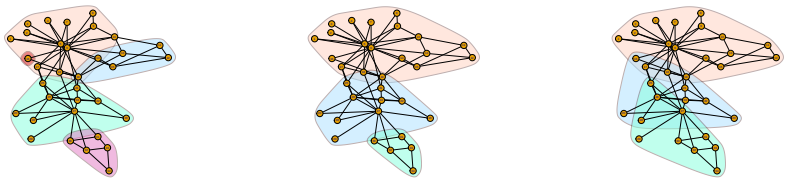
The study of those subgraphs whose vertices, which have more connections within the same subgraph than with the rest of the graph, appear related to problems such as reassignment of students in grades, trophic chains, traffic in airports, and public transport networks [3, 4, 6].

Obtaining these communities is a problem that has merited attention from different points of view. There is extensive literature on the study of communities,

mostly focused on non-directed graphs [2,3,7]. Modularity allows us to measure the goodness of the approximation obtained by the algorithm.

There are different algorithms to obtain the communities, however, depending on the criteria or properties used by each algorithm, slightly different results are obtained. So, for example, if we consider the graph of Zachary, where students of a karate club must opt for one of the two senseis, applying different algorithms we get different results, even with a similar modularity, see Figure 1.

Our interest is focused on the study of communities in directed graphs, weakly connected, with weights on the edges. The algorithms to find communities usually do not work satisfactorily in this type of graph.



(a) Modularity 0.401 (b) Modularity 0.440 (c) Modularity 0.399

Fig. 1: Communities obtained on the Zachary graph with (a) Girvan-Newman, (b) Walk Trap, and (c) Label Propagation algorithms.

2 State of the art

Recently, different alternatives have been proposed to obtain the communications in directed graphs, by the way: in 2016, Hervás et al proposed a pruning algorithm; in [12] an algorithm based on center detection.

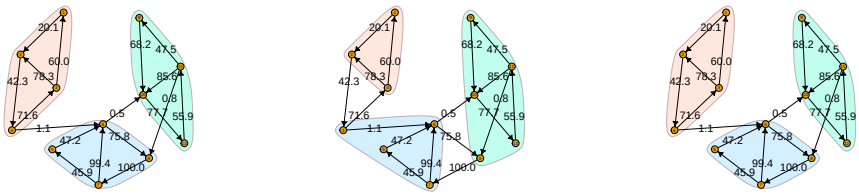
A new approach was proposed in [11] that uses convolution techniques to solve problems related to tax fraud detection. The general idea is to apply edge detection techniques in images. Detecting a border in an image consists of identifying sudden changes in the intensity of the image, in some cases, these changes are perfectly detectable, in other cases minimal changes of intensity appear that we cannot consider as borders and that we will label as noise.

The first step to detect the edges is to use filters to smooth the image, eliminating noise, the second phase is to enhance the image obtained and finally detect pixels where there are edges. Muñoz et al propose the ConvGraph algorithm that consists of five steps:

1. Obtain the linear graph of G, also known as the dual graph, to more easily detect the edges.
2. Applying a convolution filter to the graph H, in this case, applies to Gaussian and the Laplacian of the Gaussian
3. Detect edges and prune when sign changes are detected in the convolution.

4. Use a ratio measuring community weights against the total graph to validate the community.
5. Transfer the communities from the dual graph to the original graph.

The results are more than acceptable as we can see in detected communities in Figure 2 and the obtained modularity coefficients.



(a) (b) (c)

Fig. 2: (a) The desired communities to be detected. Detected communities with (b) Girvan-Newman (mod. 0.490), and with (c) ConvGraph (mod. 0.661).

3 Using fuzzy logic in the convolution

The question to ask ourselves is if there are other filters that can provide us with better results. Related to this question lies the fact that complex datasets are not free from uncertainty, noise and imprecision. So, alternative solutions can be found if we use tools able to represent and manage these concepts in the datasets. For this, one interesting tool is fuzzy logic. The Fuzzy logic has been used previously for processing, for example digital images containing acquisition or transmission noise. The development of methods and techniques to eliminate noise as accurately as possible is crucial in this area. In 2007 [9] addressed the difficulties due to the combination of Gaussian and impulsive noise, and in [10] proposed a method based on fuzzy metrics.

The proposed method is based on the concept of peer group [2]. This approach considers that a pixel in an image is surrounded by a set of neighbors similar to it, forming the so-called peer group. Establishing this similarity between pixels is not an easy process, so the authors decided to use fuzzy filters and obtain competitive results compared to other methods of noise detection. Our objective is to target communities through the use of these filters. Fuzzy logic is a fascinating tool to explore for the problem addressed in this work.

4 Proposed Algorithm

Before presenting the proposed algorithm, we introduce the concept of linear graph, and we also provide the definitions of fuzzy distance which is central element to our approach.

Line graph: For a given graph $G = (V, E)$, where V represents the set of vertices and E the set of edges, it can be obtained a line graph $L_G = (V', E_L)$, where the line graph is defined as: The vertices $V' = \{v'_1, v'_2, \ldots, v'_n\}$ of the linear graph L_G correspond to the edges E of the original graph G. Each pair of vertices in L_G are adjacent if and only if their corresponding edges in G share a common vertex.

Fuzzy metric: We define the fuzzy metric $M(source, destination, t)$ as the measure of the fuzzy distance between vertices $source = v'_i$ and $destination = v'_j$ in the linear graph with a sensitivity parameter t. The metric is defined with the function shown in Equation 1.

$$M(s, d, t) = \frac{t}{t + \left|s - d\right|} \tag{1}$$

Weighted fuzzy distance: The weighted fuzzy distance W_s for a vertex s is defined as the difference of the vertex s in the line graph and the sum of the other destination vertices different than s calculated with the next Equation 2:

$$W(s, t) = \frac{t}{t + \left|v'_s - \sum_{j=1}^{j<=n} v'_j\right|} \quad \text{where } j \neq s \tag{2}$$

The proposed algorithm detects the constituent edges of the community borders as those whose corresponding vertex in the linear Graph has an weighted fuzzy distance significantly different from the weighted fuzzy distance of the other vertices. And after that pruning those community edges the graph communities can more easily be detected.

4.1 Algorithm: ConvGraph using Fuzzy filters

Algorithm 1 ConvGraph using Fuzzy filters.

Data: Directed graph $G =< V, E >$
Result: Communities detected after applying fuzzy filters to the graph
Transform the original graph G to its graph line L_G
for each *source* vertex of the line graph vertices V' **do**
  Calculate the weighted distance $W(source, t)$
end for
Detect borders between communities and select edges to prune when the corresponding weighted distance W_i have a significant difference with the weighted distance of the rest of edges

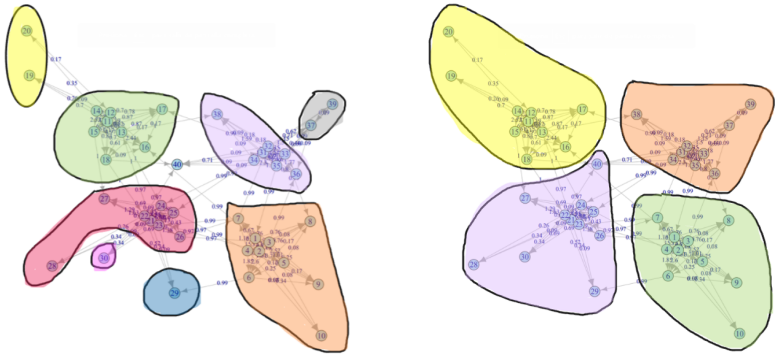
We developed an improvement for this algorithm that eases finding the borders between communities increasing the difference in their weighted fuzzy distances and reducing the influence of the parameter t. It was achieved with a

different definition of fuzzy distance and weighted distance shown in Equations 3 and 4.

$$M'(x,y,t) = \frac{t}{t + \left|(n-1)x - y\right|} \tag{3}$$

$$W(s,t) = \frac{t}{t + \left|(n-1)v'_s - \sum_{j=1}^{j<=n} v'_j\right|} \quad \text{where } j \neq s \tag{4}$$

Figure 3 shows a graph directed with four communities obtained using the generator proposed at [8], and the results when applying the two versions of the algorithm.



(a) (b)

Fig. 3: Detection of Graph Communities when using the ConvGraph algorith (a) with metric M, and (b) with metric M'.

(a) Modularity 0.699 (c) Modularity 0.699

Fig. 4: Same Graph Communities are detected in the Zachary's graph when using the ConvGraph algorith (a) with metric M, and (b) with metric M'.

5 Conclusions

The use of Fuzzy filters combined with the ConvGraph algorithm has shown efficient in obtaining communities in poorly conditioned directed graphs, obtaining more accurate results and improving the values of modularity. In fact, by applying these procedures to untreated graphs, for example, Zachary's graph in Figure 1 shows a better approach to communities and better values of modularity, as we can see in 4. Until here the objectives of this work arrive, our steps from now are directed to determine the values of the parameter t that allow us to optimize the obtaining of communities.

References

1. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. U. (2006). Complex networks: Structure and dynamics. Physics reports, 424(4-5), 175-308.
2. Deng, Y., Kenney, C.,Moore, M. S., and Manjunath, B. S. (1999). Peer group filtering and perceptutal color image quantization, in Proc. IEEE Int. Symp. Circuits Systems, 1999, vol. 4, pp. 21-24.
3. Feng S, Hu B, Nie C, Shen X (2016) Empirical Study on a Directed and Weighted Bus Transport Network in China. Physica A: Statistical Mechanics and its Applications 441(C):85-92.
4. Girvan, M.; Newman, M. E. J. (2002). Community structure in social and biological networks". Proceedings of the National Academy of Sciences. 99 (12): 7821-7826. 5
5. Hervás, A., Soriano, P. P., Capilla, R., and Peinado, J. (2016) An algorithm to obtain directional communities in a directed graph. In L. Jodar Ed. Proceedings of the Mathematical Modelling in Engineering & Human Behaviour, 347-352. SPUPV, Valencia, Spain.
6. Hervás, A., Soriano-Jiménez, P. P., Jiménez, A., Peinado, J., Capilla, R., and Montañana, J. M. (2015). A Case Study of Directional Communities in a Directd Graph: The Accessing Procedure to the Spanish Public University System. Modeling Human Behavior, 9. Nova Science Publishers, Inc. Hauppauge, NY. USA.
7. Lancichinetti A, and Fortunato S (2009) Community detection algorithms: A comparative analysis. Phys. Rev. E 80(5):056117.
8. Montañana, J. M., Hervás, A., and Soriano Jiménez, P. P. (2020). A Proposal for a Benchmark Generator of Weakly Connected Directed Graphs. Open Journal of Modelling and Simulation, 8(1), 18-34.
9. Schulte, S., Morillas, S., Gregori, V, and EE Kerre, EE. (2007). A new fuzzy color correlated impulse noise reduction method IEEE Transactions on Image Processing 16 (10), 2565-2575, 92,
10. Morillas, S., Gregori, V., and Hervás, A. (2009). Fuzzy peer groups for reducing mixed Gaussian-impulse noise from color images. IEEE Transactions on Image Processing, 18(7), 1452-1466.
11. Muñoz,H., Vicente,E., Gonzalez,I., Mateos, A., and A. Jimenez-Martin. (2021) ConvGraph: Community Detection of Homogeneous Relationships in Weighted Graphs, Mathematics, vol. 9, no 4, pag. 367.
12. You,T., Cheng,H.-M., Ning,Y.-Z., Shia, B.-C., and Zhang, Z.-Y. (2016) Community detection in complex networks using density-based clustering algorithm and manifold learning, Physica A: Statistical Mechanics and its Applications, vol. 464, pp. 221-230.

Competitiveness of Formula 1 championship from 2012 to 2022 as measured by Kendall corrected evolutive coefficient

Francisco Pedroche[1]

Institut Universitari de Matemàtica Multidisciplinària,
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain
pedroche@imm.upv.es,
WWW home page: http://personales.upv.es/pedroche/

Abstract. In this paper we analyze the FIA formula one world championships from 2012 to 2022 taking into account the drivers classifications and the constructors (*teams*) classifications of each Grand Prix. The needed data consisted of 22 matrices of sizes ranging from 25×20 to 10×19 that have been elaborated from the GP classifications extracted from the official FIA site. We have used the Kendall corrected evolutive coefficient, recently introduced, as a measure of Competitive Balance (CB) to study the evolution of the competitiveness along the years in both drivers and teams championships. In addition, we have compared the CB of F1 championships and two major European football leagues from the seasons 2012-2013 to 2022-2023.

Keywords: Kendall's tau, Formula One, Football, Competitive balance, sport rankings, contest

1 Introduction

A *ranking* naturally appears when we sort elements, being this a key action in more activities such as analysis of sport competitions [2], economic time series [14], comparison of algorithms performance [25], etc. Series of rankings can be studied from different perspectives. For example, to analyse sorting algorithms [15], to define measures of disarray [7], to use rank transformation to develop nonparametric methods in Statistics [5], to *learn to rank* in machine learning applications [4], etc. In this paper we are interesting in characterising a series of rankings by giving a coefficient that measures the disarray along the series in the classic manner of [11]. Specifically, we follow the definitions of [21], [20] and [6].

2 Kendall corrected evolutive coefficient

The Kendall corrected evolutive coefficient, denoted by $\widehat{\tau}_{ev}^{\bullet}$, was introduced in [21]. It takes as input a series of m rankings (with at most n elements) that

can be *complete* (that is, the n elements are ranked in all the rankings) or incomplete. In addition, we consider the existence or not of ties between the ranked elements. Kendall corrected evolutive coefficient can be considered as an extension of a correlation coefficient of two rankings applied to m rankings and therefore, as output, $\widehat{\tau}_{ev}^{\bullet}$ gives a real number in $[-1, 1]$.

The coefficient $\widehat{\tau}_{ev}^{\bullet}$ reduces to some particular coefficients that are well documented and can be found in the literature. For example, when $m = 2$ and the rankings are complete and with no ties, then $\widehat{\tau}_{ev}^{\bullet}$ reduces to the classical Kendall's τ coefficient of disagreement (see [11], [12], [13]) that can be written as

$$\tau = \frac{2(P - Q)}{n(n-1)} \tag{1}$$

where P is the number of pair of elements that keep its relative order from the first ranking to the second one and Q is the number of pairs of elements that change its order. For example, taking $n = 3$, the rankings $\mathbf{a} = [1, 2, 3]$ and $\mathbf{b} = [3, 2, 1]$ have an associated $\tau = -1$ and the rankings $\mathbf{a} = [1, 2, 3]$ and $\mathbf{b} = [1, 2, 3]$ have an associated $\tau = 1$. When $m = 2$ and the rankings are complete and with ties, then $\widehat{\tau}_{ev}^{\bullet}$ is related to the *Kendall distance with penalty parameter* $p \in [0, \frac{1}{2}]$ defined in [8]. When $m > 2$ and the rankings are complete and with ties, then $\widehat{\tau}_{ev}^{\bullet}$ reduces the *corrected evolutive Kendall distance with penalty parameter* p introduced in [20].

In sport competitions it is most used the term *Competitive Balance* (CB) to measure the balance between the teams [27], [19]. A high measure of CB means that the competition is highly interesting since it is very difficult to predict the result of a match (or a race, in our case), while a low measure of CB means that the competition is very predictable, and therefore *boring* (see. [18], [17], [9], [10], [2], [3]). In this regard it is more convenient to use the measure called *Normalized Strength* (borrowed from complex networks terminology, see [6], [1]), and that we define here by

$$NS = \frac{1 - \widehat{\tau}_{ev}^{\bullet}}{2} \tag{2}$$

Note that NS is a normalized index, $NS \in [0, 1]$, and its value can be considered as a measure of CB. We will use this index in our analysis. The interested reader may find the precise definition of $\widehat{\tau}_{ev}^{\bullet}$ in [21] but we omit the details for the sake of brevity.

3 Formula One World Championships

Formula One (also known as Formula 1 or F1) organised by the Fédération Internationale de l'Automobile (FIA) is a well-known international racing for cars [23]. The drivers championship began in the season of 1950, while the constructors championship began in 1958. Along the years, there has been some modifications both in the format and in the rules that the participants must accomplish.

A Formula One season consists of a series of races, each of them known as Grand Prix (denoted as GP), that take place in several countries. For example, the F1 2022 season consisted of 22 GP and participated 10 teams and 22 drivers. A GP is held on a weekend. On friday and saturday some qualifying sessions fix the starting order (*the grid*) for the GP race that occurs on Sunday. In this paper we are interested only in the ranking corresponding to this GP races. This ranking is decided based on the timing of each driver and he receives a quantity of points depending on his ranking. From 2010 to 2018 the sharing of the points was given as shown in Table 1. The points assigned to the constructors in a GP is the sum of the points of the two drivers of the team that participated in that GP.

Table 1: Points scoring sharing since 2010

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
25	18	15	12	10	8	6	4	2	1

From 2019 one additional point is given to the pilot that occupied a position in the top ten and furthermore has the fastest lap in the race. FIA has some rules to break ties between the pilots and therefore the ranking of the drivers can be considered as ranking with no ties. Note, therefore that each GP has its own classification. The final ranking (that is, the F1 Championship) of the season is made by accumulating the points of each GP, and, again, some rules are applied to break the ties, if any. Our collection of rankings are precisely the rankings of each GP in a season, both for drivers and constructors. We use these series of rankings to compute the corresponding $\widehat{\tau}_{ev}^{\bullet}$ of that season, and then the corresponding NS. We precisely describe the used rankings in the next section.

4 Description of the rankings

We have selected the F1 classifications from 2012 to 2022. Our criterium to select our dataset is based on taking the GP classifications of championships in where 1) the regulations does not vary too much, 2) the distribution of points (e.g. as given by Table 1) is quite stable, 3) the number of GP does not vary too much and 4) that the standings can be easily retrieved from the official FIA site [23]. For example, the 2012 season can be retrieved from the FIA site [24]. In Table 2 we show the number of drivers in each championship jointly with the number of GP in that year.

To describe our rankings we use the following notation (see [26], [21]). Let $V = \{v_1, v_2, \cdots, v_n\}$ be the objects to be ranked, with $n > 1$. The ranking is

Table 2: Number of drivers, teams and GP in each analyzed F1 Championship

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Drivers	25	23	24	22	24	25	20	20	23	21	22
Teams	12	11	11	10	11	10	11	10	10	10	10
GP's	20	19	19	19	21	20	21	21	17	22	22

given by

$$\mathbf{a} = [a_1, a_2, \cdots, a_n] \tag{3}$$

where a_i is the position of v_i in the ranking. Note that if $a_i = a_j$, then v_i and v_j are tied. If v_i is not ranked, then it is denoted as $a_i = \bullet$.

4.1 Drivers ranking

From the FIA site, we can retrieve the drivers classification for each GP of the considered championship. In these classifications we can see the ranking, the points obtained by each driver, and a note indicating whether the driver has finished the race or not. To construct our drivers ranking we consider that a driver that has not finished the race (or has not even start it) is an absent element in our ranking, and therefore it is indicated by $\bullet$. For example, in Table 3 we show our notation to describe the first three rankings of the 2012 championship.

4.2 Constructors ranking

From the FIA site we can retrieve the constructors classification for each GP of the considered championship. The points given to a constructor consist of the sum of the points of the two drivers of the corresponding team in each GP. In this case the FIA site offers the points obtained by each constructor. This gives us the opportunity to create two types of rankings, being the interest to see how our measure NS is affected by these types. The two considered methods are the following:

Method 1: We consider that the constructors that have 0 points are tied in the last position.

Method 2: We consider that the constructors that have 0 points are absent elements.

As an example, in Table 4 we show the constructors name, scoring and $\mathbf{a}_i$ vectors (by using Method 1 and Method 2) for the first three GP of FIA 2012 World Championship.

Table 3: Drivers' name, nationality, and $\mathbf{a}_i$ vector for three of the first GP of FIA 2012 World Championship. Elaborated from [23]. Note that • means that the driver did not start or did not finish the race. The rankings are *incomplete rankings with no ties*. The order of the drivers in the first column follows the (final) classification of the constructors championship. The drivers Raikkonen, Grosjean and D'Ambrosio belong to the same team (Lotus F1) while the rest of teams contributed with two drivers in the whole GP rankings of this championship.

Driver	Nat	GP1	GP2	GP3
Sebastien Vettel	DEU	2	11	5
Fernando Alonso	ESP	5	1	9
Kimi Raikkonen	FIN	7	5	14
Lewis Hamilton	GBR	3	3	3
Jenson Button	GBR	1	14	2
Mark Webber	AUS	4	4	4
Felipe Massa	BRA	•	15	13
Romain Grosjean	FRA	•	•	6
Nico Rosberg	DEU	12	13	1
Sergio Perez	MEX	8	2	11
Nico Hulkenberg	DEU	•	9	15
Kamui Kobayashi	JPN	6	•	10
Michael Schumacher	DEU	•	10	•
Paul Di Resta	GBR	10	7	12
Pastor Maldonado	VEN	13	19	8
Bruno Senna	BRA	16	6	7
Jean-Eric Vergne	FRA	11	8	16
Daniel Ricciardo	AUS	9	12	17
Vitaly Petrov	RUS	•	16	18
Timo Glock	DEU	14	17	19
Charles Pic	FRA	15	20	20
Heikki Kovalainen	FIN	•	18	23
Jérôme D'Ambrosio	BEL	•	•	•
Narain Karthikeyan	IND	•	22	22
Pedro De la Rosa	ESP	•	21	21

5 Results

5.1 Comparison of constructors and drivers championships

In order to compare the *competitivity balance* of the GP of drivers and constructors we have computed NS, given by (2) for the GP standings from 2012 to 2022 for drivers and for constructors (with Method 1 and Method 2). The results are shown in Table 5.

Table 4: Constructor's name, scoring and $\mathbf{a}_i$ vectors (by using Method 1 and Method 2) for the first three GP of FIA 2012 World Championship. The order of the teams in the first column follows the (final) classification of the championship.

	Score			Method 1			Method 2		
Constructors	GP1	GP2	GP3	GP1	GP2	GP3	GP1	GP2	GP3
Red Bull Racing	30	12	22	2	4	3	2	4	3
Scuderia Ferrari	10	25	2	4	1	6	4	1	6
Vodafone McLaren Mercedes	40	15	33	1	3	1	1	3	1
Lotus F1 Team	6	10	8	5	5	5	5	5	5
Mercedes AMG Petronas F1 Team	0	1	25	8	8	2	•	8	2
Sauber F1 Team	12	18	1	3	2	7	3	2	7
Sahara Force India F1 Team	1	8	0	7	6	8	7	6	•
Williams F1 Team	0	8	10	8	6	4	•	6	4
Scuderia Toro Rosso	2	4	0	6	7	8	6	7	•
Caterham F1 Team	0	0	0	8	8	8	•	•	•
Marussia F1 Team	0	0	0	8	8	8	•	•	•
HRT F1 Team	0	0	0	8	8	8	•	•	•

The data on Table 5 can be resumed on the box-and-whiskers plot shown on Figure 1. In more detail, the mean values of NS on the period 2012-2022, and the corresponding sample standard deviation, s, are as follows:
Mean value of NS for drivers: 0.2203, ($s = 0.018$).
Mean value of NS for constructors (Method 1): 0.2394, ($s = 0.035$).
Mean value of NS for constructors (Method 2): 0.2771, ($s = 0.070$).

Let us consider that NS is a random variable. By computing the Shapiro-Wilk test for normality [22] we obtain the p-values 0.61, 0.08 and 0.44 for the corresponding NS series for drivers, and constructors (Method 1 and Method 2) respectively. Therefore we cannot reject the normality of the distribution of NS of the corresponding samples. Regarding the mean values of NS for constructors by using Method 1 and Method 2, since they come from the same data (as an example, the scores in Table 4 give us the corresponding values for Method 1 and Method 2) we can use a comparison method for means coming from paired data. By using a t-test we obtain a p-value of 0.18 and therefore we cannot reject that the means are equal with a confidence interval of 95%. Since the value of the variances does not have a ratio major than 4 we can use the t-test for comparing the mean of NS by using Method 1, and the corresponding NS for drivers. We obtain that the p-value is 0.12 and therefore we cannot reject the null hypothesis that the means are equal. All in all we have the statistically the three values of NS are not different, with a confidence interval of 95%.

Table 5: NS for the series of GP of the Championships from 2012 to 2022 for drivers and constructors.

Year	NS Drivers	NS Constructors Method 1	NS Constructors Method 2
2012	0.2561	0.2456	0.4052
2013	0.2136	0.1924	0.3421
2014	0.1913	0.1616	0.3106
2015	0.2270	0.2722	0.2350
2016	0.2065	0.2218	0.2143
2017	0.2140	0.2632	0.2179
2018	0.2188	0.2559	0.1886
2019	0.2157	0.2772	0.2596
2020	0.2436	0.2562	0.3652
2021	0.2270	0.2413	0.2715
2022	0.2092	0.2455	0.2376

5.2 Comparison of competitiveness between F1 championships and two major European football leagues

A competitive balance measure like NS, based on sport ranking series, can be used to compare the CB of two different sports. For example, by computing the coefficient NS for two major European football leagues (Spanish League - commercially known as Laliga Santander in the season 2022/23-, and the English Premier league) we obtain the results shown in Table 6. We have used the series of standings from the season 2012-2013 to the season 2022-2023 for both the Spanish League (retrieving the data from the links on [28]) and Premier League (retrieving the data from [29]). The summary for the football leagues in the studied period is the following:

Mean value of NS for Spanish league: 0.059, ($s = 0.0094$).

Mean value of NS for Premier league: 0.056, ($s = 0.0062$).

As a consequence, by using the results on section 5.1 for NS of drivers and NS of constructors by using Method 1, we obtain that the mean value of NS for the F1 championships is about four times greater than the values of NS corresponding to the analyzed football leagues.

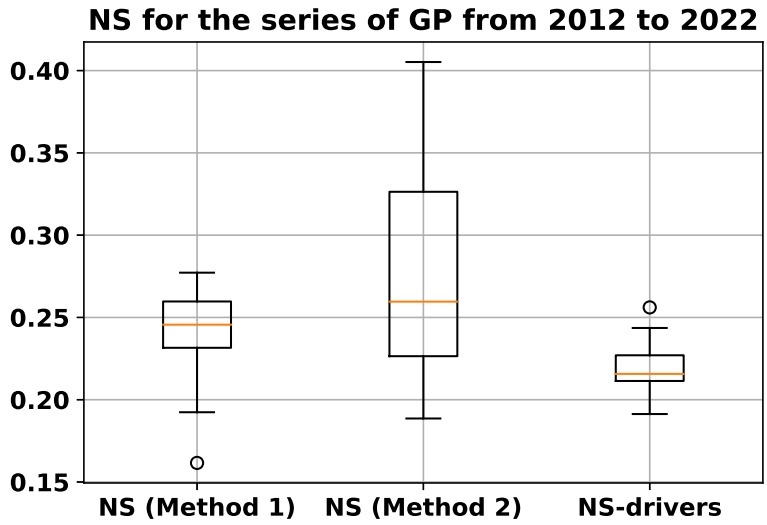


Fig. 1: Box-and-whiskers diagram for NS for the series of GP of the championships from 2012 to 2022 for drivers and constructors (by using the two methods explained on the text).

6 Conclusions

In this communication we have shown how to apply a recently introduced metric to calculate a measure of the competitive balance (CB) associated to Formula 1 championships, by taking into account the standings of the Grand Prix that compose each championship. We have introduced to methods (called Method 1 and Method 2) to compute the CB values of the F1 Constructors Championship in the period 2012-2022. We have obtained that these two methods do not offer mean values that can be considered statistically different. We think that this shows a good behaviour of our metric since both Method 1 and Method 2 are obtained by computing a linear combination from the same set of data (the F1 Drivers Championship) but with different treatment of the constructors that finish with zero points in a Grand Prix. We also have obtained that the CB of the F1 Drivers Championship and F1 Constructors Championship show similar values on the studied period, but with a slightly higher mean value for the Constructors Championship. As an example of the power of our metric, we have compared the CB of two different sports: the Formula 1 championships from 2012 to 2022 and the Spanish football league and Premier football league on the seasons 2012-2013 to 2022-2023. Our results show that the mean value of CB for the F1 championships is about four times greater than the values of CB corresponding to the analyzed football leagues.

Table 6: NS values for two European football leagues from season 2012/2013 to season 2022/2023.

Year	NS Spanish league	NS Premier league
2012	0.0615	0.0514
2013	0.0593	0.0656
2014	0.0546	0.0597
2015	0.0613	0.0563
2016	0.0435	0.0589
2017	0.0589	0.0550
2018	0.0688	0.0489
2019	0.0600	0.0643
2020	0.0757	0.0583
2021	0.0440	0.0461
2022	0.0595	0.0515

References

1. Barrat, A., Barthélemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks, Proceedings of the National Academy of Sciences, 101(11), 3747-3752 (2004). doi:10.1073/pnas.0400087101
2. Basini, F., Tsouli, V., Ntzoufras, I., Friel, N.: Assessing competitive balance in the English Premier League for over forty seasons using a stochastic block model, Journal of the Royal Statistical Society Series A: Statistics in Society, qnad007 (2023). doi:10.1093/jrsssa/qnad007
3. Budzinski, O., Feddersen, A.: Measuring Competitive Balance in Formula One Racing, Ilmenau Economics Discussion Papers, Vol. 25, No. 121, (2019). https://www.db-thueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00044434/Diskussionspapier_Nr_121.pdf
4. Cheng, W., Rademaker, M., De Baets, B., Hüllermeier, E.: Predicting Partial Orders: Ranking with Abstention. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science(), vol 6321. Springer, Berlin, Heidelberg. doi:org/10.1007/978-3-642-15880-3_20
5. Conover, W.J.:The rank transformation—an easy and intuitive way to connect many nonparametric methods to their parametric counterparts for seamless teaching introductory statistics courses. WIREs Comp Stat, 4: 432-438. (2012). doi: 10.1002/wics.1216

6. Criado, R., García, E., Pedroche, F., Romance, M.: A new method for comparing rankings through complex networks: Model and analysis of competitiveness of major European soccer leagues Chaos: An Interdisciplinary Journal of Nonlinear Science. 23, 043114 (2013) doi:10.1063/1.4826446
7. Diaconis, P., Graham, R.L.: Spearman's Footrule as a Measure of Disarray. J. R. Stat. Soc. B Met. 39, 262–268 (1977). doi:10.1111/j.2517-6161.1977.tb01624.x
8. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E.: Comparing Partial Rankings. SIAM J. Discrete Math. 20, 628–648 (2006). doi:10.1137/05063088X
9. Garcia-del-Barrio, P., Reade, J.J.: Does certainty on the winner diminish the interest in sport competitions? The case of formula one. Empir Econ 63, 1059–1079 (2022). https://doi.org/10.1007/s00181-021-02147-8
10. Judde, C., Booth, R., Brooks, R.: Second Place Is First of the Losers: An Analysis of Competitive Balance in Formula One. Journal of Sports Economics, 14(4), 411–439 (2013) https://doi.org/10.1177/1527002513496009
11. Kendall, M.G.: A New Measure of Rank Correlation. Biometrika 30, 81–89 (1938). doi:10.2307/2332226
12. Kendall, M.G.: *Rank Correlation Methods*, 4th ed.; Griffin: London, UK, 1970.
13. Kendall, M.G., Babington-Smith, B.: The Problem of m Rankings. Ann. Math. Stat. 10, 275–287 (1939). doi:10.1214/aoms/1177732186
14. Kendall, M. G., Hill, A. B.: The Analysis of Economic Time-Series-Part I: Prices. Journal of the Royal Statistical Society. Series A (General), 116(1), 11–34.(1953) doi:10.2307/2980947
15. Knuth, D. The art of computer programming. Vol 3. (2nd. Ed) 1988. Addison Wesley Longman
16. Krauskopf, T., Langen, M., Bünger, B.: The search for optimal competitive balance in formula one, CAWM Discussion Papers 38, University of Münster, Münster Center for Economic Policy (MEP). 2010.
17. Manasis, V., Ntzoufras, I., Reade, J. J.: Competitive balance measures and the uncertainty of outcome hypothesis in European football, IMA Journal of Management Mathematics, (33), 1, 19–52 (2022) doi.org/10.1093/imaman/dpab027
18. Mastromarco, C., Runkel, M.: Rule changes and competitive balance in Formula One motor racing, Applied Economics, 41(23), 3003–3014 (2009) doi:10.1080/00036840701349182
19. Owen, P.D., Ryan, M., Weatherston, C.R.: Measuring Competitive Balance in Professional Team Sports Using the Herfindahl-Hirschman Index. Rev. Ind. Organ. 31, 289–302 (2007). doi.org/10.1007/s11151-008-9157-0
20. Pedroche, F., Criado, R., García, E., Romance, M., Sánchez, V.E. Comparing series of rankings with ties by using complex networks: An analysis of the Spanish stock market (IBEX-35 index). Netw. Heterog. Media. 10, 101–125 (2015). doi:10.3934/nhm.2015.10.101
21. Pedroche, F., Conejero, J.A. Corrected Evolutive Kendall's τ Coefficients for Incomplete Rankings with Ties: Application to Case of Spotify Lists. Mathematics, 8(10), 1828, (2020). doi:10.3390/math8101828
22. Shapiro, S. S., Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591–611 (1965) doi.org/10.2307/2333709
23. https://www.fia.com
24. https://www.fia.com/events/fia-formula-one-world-championship/season-2012/classifications
25. Veček, N., Mernik, M., Črepinšek, M: A chess rating system for evolutionary algorithms: A new method for the comparison and ranking of evolutionary algorithms, Information Sciences, 277, 656-679 (2014). doi:10.1016/j.ins.2014.02.154.

26. Yoo, Y.; Escobedo, A.R.; Skolfield, J.K.: A new correlation coefficient for comparing and aggregating non-strict and incomplete rankings. Eur. J. Oper. Res. 285, 1025–1041 (2020). doi.org/10.1016/j.ejor.2020.02.027
27. Zimbalist, A. S.: Competitive Balance in Sports Leagues: An Introduction. Journal of Sports Economics, 3(2), 111–121. (2002). doi:10.1177/152700250200300201
28. Primera División de España. Wikipedia. Retrieved 5/5/2023. https://es.wikipedia.org/wiki/Primera_DivisindeEspaa
29. Premier League. Retrieved 5/5/2023. https://www.ceroacero.es/edicao.php?id_edicao=175435

Part VIII

Mathematical Models in Population Dynamics

Constructing exact numerical solutions and nonstandard difference schemes for second order linear delay differential equations

Carlos Julio Mayorga[1,2], María Ángeles Castro[1], Antonio Sirvent[1], and Francisco Rodríguez[1,3]

[1] Dept. Applied Mathematics, University of Alicante, Apdo. 99, Alicante, Spain,
[2] Central University of Ecuador, Quito, Ecuador,
[3] Multidisciplinary Institute for Environmental Studies (IMEM), University of Alicante, Apdo. 99, Alicante, Spain.

Abstract. Previous works have presented exact numerical solutions for first order delay differential equations and systems. For linear systems with non-commuting matrix coefficients, an expression involving infinite sums was truncated to produce numerical approximations, defining nonstandard numerical schemes of arbitrary order. It was suggested that it could be possible to reduce these infinite sums to finite expressions for particular problems, depending on the specific structures of the coefficient matrices. In this communication, we discuss how this approach can be effective for some second order linear delay differential problems, by transforming them to first order systems, in general with non-commuting matrix coefficients. We show that, taking advantage of the simple structures of the resulting matrix coefficients, finitely expressed exact numerical solutions can be obtained, and high order numerical schemes can be derived from them that are dynamically consistent with the asymptotic stability of the continuous solutions.

Keywords: second order initial-value delay differential problems, exact schemes, nonstandard numerical methods, dynamic consistency

1 Introduction

Exact numerical schemes have been proposed for a wide variety of differential equations and systems, being useful to suggest basic rules to construct nonstandard finite difference (NSFD) methods that show good dynamic consistency properties [1]. However, the construction of exact numerical solutions for delay differential equations (DDE) has been much limited. The first partial result was due to Garba *et al.* [2, 3], who proposed a method that was exact in the first delay interval, and then provided the basis to construct a second order NSFD method, for the retarded scalar initial value problem

$$x'(t) = \alpha x(t) + \beta x(t-\tau), \qquad t > 0, \tag{1}$$

$$x(t) = \varphi(t), \qquad -\tau \le t \le 0. \tag{2}$$

García *et al.* extended this result in [3], by constructing an exact numerical solution for problem (1)-(2) in the whole domain, and proposing a family of NSFD methods of arbitrary order based on it, and then in [4], where they obtained similar results for the coupled system

$$X'(t) = AX(t) + BX(t-\tau), \qquad t > 0, \tag{3}$$
$$X(t) = F(t), \qquad -\tau \le t \le 0, \tag{4}$$

with $X(t)$ and $F(t)$ being d-dimensional real vector functions, and A and B commuting $d \times d$ real matrices.

The general case of non-commuting matrix coefficients was considered in [5], where an expression for the numerical solution of (3)-(4) involving infinite sums was presented, suggesting that it could be reduced to finitely expressed closed forms for particular problems. Based on the expected form of the exact schemes obtained for retarded delay equations, an exact numerical solution for the scalar neutral equation

$$x'(t) - \gamma x'(t-\tau) = \alpha x(t) + \beta x(t-\tau), \qquad t > 0, \tag{5}$$

was also provided in [6].

In this communication we discuss how the results in [5] can be used to obtain both, exact numerical solutions and NSFD schemes based on them, for second order DDEs, and present some examples for the problem

$$x''(t) = -\alpha^2 x(t) + \beta x(t-\tau), \qquad t > 0, \tag{6}$$
$$x(t) = f(t), \qquad -\tau \le t \le 0, \tag{7}$$

which has been dealt with in [7].

2 Results and discussion

Theorem 1 in [5] provided an expression involving infinite sums for the exact solution of the general problem (3)-(4), as recalled in the next lemma.

Lemma 1. *Consider problem (3)-(4) with $F \in C^1[-\tau, 0]$. Let $I \in \mathbb{R}^{d\times d}$ be the identity matrix, $C = A^{-1}B$, and assume A and $I + C$ invertible. Write*

$$Q_1(t) = \left(\mathrm{e}^{At} - I\right)(I + C), \qquad Q_m(t) = \int_0^t \mathrm{e}^{A(t-s)} B Q_{m-1}(s)\, ds, \quad m > 1,$$

and define the matrix constants $K^m_{r,p}$, $\forall m \ge 1$, by

$$K^m_{r,p} = 0,\ r < p; \qquad K^m_{r,0} = A^r,\ r \ge 0;$$
$$K^m_{r+1,p} = AK^m_{r,p} + BK^m_{r,p-1},\ 1 \le p \le m-1; \qquad K^m_{r+1,m} = K^m_{r,m-1}B.$$

Let $X(t) = F(t)$ for $t \in [-\tau, 0]$ and

$$X(t+h) = e^{Ah}X(t) + \sum_{p=1}^{m} G_p(h)X(t-p\tau) + \int_0^h Q_m(h-s)(I+C)^{-1}CF'(t-m\tau+s)\, ds,$$

for $m \geq 1$ and $(m-1)\tau \leq t < t+h \leq m\tau$, with $G_p(h) = \sum_{r=p}^{\infty} \frac{h^r}{r!} K_{r,p}^m$. Then, $X(t)$ is a well-defined function satisfying (3) and (4).

An exact numerical scheme follows immediately by considering a mesh of amplitude $h = \tau/N$, $N \geq 1$, denoting $t_n \equiv nh$ and $X_n \equiv X(t_n)$, for $n \geq -N$, and computing $X_n = F(t_n)$, for $-N \leq n \leq 0$, and

$$X_{n+1} = e^{Ah} X_n + \sum_{p=1}^{m} G_p(h) X_{n-pN} + \int_0^h Q_m(h-s)(I+C)^{-1} C F'(t_{n-mN}+s) ds, \tag{8}$$

for $(m-1)\tau \leq nh < m\tau$ and $m \geq 1$.

A linear high order scalar DDE can be converted in the usual way into a system as given in (3), with special forms of the matrices A and B. Thus, if these special structures would allow to derive explicit expressions for the matrices $K_{r,p}^m$ in Lemma 1, it could be possible to evaluate the sums defining $G_p(h)$, and so obtaining a finitely expressed numerical scheme as given in (8).

In the case of the second order DDE (6), the corresponding coefficient matrices are given by

$$A = \begin{pmatrix} 0 & a \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix},$$

with $X(t)$ and $F(t)$ in (3)-(4) given by

$$X(t) = \left(x'(t)\ x(t)\right)^T, \quad F(t) = \left(f'(t)\ f(t)\right)^T.$$

Using the special forms of these matrix coefficients, it has been shown in [7] that the matrices $G_p(h)$ can in fact be evaluated and expressed in terms of Bessel and hypergeometric functions, confirming that the strategy of using the results in [5] to tackle scalar higher order problems can yield positive results.

The exact solution for problem (6)-(7) with coefficients $\alpha = 1.5$, $\beta = 0.75$, delay $\tau = 1$ and initial function $f(t) = (t+1)$, computed using an analytical expression that was presented in [8], and the exact numerical solution derived from (8), as given in [7], are presented in Figure 1.

The expressions for the exact numerical solutions of DDEs presented so far include the presence of an integral term, which seems unavoidable given the infinite dimensional character of delay equations. This integral term depends on the initial function, and it can be analytically computed only in particular cases, so that in general it has to be numerically approximated. However, as shown in [3–5, 7], two families of NSFD schemes can be defined by computing the exact numerical solution in the first M intervals and then dropping the integral term, computing the finite sum in (8) either in full ($\mathcal{F}_M$ schemes) or truncated up to the M term ($\mathcal{T}_M$ schemes), both type of schemes being of the same order.

In the case of problem (6)-(7), the corresponding $\mathcal{F}_M$ and $\mathcal{T}_M$ schemes can be shown to be of order $2M$, which is confirmed by numerical experiments. Table 1 shows values of maximum absolute errors, E_h, for the numerical solutions of the problem in Figure 1 computed with $\mathcal{T}_M$ schemes of different orders and three mesh sizes, and computational orders estimated as $\left(\ln E_h - \ln E_{h/2}\right) / \ln 2$.

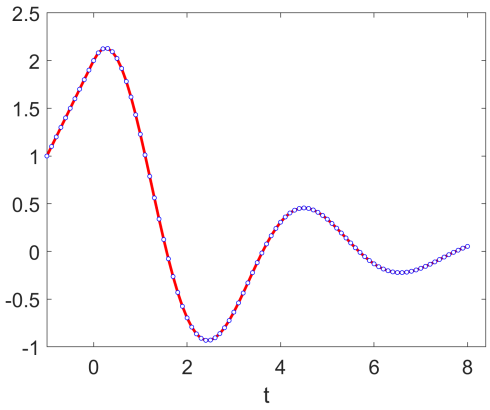
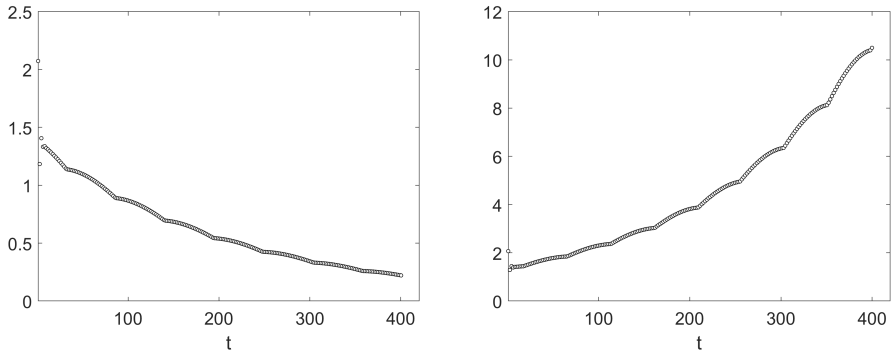


Fig. 1: Continuous solution (line) and exact numerical solution computed with the exact scheme (points) for problem (6)-(7) with coefficients $\alpha = 1.5$, $\beta = 0.75$, delay $\tau = 1$ and initial function $f(t) = (t + 1)$.

It can be proved that the family of $\mathcal{F}_M$ schemes preserve the delay-dependent stability of the continuous problem. Although a similar proof might not be possible for the truncated $\mathcal{T}_M$ schemes, they also seem to possess good dynamic consistency properties, as shown by numerical experiments.

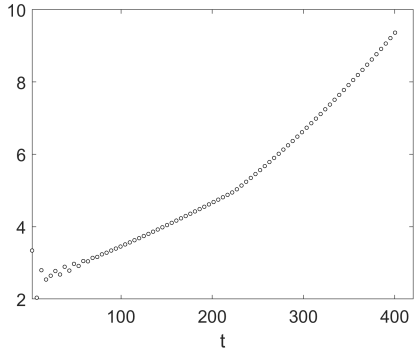
Equation (6) with coefficients as in Figure 1, $\alpha = 1.5$ and $\beta = 0.75$, is stable without delay, and there are three stability switches as the delay is increased, becoming unstable at $\tau_1 \approx 1.81$, stable at $\tau_2 \approx 5.13$, and remaining unstable from $\tau_3 \approx 5.44$, which is well reproduced by $\mathcal{T}_M$ schemes, as shown in Figure 2.

It is expected that in future works the approach presented here could also be applied to more general second order DDEs, and hopefully also to higher order delay equations.

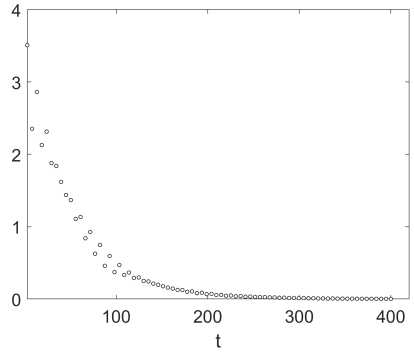


(a) $\tau = 1.80$.

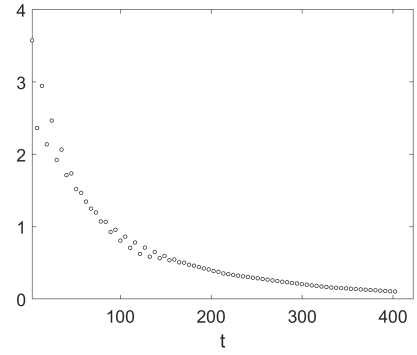
(b) $\tau = 1.83$.



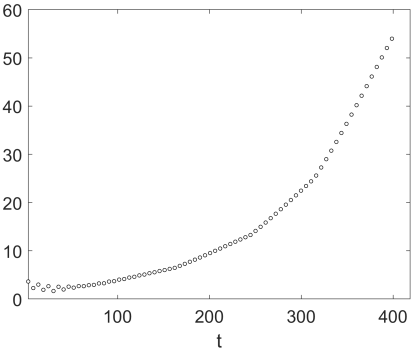
(c) $\tau = 5.10$.



(d) $\tau = 5.30$.



(e) $\tau = 5.40$.



(f) $\tau = 5.50$.

Fig. 2: Dynamic consistency behaviour of the $\mathcal{T}_M$ method ($M = 3$). Trends in maximum absolute values, in each interval of τ amplitude for problem (6)-(7) with the same coefficients and initial function as in Figure 1, show stable and unstable behaviours with increasing values of delays, from (**a**) to (**f**), consistent with the delay-dependent changes in stability of the continuous solution.

Table 1: Maximum absolute errors (upper values) and order estimates (lower values) for numerical solutions of problem (6)-(7) as in Figure 1, computed with $\mathcal{T}_M$ schemes of three different orders ($M = 2$, $M = 3$, and $M = 4$), for different mesh sizes.

$\mathbf{h} = \tau/\mathbf{N}$	$\mathbf{M = 2}$	$\mathbf{M = 3}$	$\mathbf{M = 4}$
$h = 0.025$	2.32×10^{-9}	2.59×10^{-14}	1.69×10^{-19}
	-	-	-
$h = 0.05$	3.71×10^{-8}	1.66×10^{-12}	4.31×10^{-17}
	3.9992	5.9994	7.9996
$h = 0.1$	5.93×10^{-7}	1.06×10^{-10}	1.10×10^{-14}
	3.9992	5.9992	7.9992

References

1. Mickens, R.E. *Nonstandard Finite Difference Schemes: Methodology and Applications*; World Scientific: Singapore, 2021.
2. Garba, S.M.; Gumel, A.B.; Hassan, A.S.; Lubuma, J.M.-S. Switching from exact scheme to nonstandard finite difference scheme for linear delay differential equation. *Appl. Math. Comput.* **2015**, *258*, 388–403.
3. García, M.A.; Castro, M.A.; Martín, J.A.; Rodríguez, F. Exact and nonstandard numerical schemes for linear delay differential models. *Appl. Math. Comput.* **2018**, *338*, 337–345.
4. García M.A.; Castro M.A.; Martín J.A.; Rodríguez F. Exact and nonstandard finite difference schemes for coupled linear delay differential systems. *Mathematics.* **2019**, *7*, 1038.
5. Castro, M.A.; Sirvent, A.; Rodríguez, F. Nonstandard finite difference schemes for general linear delay differential systems. *Math. Meth. Appl. Sci.* **2021**, *44*, 3985–3999.
6. Mayorga, C.J.; Castro M.A.; Sirvent, A.; Rodríguez F. On the construction of exact numerical schemes for linear delay models. *Mathematics.* **2023**, *11*, 1836.
7. Castro, M.A.; Mayorga, C.J.; Sirvent, A.; Rodríguez, F. Exact numerical solutions and high order nonstandard difference schemes for a second order delay differential equation. *Math. Meth. Appl. Sci.* **2023**, DOI: 10.1002/mma.9540.
8. Rodríguez, F.; Roales, M.; Martín, J.A. Exact solutions and numerical approximations of mixed problems for the wave equation with delay. *Appl. Math. Comput.* **2012**, *219*, 3178–3186.

Modeling interference on interference competition models

M. C. Vera[1], M. Marvá[1], R. Escalante[1], and V. García[1]

Universidad de Alcalá,
Departamento de Física y Matemáticas,
Alcalá de Henares 28807, Spain.
carmen.vera@uah.es

Abstract. Paradoxically, prior interference competition models did not account for the effect of both intra- and inter-species interference. We do so by adapting the works of Beddington [2] and DeAngelis [7] on predator-prey models to the classical Gause interference competition model [8]. The established theory estates that global species coexistence (i.e., regardless of the initial amount of individuals) is only possible, roughly, when the ratio of the inter-species effects over the intra-species effects is less than 1. This feature was intended to support the Gause's Competitive Exclusion Principle (two species competing for the same resource can not -hardly- coexist) and the Coexistence Paradox (this hypothesis is at odds with Nature). We have found that taking into account intra-species interference in competition allows competing species to global coexistence even if the above mentioned ratio is larger than 1. This feature was not allowed in previous works on interference competition that introduced herd-type behavior [1], [3], [13], the time spent in competition [4] or group defense [5]. We have also found multi-stability scenarios not allowed by the classical model [8] but found in the above mentioned references which, in turn, here are feasible in a wider range of the parameters space due exclusively to intra-species interference when competing heterospecifics. Therefore, accounting for interference contributes to unveil the Paradox of coexistence.

Keywords: interference competition, interfering time, species competition

1 Introduction

The Competitive Exclusion Principle [9] states, roughly, that two species that compete for the same resource can not coexist. Two classical works support this Principle: the Gause [8] (Lotka-Volterra like) competition model was derived for ordinary differential equations based on laboratory experiments with *Paramecium*. The Leslie-Gower difference equations competition model [12] was inspired on the famous experiments with *Tribolium* carried out by Park and collaborators.

Both models share a handful of features. Let $x_i(t)$ be the number of individuals of species $i = 1, 2$ at time t. The Gause model reads as

$$x_i'(t) = r_i x_i(t) - a_{ii} x_i^2(t) - a_{ij} x_i(t) x_j(t), \tag{1}$$

while the Lelie-Gower model is

$$x_i(t+1) = \frac{r_i x_i(t)}{1 + a_{ii}x_i(t) + a_{ij}x_j(t)}, \tag{2}$$

where r_i stands for the intrinsic growth rate, a_{ii} for the intra-species competition coefficient, while a_{ij} measures the competitive effect of species j on species i.

The nullclines of both models are straight lines. Also, assuming that the trivial equilibrium point $E_0^* = (0, 0)$ is unstable, both models allow for four different competition outcomes: either species 1 or species 2 wins regardless of the initial values, global coexistence, or priority effects: one species will go extinct depending on the initial amount of individuals. Interestingly, the competition outcome depends essentially[1] on the same combination of parameters

$$\text{Gause: } \frac{a_{ij}}{a_{jj}}\frac{r_j}{r_i}, \qquad\qquad \text{Leslie-Gower: } \frac{a_{ij}}{a_{jj}}\frac{r_j - 1}{r_i - 1}. \tag{3}$$

The above expressions, that we denote indistinctly by c_{ij}, are interpreted as follows: *forget for a moment* of the r's ratio. Coefficient $c_{ij} < 1$ means that the effect of species j on species i is softer than the effect of species j on species j. Recall that species j would survive in the absence of species i. Thus, $c_{ij} < 1$ means that species j can not drive species i to extinct. Note that condition $a_{ij}/a_{jj} < 1$ can be reversed by multiplying by the ratio of the intrinsic growth rates, i.e., being not harmful enough can be compensated by a sufficiently larger reproduction rate, which must be taken into account [11].

The outcomes of both the Gause (1) and the Leslie-Gower model (2) are summarized in Figure 1 in terms of the above defined c_{ij}, $i \neq j$ coefficients, the so-called *competitive strengths* [4]. In terms of c_{ij}, coexistence or extinction depends on the balance between intra- and inter-species competition [14], [15].

Coexistence seems to be much more common in Nature than species exclusion, which is at odds with the Competitive Exclusion Principle and gives rise to the Paradox of Coexistence. Ecologist have done many work to explain this contradiction (find a recent review in [4]). However, from the deterministic models viewpoint not too much work has been done apart from the recent works assuming herd behavior [1], [3], [13], accounting for the time spent on inter-species competition (individuals interference) [4], and group defense [5].

In this work we set a model that accounts for intra-species interference in competing species in Section 2. Then we summarize the possible outcomes of the model in Section 3 and briefly discuss the results in Section 4.

[1] If λ_i are the eigenvalues of the Jacobian at an equilibrium point E^*, it is asymptotically stable if $\lambda_i < 0$ ($|\lambda_i| < 1$, respt.) for differential equations (difference equations, resp.)

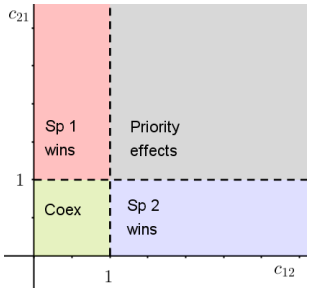


Fig. 1: Outcomes of the classical competition model (1) as function of the competitive stregths c_{12} and c_{21}.

2 The model

Beddington [2] and DeAngelis [7] modelized the effect of interference between predators when hunting preys. This effect can be easily adapted to competing species.

Let us make a comment to clarify the effect we are accounting for. Note that we deal with interference competition (also known as contest competition) which means that, in contrast to exploitative (or scramble) competition, after disputing for a resource one of the competitors will take the whole resource they are competing for. Two heterospecifics may compete for a resource, and the effect of the resulting interference was addressed in [4]. Instead, here, we account for the mutual interference between conspecifics that compete for a resource with a heterospecific. This effect is different from that due to logistic behavior, i.e., that of two individuals of the same species competing for a resource in the absence of an individual of the other species.

The complete competition interference model reads as follows

$$x_i' = r_i x_i - a_{ii} x_i^2 - \frac{a_{ij} x_i x_j}{1 + a_i x_i + \tilde{a}_j (x_j - 1)} \qquad i \neq j, \qquad i, j = 1, 2. \tag{4}$$

The model analyzed in [4] is (4) with $\tilde{a}_j = 0$. Here, instead, we are aimed to understand the *net effect* of intra-species interference when competing heterospecifs. Thus, we consider that the inter-species interference in competition is negligible, so that $a_i = 0$ and system (4) becomes

$$x_i' = r_i x_i - a_{ii} x_i^2 - \frac{a_{ij} x_i x_j}{1 + \tilde{a}_j (x_j - 1)} \qquad i \neq j, \qquad i, j = 1, 2. \tag{5}$$

The analysis of the complete model (4) is beyond the scope of this contribution and will be available somewhere. We rewrite system (5) accordingly to $u_i = a_{ii} x_i / r_i$, $c_{ij} = a_{ij} r_j / (r_i a_{jj})$, and $K_i = r_i / a_{ii}$, that yields

$$u_i' = r_i \left(u_i - u_i^2 - \frac{c_{ij} u_i u_j}{1 + \tilde{a}_j (K_j u_j - 1)} \right), \qquad i \neq j, \qquad i, j = 1, 2. \tag{6}$$

where c_{ij} is the competitive strength as defined in the left hand side of (3), K_i is the carrying capacity, and a_{ii}/r_i is the absolute competition coefficient [6] of species, $i = 1, 2$. It is clear that $\tilde{a}_j = 0$ in (5) yields the classical system (1). Assuming $\tilde{a}_j > 0$ gives rise to a new model, the competition model with Beddington-DeAngelis competitive response.

3 Results

We analyze system (6) by assuming that $0 < \tilde{a}_i < 1$ for $i = 1, 2$. From now on we assume that $r_i > 0$ for $i = 1, 2$, which implies that the trivial equilibrium point $E_0^* = (0, 0)$ is unstable. Thus, in the absence of species j, species i behaves according to the logistic equation

$$u_i' = r_i u_i (1 - u_i), \tag{7}$$

and $u_i^* = 1$ (that is equivalent to $x_i^* = r_i/a_{ii}$, the corresponding carrying capacity) is a global attractor for the solutions of (7).

A first result states that the dynamics of system (5) evolves eventually in the region $[0, 1] \times [0, 1]$:

Proposition 1. *The non-negative cone is invariant for system (6). Besides, any solution of such a system with positive initial values will eventually enter the region* $[0, 1] \times [0, 1]$ *and will not leave it.*

Proof. It follows from direct calculations using the fact that the solution of $u_i' = r_i(u_i - u_i^2)$ upper bounds the solution of the corresponding equation in system (6).

Indeed,

Proposition 2. *The solutions of system (6) converge eventually monotonically to an equilibrium point.*

Proof. It follows from direct calculations computing the conditions stated in [10].

Proposition 3. *Consider the semi-trivial equilibrium points.* $E_1^* = (1, 0)$ *and* $E_2^* = (0, 1)$. *Then,* E_j^* *is locally asymptotically stable if* $c_{ij} > \tilde{c}_{ij}^*$ *for* $i \neq j$, $i, j = 1, 2$ *respectively, where*

$$\tilde{c}_{ij}^* = 1 + \tilde{a}_j(K_j - 1), \qquad i = 1, 2, \quad i \neq j. \tag{8}$$

Proof. The existence of E_i^* follows from direct calculations. The stability conditions follow from a standard analysis of the eigenvalues of the Jacobian matrix.

We next classify the possible outcomes of system (6) in terms of the coefficients of the model. It will turn out that the quantities

$$\Gamma_i = \frac{1 + \tilde{a}_i(K_i - 1)}{\tilde{a}_i K_i} = 1 + \frac{1 - \tilde{a}_i}{\tilde{a}_i K_i}, \qquad i = 1, 2 \tag{9}$$

play a key role. Note that $K_i \geq 1$ since they are carrying capacities. Indeed, we divide the $c_{12}c_{21}$ positive cone into four regions

$$\begin{aligned} R_{coex} &:= \{0 < c_{12} < \tilde{c}_{12}^*,\, 0 < c_{21} < \tilde{c}_{21}^*\} \\ R_{pe} &:= \{\tilde{c}_{12}^* < c_{12},\, \tilde{c}_{21}^* < c_{21}\} \\ R_1 &:= \{0 < c_{12} < \tilde{c}_{12}^*,\, \tilde{c}_{21}^* < c_{21}\} \\ R_2 &:= \{\tilde{c}_{12}^* < c_{12},\, 0 < c_{21} < \tilde{c}_{21}^*\} \end{aligned} \tag{10}$$

that define the possible outcomes of system (6) (see Figure 2).

Proposition 4. *Consider system (6). Then:*

1. *Global coexistence. There exists an equilibrium point in the non-negative cone that is GAS for any* $(c_{12}, c_{21}) \in R_{coex}$.
2. *Priority effects. There exists a saddle equilibrium point in the non-negative cone that is unstable for any* $(c_{12}, c_{21}) \in R_{pe}$. *Indeed,* E_1^* *and* E_2^* *are locally asymptotically stable, and the stable manifold of the positive (componentwise) equilibrium defines the basins of attraction of each semi-trivial equilibrium point.*

Proof. The non-trivial equilibrium points are the solutions to the equation resulting from equating the nullclines of system (6). The number of solutions included in the non-negative cone follows from applying the Descartes' rule of signs to that equation. The stability conditions of the semi-trivial equilibrium points E_1^* and E_2^* were proved in Proposition 3. The stability of the non-triVial equilibrium points follows from the above considerations and Proposition 2.

Proposition 5. *Consider system (6). Then:*

1. *Assume now that* $\Gamma_1 = \Gamma_2$, *that is*

$$\frac{K_2}{K_1} = \frac{\tilde{a}_1(1-\tilde{a}_2)}{\tilde{a}_2(1-\tilde{a}_1)} \tag{11}$$

   Then, it follows that:
   (a) *Species 1 wins:* E_1^* *is GAS whenever* $(c_{12}, c_{21}) \in R_1$.
   (b) *Species 2 wins:* E_2^* *is GAS whenever* $(c_{12}, c_{21}) \in R_2$.
2. *Instead, if* $\Gamma_1 < \Gamma_2$, *that is equivalent to*

$$\frac{K_2}{K_1} < \frac{\tilde{a}_1(1-\tilde{a}_2)}{\tilde{a}_2(1-\tilde{a}_1)} \tag{12}$$

   Then, it follows that:
   (a) *Species 1 wins:* E_1^* *is GAS whenever* $(c_{12}, c_{21}) \in R_1$.
   (b) *Consider the second degree equation on* u_1 *that raises from equating the nullclines of system (6). Then, there exists a curve,* Ψ_+, *arising from solving on* c_{21} *the result of equating to zero the discriminant of the solution of the above-mentioned second degree equation such that*

i. Species 2 wins: E_2^* is GAS whenever

$$\{(c_{12}, c_{21}) \in R_2\} \cap \{(c_{12}, c_{21});\, c_{21} < \Psi_+(c_{12})\} \tag{13}$$

ii. Conditional coexistence in favour of species 2. On the contrary, if

$$\{(c_{12}, c_{21}) \in R_2\} \cap \{(c_{12}, c_{21});\, \Psi_+(c_{12}) < c_{21}\} \tag{14}$$

then E_2^* is locally asymptotically stable and E_1^* unstable. In addition, there exist two equilibrium points in the positive cone, one locally asymptotically stable and one unstable. The latest is a saddle equilibrium point whose stable manifold separates the basins of attraction of E_2^* and the positive (coexistence) equilibrium point.

3. Finally, if $\Gamma_1 > \Gamma_2$, that is equivalent to

$$\frac{K_2}{K_1} > \frac{\tilde{a}_1(1-\tilde{a}_2)}{\tilde{a}_2(1-\tilde{a}_1)} \tag{15}$$

Then, it follows that:

(a) Species 2 wins: E_2^* is GAS whenever $(c_{12}, c_{21}) \in R_2$.

(b) Consider the second degree equation on u_1 that raises from equating the nullclines of system (6). Then, there exists a curve, Ψ_-, arising from solving on c_{21} the result of equating to zero the discriminant of the solution of the above-mentioned second degree equation such that

i. Species 1 wins: E_1^* is GAS whenever

$$\{(c_{12}, c_{21}) \in R_1\} \cap \{(c_{12}, c_{21});\, \Psi_-(c_{12}) < c_{21}\} \tag{16}$$

ii. Conditional coexistence in favour of species 1. On the contrary, if

$$\{(c_{12}, c_{21}) \in R_1\} \cap \{(c_{12}, c_{21});\, c_{21} < \Psi_-(c_{12})\} \tag{17}$$

then E_1^* is locally asymptotically stable and E_2^* unstable. In addition, there exist two equilibrium points in the positive cone, one locally asymptotically stable and one unstable. The latest is a saddle equilibrium point whose stable manifold separates the basins of attraction of E_1^* and the positive (coexistence) equilibrium point. See the right panel of Figure 3.

Proof. When equating the nullclines of system (6) we get a second degree equation for u_1. The solutions of such an equation are the u_1 component of the equilibrium points of system (6). Letting the discriminant of the solution of that equation equal to zero, the curves Ψ_+ and Ψ_- are obtained. These curves bound the regions on the $c_{21}c_{12}$ plane where there are two, one or none equilibrium points (that is, the algebraic equation has either real or complex solutions).

Signs of the coordinates of the equilibrium points are determined by using the Descartes' rule of signs. The number of equilibrium points inside the non-negative cone, in addition with stability of the semi-trivial equilibrium points (Proposition (2)) yield the stability of the non-trivial equilibrium points.

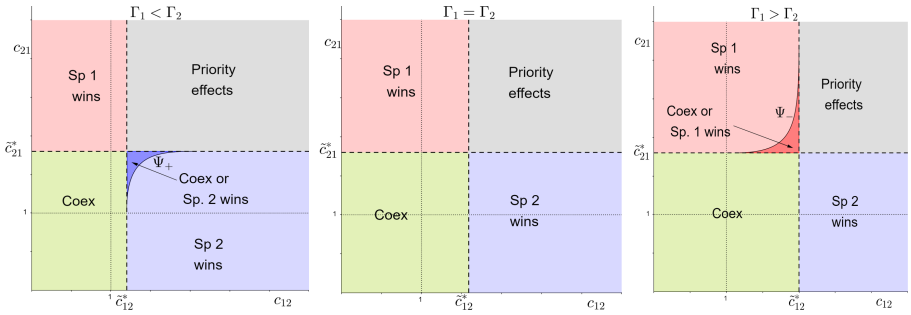


Fig. 2: Competition outcomes of system (6) as function of the competitive strengths c_{12}, c_{21} for increasing values of $\tilde{a_2} = 0.1, 0.3, 0.5$ (from left to right). Other parameter values are: $r_1 = 6.8$, $r_2 = 2$, $K_1 = 5$, $K_2 = 3$, $\tilde{a_1} = 0.2$. The code colour is the same as in Figure 1 except for the dark blue and dark red regions, that represent conditional coexistence in favour of species 2 or 1, respectively. Note $\tilde{a}_1 = 0.2$ is kept fixed in the three figures and $\tilde{a}_2$ varies. As a consequence, $\tilde{c}^*_{21}$ remains the same while $\tilde{c}^*_{21}$ varies accordingly. Fixing $\tilde{a}_2$ and varying $\tilde{a}_1$ would let fixed $\tilde{c}^*_{21}$ and change $\tilde{c}^*_{12}$.

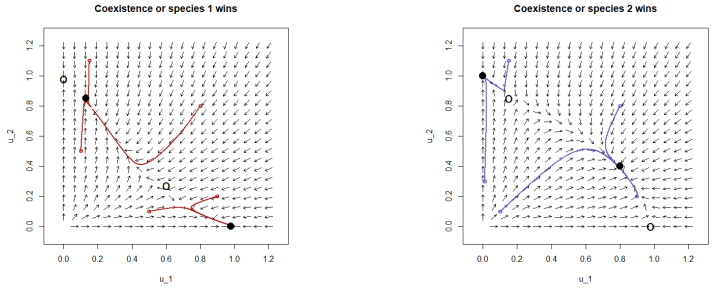


Fig. 3: Conditional coexistence in favour of species 1 (left panel, either species 1 wins or there is species coexistence) or species 2 (right panel, either species 2 wins or there is species coexistence).

4 Discussion

It is intuitive that spending time on interfering with conspecifics softens the competitive effect on heterospecifics. In this work we are able to translate into numbers such a consequences as described in Proposition 4.

Qualitatively, the new dynamical scenarios are equivalent to those found in the recent literature. However, the driving process is different: it is not interference between heteropecifics [4], neither group defense [5] or herd behavior [13] nor the result of a competitive-cooperative balance [15].

A key result is that the global coexistence region is larger than that in [1], [4], [5], [13] or [15]. This feature constitutes an explanation (not unique) of the Paradox of Coexistence. Also, conditional coexistence (items 2b) and 3b) in Proposition 5) expands the scenarios permitted by the classical model allowing for coexistence.

It is of full interest to analyze system (4) for $a_i > 0$ and $\tilde{a}_i > 0$ to fully understand the combined effect of interfering with con- and hetero-specifics when competing.

Acknowledgments: this work is supported by the research project PIUAH22CC-041 funded by the universidad de Alcalá, Spain. M. Marvá is also partially supported by the Programa de Recualificación del Profesorado Universitatio funded by the Ministerio de Universidades (Spain) and Next Generation funds (European Union).

References

1. Ajraldi, V., Pittavino, M., Venturino, E.: Modeling herd behavior in population systems, Nonlinear Anal. Real-World App. 12(4), 2319-2338 (2011). doi:10.1016/j.nonrwa.2011.02.002
2. Beddington, J. R.: Mutual interference between parasites or predators and its effect on searching efficiency. J. Animal Ecol. 44(1) 331-340 (1975). doi:10.2307/3866
3. Bulai, I.M., Venturino, E.: Shape effects on herd behavior in ecological interacting population models. Math. Comp. Simul. 141, 40-55 (2017). doi:10.1016/j.matcom.2017.04.009
4. Castillo-Alvino, H., Marvá, M.: The competition model with Holling type II competitive response to interfering time. J. Biol. Dyn. 14(1), 222-242 (2020). doi:10.1080/17513758.2020.1742392
5. Castillo-Alvino, H., Marvá, M.: Group defense promotes coexistence in interference competition: The Holling type IV competitive response. Math. Comp. Simul. 198, 426-445 (2022). doi:10.1016/j.matcom.2022.02.031
6. CHesson, P.: Mechanisms of maintenance of species diversity Annu. Rev. Ecol. Syst. 31, 343-66 (2000). 10.1146/annurev.ecolsys.31.1.343
7. DeAngelis, D. L., Goldstein, R. A., O'Neill, R. V.: A model for tropic interaction. Ecol, 56(4), 881-892 (1975). doi:10.2307/1936298
8. Gause, G.F.: The Struggle for Existence. Dover Books on Biology, Hafner, Baltimore, New York (1934)
9. Hardin, G.: The competitive exclusion principle, Science 131(3409), 1292-1297 (1960). doi:10.1126/science.131.3409.1292

10. Hirsch, M.W.: Systems of Differential Equations Which Are Competitive or Cooperative: I. Limit Sets. SIAM j. math. anal. 13(2), 167-179 (1989). https://doi.org/10.1137/0513013
11. Kuno E.: Some strange properties of the logistic equation defined with r and K: inherent defect or artifacts? Res. Popul. Ecol. 33, 33-39 (1991). doi:10.1007/BF02514572
12. doi:10.1126/science.131.3409.1292 Leslie, P. H., Gower J. C.: The Properties of a Stochastic Model for for two Competing Species. Biometrika. 47(3/4), 219-234 (1958). doi:10.2307/2333181
13. Melchionda, D., Pastacaldi, E., Perri, C. Venturino., E, Chatoppaday, J.: Social behaviorinduced multistability in minimal competitive ecosystems, J. Theor. Biol. 439, 24-38 (2018). doi:10.1016/j.jtbi.2017.11.016
14. M. Ritchie, Competition and coexistence of mobile animals, in Competition and Coexistence, U. Sommer and B. Worm, eds., Springer-Verlag, Berlin, 2002, pp. 109–131
15. Zhang, Z.: Mututalism or cooperation among competitors promotes coexistence and competitive ability, Ecol. Model. 164, 271-282 (2003). doi:10.1016/S0304-3800(03)00069-3

Delay effects on a classical dryland vegetation model

Ikram Medjahdi[1], Fatima Zohra Lachachi[1], María Ángeles Castro[1], and Francisco Rodríguez[1,2]

[1] Dept. Applied Mathematics, University of Alicante, E-03080, Alicante, Spain,
[2] Multidisciplinary Institute for Environmental Studies (IMEM), University of Alicante, E-03080, Alicante, Spain

Abstract. Vegetation in semiarid areas exhibit spatial discontinuities and complex temporal dynamics, and different models have been proposed in the literature to describe the dynamic and spatial characteristic of semiarid vegetation. Klausmeier proposed a model consisting in a system of two partial differential equations relating plant growth and soil water, showing that, under water limitation, increased infiltration of water by plants could produce characteristic spatial patterns of vegetation found in drylands. This classical model has been analysed and extended in later years in different aspects, including recently the incorporation of a discrete delay to account for the lag between water infiltration into the soil and the following water uptake by plants. In this communication, we consider a more ecologically realistic distributed delay for soil water availability, and analyse its effect on the stability and bifurcations of a mean field Klausmeier-Gray-Scott model.

Keywords: semiarid vegetation, mean field model, Gamma distributed delay, stability and bifurcations

1 Introduction

Klausmeier model [1] is a classical reference in dryland vegetation modelling, as it was able to reproduce spatial patterns similar to those found in different semiarid ecosystems with a relatively simple system of two partial differential equations, relating plant biomass (N) and soil water (W). In its original formulation, it included a two-dimensional spatial region with a diffusion term representing spreading of vegetation and a gradient term, representing unidirectional flow of water down the slope,

$$\begin{cases} \dfrac{\partial W}{\partial T} = V\dfrac{\partial W}{\partial Y} + A - LW - RWN^2, \\ \\ \dfrac{\partial N}{\partial T} = D\left(\dfrac{\partial^2}{\partial X^2} + \dfrac{\partial^2}{\partial Y^2}\right)N + RJWN^2 - MN, \end{cases} \tag{1}$$

with parameters water input (A), water evaporation (L), water velocity downhill (V), mortality (M), basic water intake rate (R), plant dispersal (D), and plant biomass yield per unit water (J).

This classical model has been analysed and extended in later years in different aspects (e.g., [2, 3]). With suitable change of variables, the number of parameters in the nondimensionalized model can be greatly reduced, and, for a diffusive model for both water and vegetation, usually called Klausmeier-Gray-Scottmodel, in the unidimensional case, one has

$$\begin{cases} \dfrac{\partial}{\partial t}w(x,t) = d_1\dfrac{\partial^2}{\partial x^2}w(x,t) + a - w(x,t) - w(x,t)n^2(x,t), \\ \\ \dfrac{\partial}{\partial t}n(x,t) = d_2\dfrac{\partial^2}{\partial x^2}n(x,t) + w(x,t)n^2(x,t) - mn(x,t), \end{cases} \tag{2}$$

with rescaled variables water (w) and vegetation biomass (n), and rescaled parameters water input (a) and mortality (m), where d_1 and d_2 are diffusion coefficients.

The term $w(x,t)n(x,t)$ represents the amount of water infiltrating into soil increased by the presence of vegetation. In [4], a discrete delay was incorporated into this term, $w(x,t-\tau)n(x,t-\tau)$, to account for the expected lag between water infiltration into the soil and the subsequent water uptake by plants.

In this communication, we focus on the dynamics of a mean field Klausmeier-Gray-Scott model with a more ecologically realistic distributed delay for soil water availability. By considering a Gamma distributed delay kernel, we use the so called *linear chain trick* [5, 6], which allows to convert a delay differential system into an ordinary differential system of higher dimension.

2 Results and discussion

We consider the following non-spatial Klausmeier–Gray–Scott model with distributed delay in the form of a weak Gamma kernel,

$$\begin{cases} \dfrac{dw(t)}{dt} = a - w(t) - w(t)n(t)^2, \\ \\ \dfrac{dn(t)}{dt} = n(t)\int_{-\infty}^{t} g_\alpha(t-s)w(s)n(s)ds - mn(t), \end{cases} \tag{3}$$

where $g_\alpha(t) := \alpha e^{-\alpha t}$, with $\alpha > 0$. Since the expected value of $g_\alpha(t)$ is $1/\alpha$, the parameter α represents the inverse of the mean delay.

Defining the new variable

$$z(t) = \int_{-\infty}^{t} g_\alpha(t-s)w(s)n(s)ds,$$

system (3) can be transformed into a system of three ordinary differential equations,

$$\begin{cases} \dfrac{dw(t)}{dt} = a - w(t) - w(t)n(t)^2, \\ \dfrac{dn(t)}{dt} = n(t)z(t) - mn(t), \\ \dfrac{dz(t)}{dt} = \alpha w(t)n(t) - \alpha z(t). \end{cases} \tag{4}$$

Systems (2) and (3), or equivalently (4), have the same equilibrium points (w_*, n_*), a stable trivial equilibrium $P_0 = (a, 0)$, corresponding to bare soil, a double equilibrium $P_1 = (m, 1)$ when $a = 2m$, which is a saddle-node bifurcation point, and two different equilibria with positive vegetation when $a > 2m$, P_2 and P_3, with $P_2 < P_3$,

$$P_2 = \left(\frac{a + \sqrt{a^2 - 4m^2}}{2}, \frac{2m}{a + \sqrt{a^2 - 4m^2}} \right),$$

which is unstable, and

$$P_3 = \left(\frac{a - \sqrt{a^2 - 4m^2}}{2}, \frac{2m}{a - \sqrt{a^2 - 4m^2}} \right),$$

whose stability depends on the parameters a, m, and also α in the model with delay.

It can be shown that when $a > 2m$, so that the equilibrium P_3 exists, in the model without delay this positive vegetation equilibrium is stable if $m < 2$ or if $m > 2$ and $a > \frac{m^2}{\sqrt{m-1}}$ (Fig. 1, top).

In the model given by systems (3) or (4), the presence of a distributed delay may reduce the region of stability of P_3, resulting in the degradation of the system to the bare soil state in conditions where vegetation could be sustained without delay. In this case, when $m > 2$, there is a region delimited by the curves

$$a = \frac{m^2}{\sqrt{m-1}}$$

and

$$a = \frac{m\left(1 + L^2\right)}{L},$$

with

$$L = \frac{\sqrt{2m} + \sqrt{2m-4}}{2},$$

where the stability depends on the value of the delay parameter α (Fig. 1, bottom).

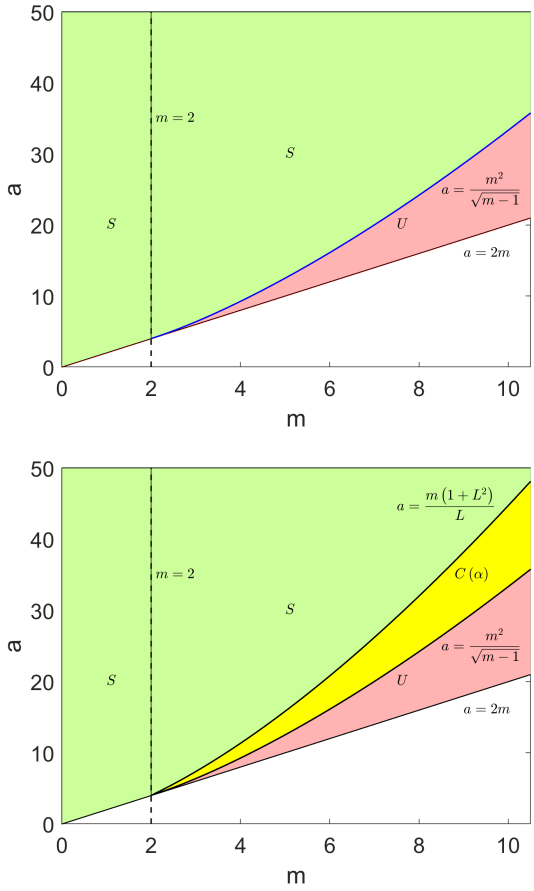
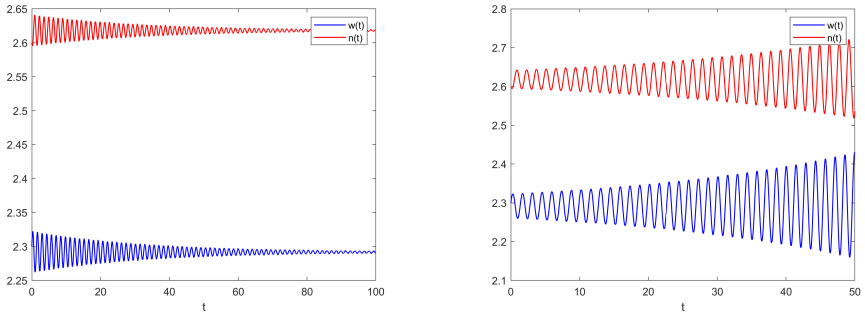
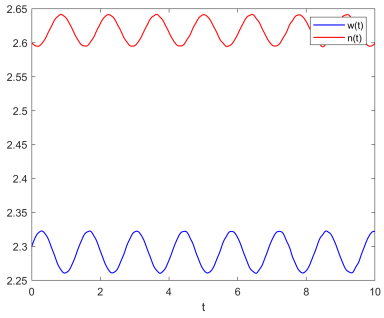


Fig. 1: Regions where the higher positive vegetation equilibrium is stable (S, green) or unstable (U, red) in terms of mortality, m, and water input, a. Top: Model without delay. Bottom: Model with delay (stability in the yellow region depends on the distributed delay parameter α).



(a) $\alpha = 1.1\alpha_c$ (b) $\alpha = 0.9\alpha_c$



(c) $\alpha = \alpha_c$

Fig. 2: Stable (**a**), unstable (**b**), and periodic behaviours (**c**) of the mean field Klausmeier-Gray-Scott model dynamics depending on the distributed delay parameter α.

In this case, there is a critical value of the parameter α,

$$\alpha_c = \frac{2mn_*^2 - (n_*^2+1)^2}{n_*^2 - m + 1},$$

where n_* is the equilibrium biomass in P_3, such that the system is stable for $\alpha > \alpha_c$ and unstable for $\alpha < \alpha_c$, with a Hopf bifurcation at the critical value (Fig. 2).

Work in progress includes analysing the effects of a distributed delay on the local equilibria in the spatial Klausmeier-Gray-Scott model, as well as considering different types of kernels that define the distributed delay effects.

References

1. Klausmeier, C.A., Regular and irregular patterns in semiarid vegetation, *Science*, 284: 1826–1828, 1999.
2. van der Stelt, S., Doelman, A., Hek, G., Rademacher, J.D.M. Rise and fall of periodic patterns for a generalized Klausmeier–Gray–Scott Model. *Journal of Nonlinear Science*, 23: 39–95, 2013.
3. Wang, X., Zhang, G. The influence of infiltration feedback on the characteristic of banded vegetation pattern on hillsides of semiarid area. *PLoS ONE*, 14:e0205715, 2019.
4. Li J, Sun G-Q, Guo Z-G. Bifurcation analysis of an extended Klausmeier–Gray–Scott model with infiltration delay, *Studies in Applied Mathematics*, 148: 1519–1542, 2022.
5. Cushing, J.M. *Integrodifferential Equations and Delay Models in Population Dynamics*. Springer, Heidelberg 1977.
6. Smith, H. *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Springer, New York, 2011.

Part IX

Recent Advances in the Approximation of Matrix Functions

On the use of Euler polynomials to approximate the matrix cosine

J. M. Alonso[1], E. Defez[2], J. Ibáñez[2], and J. Sastre[3]

[1] Instituto de Instrumentación para Imagen Molecular,
[2] Instituto de Matemática Multidisciplinar,
[3] Instituto Universitario de Telecomunicación y Aplicaciones Multimedia.
Universitat Politècnica de València
Camino de Vera s/n, 46022, Valencia. Spain
jmalonso@dsic.upv.es, edefez@imm.upv.es, jjibanez@dsic.upv.es,
jsastrem@upv.es

Abstract. This paper presents an Euler polynomials-based method to approximate the matrix cosine. According to two different series expansions in terms of the Euler polynomials, algorithms devoted to the matrix cosine computation have been developed. Numerical experiments have been performed, where both algorithms have been compared with a code based on Padé approximants.

Keywords: Matrix cosine, Euler polynomials, matrix functions.

1 Introduction

Matrix functions have proven to be an efficient tool in many applications, such as image denoising [1], neural networks [2], or reduced order models [3], [4, pp. 275–303], among others. Thus, its study is an area of applied mathematics that has been remarkably developed in recent years.

Among the different matrix functions, trigonometric ones must be highlighted. Their computation has received significant attention in the last decades due to their usefulness in the solution of systems of partial differential problems. Recently, several state-of-the-art algorithms have been provided for computing these matrix functions [5–8], in particular for the matrix cosine function. Furthermore, the generalization of some well-known classical special functions to the matrix environment has been a very active area of research for decades. The extension to the matrix framework of the orthogonal Laguerre, Hermite, Chebyshev, or Jacobi polynomials have proven to be a valuable tool in various fields of engineering, statistics, physics and telecommunications.

2 Euler polynomials and matrix cosine approximation

Euler polynomials $E_n(x)$ are defined in [9] as the coefficients of the generating function

$$g(x,t) = \frac{2e^{tx}}{e^t+1} = \sum_{n\geq 0} \frac{E_n(x)}{n!} t^n \ , \ |t| < \pi. \tag{1}$$

They have the explicit expression

$$E_n(x) = \sum_{k=0}^{n} \binom{n}{k} \frac{\mathcal{E}_k}{2^k} \left(x - \frac{1}{2}\right)^{n-k}, \tag{2}$$

where the *Euler numbers* are defined by $\mathcal{E}_n = 2^n E_n(1/2)$, satisfying the explicit expression

$$\mathcal{E}_{2n} = 1 - \sum_{k=1}^{n} \binom{2n}{2k-1} \frac{2^{2k}(2^{2k}-1)}{2k} \mathcal{B}_{2k} \ , \ \mathcal{E}_{2n+1} = 0, n \geq 0, \tag{3}$$

and $\mathcal{B}_i$ is the $i-$th *Bernoulli number*.

In [10], Euler polynomials were generalized to the matrix environment. Thus, for a matrix $A \in \mathbb{C}^{r\times r}$, we define the n-th Euler polynomial by the expression

$$E_n(A) = \sum_{k=0}^{n} \binom{n}{k} \frac{\mathcal{E}_k}{2^k} \left(A - \frac{1}{2}I\right)^{n-k}. \tag{4}$$

These matrix polynomials appear in the series expansion

$$e^{At} = \frac{e^t + 1}{2} \sum_{n\geq 0} \frac{E_n(A)t^n}{n!} \ , \ |t| < \pi. \tag{5}$$

Moreover, the use of expansion (5) to approximate the matrix exponential function with satisfactory results in terms of accuracy and computational cost can be found in [10]. Given a matrix $A \in \mathbb{C}^{r\times r}$, and using expression (5), we obtain

$$\cos(A) = \frac{1}{2}(\cos(1) + 1) \sum_{n\geq 0} \frac{(-1)^n E_{2n}(A)}{(2n)!} - \frac{1}{2}\sin(1) \sum_{n\geq 0} \frac{(-1)^n E_{2n+1}(A)}{(2n+1)!}. \tag{6}$$

Taylor or Hermite polynomials are even or odd, depending on the parity of the polynomial degree n. However, Euler matrix polynomials do not verify this property. In fact, all Euler polynomials are needed in the development of $\cos(A)$, not just the even-numbered. Be that as it may, it is also possible by operating to obtain an alternative approximation to the matrix cosine where only polynomials of even degree appear, as follows

$$\cos(A) = \cos(1/2) \sum_{n\geq 0} \frac{(-1)^n E_{2n}\left(A + \frac{1}{2}I\right)}{(2n)!}. \tag{7}$$

3 The proposed algorithms

In this work, two algorithms based on the approximations (6) and (7) for the matrix cosine, in combination with the scaling and squaring technique, have been implemented and compared, attempting to choose the most cost-efficient

and accurate alternative. MATLAB implementations of these algorithms correspond respectively to codes called as *cosm_euler_at* (6) and *cosm_euler_et* (7). The degree of the matrix polynomials m employed satisfies $36 \leq m \leq 42$ for *cosm_euler_at* and $30 \leq m \leq 36$ for *cosm_euler_et*. For each matrix to be computed, the most appropriate degree m and the scaling factor s were decided from the Algorithm 3 described in [12].

4 Numerical tests

Both proposed algorithms have been numerically compared with MATLAB function *cosm*, the code in charge of computing the matrix cosine by means of a Padé approximation [11]. In our numerical experiments, we have conformed a test battery composed of the three following sets of matrices:

Set 1: 100 diagonalizable real matrices of dimension 128×128. These matrices are obtained as the result of $A = V \cdot D \cdot V^{-1}$, where D is a diagonal matrix with real and complex eigenvalues and V is an orthogonal matrix such as $V = H/\sqrt{128}$, where H is a Hadamard matrix. In this case, we have that $2.18 \leq \|A\|_2 \leq 225.71$.

Set 2: 100 non-diagonalizable complex matrices of size 128×128. These matrices have been generated as $A = V \cdot J \cdot V^{-1}$, where J is a Jordan matrix with complex eigenvalues whose modules are less than 10 and the algebraic multiplicity is randomly generated between 1 and 5. V is an orthogonal random matrix with elements in the interval $[-0.5, 0.5]$. These matrices fulfills that $83.996 \leq \|A\|_2 \leq 97.806$.

Set 3: 41 matrices from the Matrix Computation Toolbox [13] and 11 from the Eigtool MATLAB Package [14], all of them with an order equal to 128×128. These matrices satisfy that $1 \leq \|A\|_2 \leq 398423$.

All numerical tests were carried out thanks to the MATLAB version 2023a. The percentage of cases in which the relative error incurred in the cosine calculation by our algorithms is lower or greater than that of code *cosm* is given in Table 1. Our two codes improved similarly *cosm* for the matrices integrating the first and third sets. In the case of the second set, the improvement is much more significant if we refer to the function *cosm_euler_at*. These numerical data are also corroborated by those ones shown as graphs. On the one hand, Figures 1a, 1c, and 1e, corresponding to the graphical representation of the relative error committed by all the codes for the three matrix sets, reflect the higher accuracy of the results provided by the two codes proposed. In particular, it can be noticed how the values related to the errors incurred by code *cosm_euler_at* always occupied the lowest part of Figure 1c. On the other hand, Figures 1b, 1d, and 1f, concerning the performance profile, indicate that Euler polynomial-based codes are the most reliable, with their values located at the top of the pictures 1b and 1f. Clearly in these cases, *cosm* provided the poorest results. Regarding the matrices from Set 2, codes *cosm_euler_et* and *cosm* delivered almost identical results, being largely outperformed by function *cosm_euler_et* in the initial part

of the plot. Obviously, it means that it is the most accurate code since, in the 100 percent of the matrices, it offered the closest results to the exact solution. In terms of computational cost, the number of matrix products required for each code is reported in Table 2. As expected among our codes, the highest number of products corresponded to *cosm_euler_at*, since the polynomial to be evaluated as approximation to the matrix cosine is composed of all terms (even and odd). In opposition, the number of multiplications of codes *cosm_euler_et* and *cosm* were smaller and very similar between them. Nevertheless, as detailed in Table 3, *cosm* demanded the highest execution times. Very small difference in time could be appreciated between codes *cosm_euler_et*, which resulted to be the most efficient, and *cosm_euler_at*.

Table 1: Improvement percentage among the different codes.

	Set 1	Set 2	Set 3
$E(cosm) < E(cosm\_euler\_at)$	3.0%	0.0%	20.00%
$E(cosm) > E(cosm\_euler\_at)$	97.0%	100.0%	80.00%
$E(cosm) < E(cosm\_euler\_et)$	6.0%	26.0%	16.36%
$E(cosm) > E(cosm\_euler\_et)$	94.0%	74.0%	83.64%

Table 2: Number of matrix products required by the codes in comparison.

	Set 1	Set 2	Set 3
$P(cosm)$	1129	1100	604
$P(cosm\_euler\_at)$	1377	1300	720
$P(cosm\_euler\_et)$	1170	1100	626

Table 3: Elapsed time, in seconds, in the execution of the three codes.

	Set 1	Set 2	Set 3
$T(cosm)$	0.32	0.56	0.20
$T(cosm\_euler\_at)$	0.14	0.41	0.10
$T(cosm\_euler\_et)$	0.12	0.34	0.09

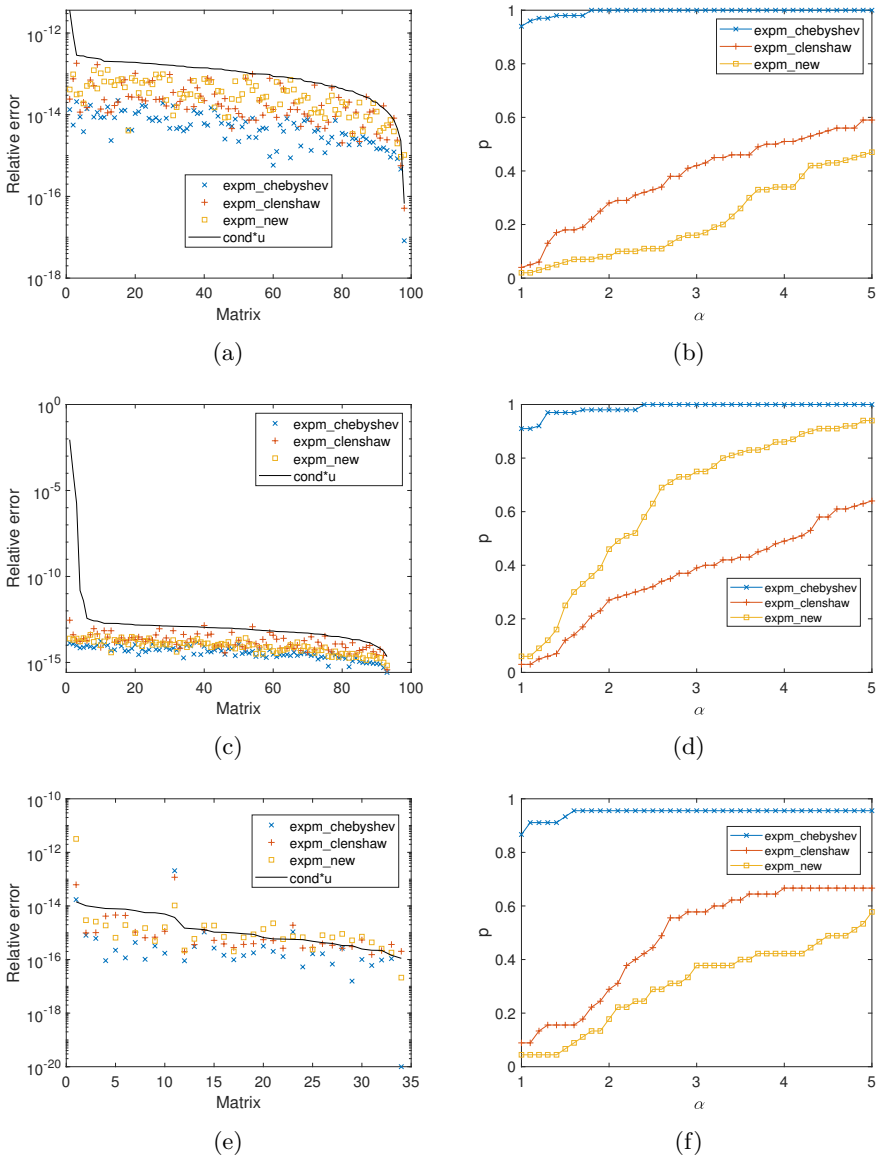


Fig. 1: Normwise relative errors for sets 1 (a), 2 (c), and 3 (e), and performance profiles for the same sets (b, d and f).

5 Conclusions

This paper describes and compares two Euler polynomials-based algorithms devoted to the matrix cosine computation. Using a large and heterogeneous battery of test matrices, it has been shown that both algorithms are more accurate and faster than *cosm*, the best known and most widely employed code for calculating this matrix function by means of the Padé rational approximation.

Acknowledgements

This work has been supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21 and PAID-11-22).

References

1. May, V., Keller, Y., Sharon, N., Shkolnisky, Y: An algorithm for improving non-local means operators via low-rank approximation. IEEE Trans. Image Process. 25 (3), 1340–1353, 2016.
2. Levie, R., Monti, F., Bresson, X., Bronstein, M.M.: Cayleynets: Graph convolutional neural networks with complex rational spectral filters. IEEE Trans. Image Process. 67 (1), 97–109, 2018.
3. Druskin, V., Mamonov, A., Zaslavsky, M.: Cayleynets: Multiscale S-fraction reduced-order models for massive wavefield simulations. Multiscale Model. Simul. 15 (1), 445–475, 2017.
4. Frommer, A., Simoncini, V.: Matrix functions. In: Model Order Reduction: Theory, Research Aspects and Applications, Mathematics in Industry, Wil HA Schilders and Henk A. van der Vorst eds. Springer, 2008.
5. Serbin, S.M., Blalock, S.A.: An algorithm for computing the matrix cosine. SIAM J. Sci. Statist. Comput. 1 (2), 198–204, 1980.
6. Dehghan, M., Hajarian, M.: Computing matrix functions using mixed interpolation methods. Math. Comput. Modell. 52 (5–6), 826–836, 2010.
7. Higham N.J.: Functions of Matrices: Theory and Computation. SIAM, Philadelphia, PA, USA, 2008.
8. Alonso-Jordá, P., Peinado, J., Ibáñez, J., Sastre, J., Defez, E.: Computing matrix trigonometric functions with GPUs through Matlab. J. Supercomput. 75, 1227–1240, 2019.
9. Olver, F.W.J., Lozier, D.W., Boisvert, R.F. and Clark, C.W.: NIST handbook of mathematical functions hardback and CD-ROM, Cambridge University Press 2010.
10. Alonso, J.M., Ibáñez, J., Defez, E., Alonso-Jordá, P.: Euler polynomials for the matrix exponential approximation. J. Comput. Appl. Math. 425, 115074, 2023.
11. Al-Mohy, A.H., Higham, N.J., Relton, S.D.: New algorithms for computing the matrix sine and cosine separately or simultaneously. SIAM J. Sci. Comput. 37, A456–A487, 2015.
12. Defez, E., Ibáñez, J., Alonso, J.M., Alonso-Jordá, P.: On Bernoulli series approximation for the matrix cosine. Math. Meth. Appl. Sci 45(6), 3239–3253, 2022.
13. Higham, N.J.: The matrix computation toolbox, 2002, http://www.ma.man.ac.uk/~higham/mctoolbox.
14. Wright, T.G.: Eigtool, version 2.1, 2009, http://www.comlab.ox.ac.uk/pseudospectra/eigtool.

An efficient method to compute the matrix exponential based on Chebyshev polynomials

E. Defez[1], J. Ibáñez[1], J. M. Alonso[2], and J. Peinado[3]

[1] Instituto de Matemática Multidisciplinar,
[2] Instituto de Instrumentación para Imagen Molecular,
[3] Departamento de Sistemas Informáticos y Computación.
Universitat Politècnica de València
Camino de Vera s/n, 46022, Valencia. Spain
edefez@imm.upv.es,{jjibanez, jmalonso, jpeinado}@dsic.upv.es

Abstract. In this paper, two algorithms based on Chebyshev polynomials are presented to approximate the exponential of a matrix. One of them uses the Clenshaw algorithm to evaluate the polynomial that approximates the matrix exponential function. The other one employs the Paterson-Stockmeyer method, after expressing the exponential function in terms of powers of a matrix. Both algorithms have been compared numerically and computationally with the code that computes the exponential function by means of Padé approximants.

Keywords: Matrix exponential, Chebyshev polynomials, Clenshaw algorithm.

1 Introduction

The matrix exponential is one of the most important matrix functions due to its utility in various areas of science and technology. But its study is also important due to the difficulty of its computation [1, 2]. Among the recent works that require the computation of the exponential matrix, we can cite, e.g. [3–5]. For this reason, several methods have been provided for computing this matrix function. In references [7, 8], authors study the use of Pade's approximants while, in references [9–11], the Taylor polynomial approximation is employed. Recently, new approximations of the exponential matrix have been proposed based on other families of polynomials, such as Hermite, Bernoulli or Euler, see [12–14].

2 Chebyshev polynomials and matrix exponential approximation

In the scalar case, the Chebyshev polynomials are a useful tool in practically all areas of applied mathematics. Chebyshev polynomials $\{T_n(x)\}_{n\geq 0}$ of the first

kind [15] are defined following explicit expressions:

$$\left.\begin{aligned} T_0(x) &= 1 \\ T_n(x) &= \sum_{k=0}^{[\frac{n}{2}]} \binom{n}{2k} \left(x^2 - I_r\right)^k x^{n-2k} \\ &= n \sum_{k=0}^{n} \frac{\left((-2)^k (k+n-1)!\right)}{(2k)!(n-k)!} (1-x)^k \ , \ n \geq 1 \end{aligned}\right\}. \tag{1}$$

Another alternative way of defining Chebyshev polynomials is using the three-term-recurrence formula:

$$\left.\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x), k \geq 1 \end{aligned}\right\}. \tag{2}$$

Among many other properties, Chebyshev polynomials are orthogonal with respect to the weight function $w(x) = (1-x^2)^{-\frac{1}{2}}$ on the interval $[-1, 1]$. In this sense, a function $f(x)$ for $|x| \leq 1$ can be developed in a series of Chebyshev polynomials of the form

$$f(x) = \sum_{k \geq 0} a_k T_k(x), a_k = \frac{2}{\pi} \int_{-1}^{1} f(x) T_k(x) (1-x^2)^{-\frac{1}{2}} dx, \tag{3}$$

where $T_k(x)$ is the kth Chebyshev polynomial, see [16–18] for details. For the particular case of $f(x) = e^x$, coefficients take the form $a_0 = J_0(i)$ and $a_k = 2i^k J_k(-i)$ for all $k \geq 1$, where i is the imaginary unit and $J_k(x)$ is the Bessel function of the first kind of order k. When we consider a matrix $A \in \mathbb{C}^{r \times r}$, $\|A\| \leq 1$, we have the development, similar to that used in [19], of the exponential matrix given by

$$e^A = \sum_{k \geq 0} a_k T_k(x) = J_0(i) I_r + 2 \sum_{k \geq 1} i^k J_k(-i) T_k(A), \ \|A\| \leq 1. \tag{4}$$

Formula (4) that has been used by different authors in the field of computational physics and quantum chemistry, see for example [19–22], or for time integration of the Schrödinger equation, see [23].

All these authors have obtained a polynomial approximation of the exponential matrix that must be evaluated on a matrix. It is well known that the Paterson-Stockmeyer method [24] is the most widely used technique to evaluate matrix polynomials for the reasons explained in [25], although other alternatives with a lower computational cost have recently appeared [26].

Nevertheless, in the literature related to polynomials Chebyshev, it is usual to use the recurrence given in [16, p.125–126], known as the *Clenshaw algorithm*, to evaluate polynomial approximations with Chebyshev polynomials. As described in the following proposition, this recurrence relation provides a way of computing finite sums of scalar Chebyshev polynomials.

Proposition 1. *Let* $\{T_n(x)\}_{n\geq 0}$ *be the sequence of Chebyshev polynomials and let* $Q_n(x)$ *be a polynomial of degree* n *to be evaluated defined by*

$$Q_n(x) = \sum_{k=0}^{n} a_k T_k(x). \tag{5}$$

Let $\bar{x}$ *be a real number and let us consider the following sequence* $\{b_k\}_{k=0}^{n+2}$ *defined by*

$$b_{n+2} = b_{n+1} = 0,$$

$$b_k = a_k + 2\bar{x}b_{k+1} - b_{k+2},\ k = n, \ldots, 1$$

$$b_0 = 2a_0 + 2\bar{x}b_1 - b_2.$$

Then, $Q_n(\bar{x}) = \dfrac{1}{2}(b_0 - b_2)$.

Alternate versions of this algorithm can be found in [27,28]. If coefficients of $Q_n(x)$ defined by (5) are scalars and $\bar{x}$ is a square matrix A, the computational cost of evaluating $Q_n(A)$, expressed in terms of matrix products, is n.

Note that by applying the expansion of the exponential function (4), we can approximate e^A as the evaluation of a polynomial $Q_n(A)$ expressed in (5). Thus, the Clenshaw algorithm can be applied or, alternatively, we can write

$$e^A \approx Q_n(A) = \sum_{k=0}^{n} a_k T_k(A) = \sum_{k=0}^{n} \alpha_k A^k, \tag{6}$$

and then evaluating (6) using the previously mentioned Paterson-Stockmeyer method. As a result, fewer matrix products will be required and the results could be more accurate.

3 The proposed algorithms

Two algorithms, called *expm_clenshaw* and *expm_chebyshev*, have been developed for computing the matrix exponential based on formula (4) together with the scaling and squaring technique. The method proposed in [14] to calculate the degree of the approximation polynomial n and the scaling factor s has been used by both of them.

In the first algorithm, the approximation described in (4) to the matrix exponential is carried out using the matrix version of the Clenshaw algorithm. In contrast, in the second algorithm, matrix A power series and the Paterson-Stockmeyer method are employed for the cited approximation.

Both implementations have been compared with the code *expm_new*, based on the Padé rational approximation for computing the matrix exponential [29].

4 Numerical experiments

Numerical experiments have been performed on a testbed composed of heterogeneous matrices from three different categories:

- **Set 1:** 100 diagonalizable matrices of dimension 128×128. They have been obtained as the result of $A = V \cdot D \cdot V^{-1}$, where D is a diagonal matrix with complex eigenvalues and V is an orthogonal matrix such as $V = H/\sqrt{n}$, with H being a Hadamard matrix and n its number of rows or columns. As 2-norm, we have that $0.1 \leq \|A\|_2 \leq 350$.
- **Set 2:** 100 non-diagonalizable complex matrices of size 128×128. These matrices are computed as $A = V \cdot J \cdot V^{-1}$, where J is a Jordan matrix with complex eigenvalues whose modules are less than 5 and the algebraic multiplicity is randomly generated between 1 and 3. V is an orthogonal matrix identically obtained to the previous set. As 2-norm, we have obtained that $3.58 \leq \|A\|_2 \leq 330.79$.
- **Set 3:** 36 matrices from the Matrix Computation Toolbox [30] and 9 from the Eigtool MATLAB Package [31]. For that matrices, it results that $1 \leq \|A\|_2 \leq 50708.3$.

All numerical executions have been carried out by means of the MATLAB version 2023a. Table 1 collects the percentage of cases in which the normwise relative error committed by algorithm *expm_chebyshev* is lower or higher than that of the codes *expm_clenshaw* and *expm_new*. Respectively for the matrices of each of our 3 sets, algorithm *expm_chebyshev* improved *expm_clenshaw* in 96%, 97%, and 88.89% of the cases. It also outperformed *expm_new* in 98%, 93%, and 93.33% of the matrices. Clearly, these percentages indicate that *expm_chebyshev* is the code that offered the most accurate results.

Table 1: Improvement percentage among the different codes for the 3 sets of matrices.

	Set 1	Set 2	Set 3
$E(expm\_chebyshev) < E(expm\_clenshaw)$	96.0%	97.0%	88.89%
$E(expm\_chebyshev) > E(expm\_clenshaw)$	4.0%	3.0%	11.11%
$E(expm\_chebyshev) < E(expm\_new)$	98.0%	93.0%	93.33%
$E(expm\_chebyshev) > E(expm\_new)$	2.0%	7.0%	6.67%

Regarding the computational cost, Table 2 includes the total number of matrix products required by the three codes in comparison. It is easy to appreciate that *expm_new* demanded the smallest number of products, closely followed by

expm_chebyshev. The quantity of matrix products required by *expm_clenshaw* were much higher. According to this number of products, Table 3 stores the time spent on calculating the exponential of the matrices of each test set. As expected, the highest time was for *expm_clenshaw*. The time involved by *expm_chebyshev* was intermediate among those of the other two codes.

Table 2: Number of matrix products required by *expm_chebyshev*, *expm_clenshaw*, and *expm_new* for the test battery solution.

	Set 1	Set 2	Set 3
$P(expm\_chebyshev)$	1686	1687	574
$P(expm\_clenshaw)$	6531	6550	2207
$P(expm\_new)$	1498	1494	466

Table 3: Elapsed time, in seconds, in the execution of the distinct codes.

	Set 1	Set 2	Set 3
$T(expm\_chebyshev)$	0.43	0.45	0.45
$T(expm\_clenshaw)$	0.90	0.92	2.12
$T(expm\_new)$	0.49	0.49	0.24

Graphics for the normwise relative errors and the performance profile for the different sets of matrices are shown in Figure 1. As can be appreciated in Figures 1a, 1c, and 1e, the three codes are numerically stable, since the relative errors they present are below, or slightly above, the continuous line. Clearly, the lowest relative errors corresponded to the code *expm_chebyshev*.

Very significant are also the results related to the performance profiles, as shown in Figures 1b, 1d and 1f, where *expm_chebyshev* far outperformed the other two codes for any of the matrix sets. On the other hand, *expm_clenshaw* performed better than *expm_new* for Sets 1 and 3, but not for the matrices from Set 2. In light of the results, the strategy of first approximating the exponential function as a series of matrix powers from the Chebyshev polynomials and then applying the Paterson-Stockmeyer method is preferable to employ the Clenshaw algorithm. This conclusion can be also stated with respect to the possibility of employing Padé approximants.

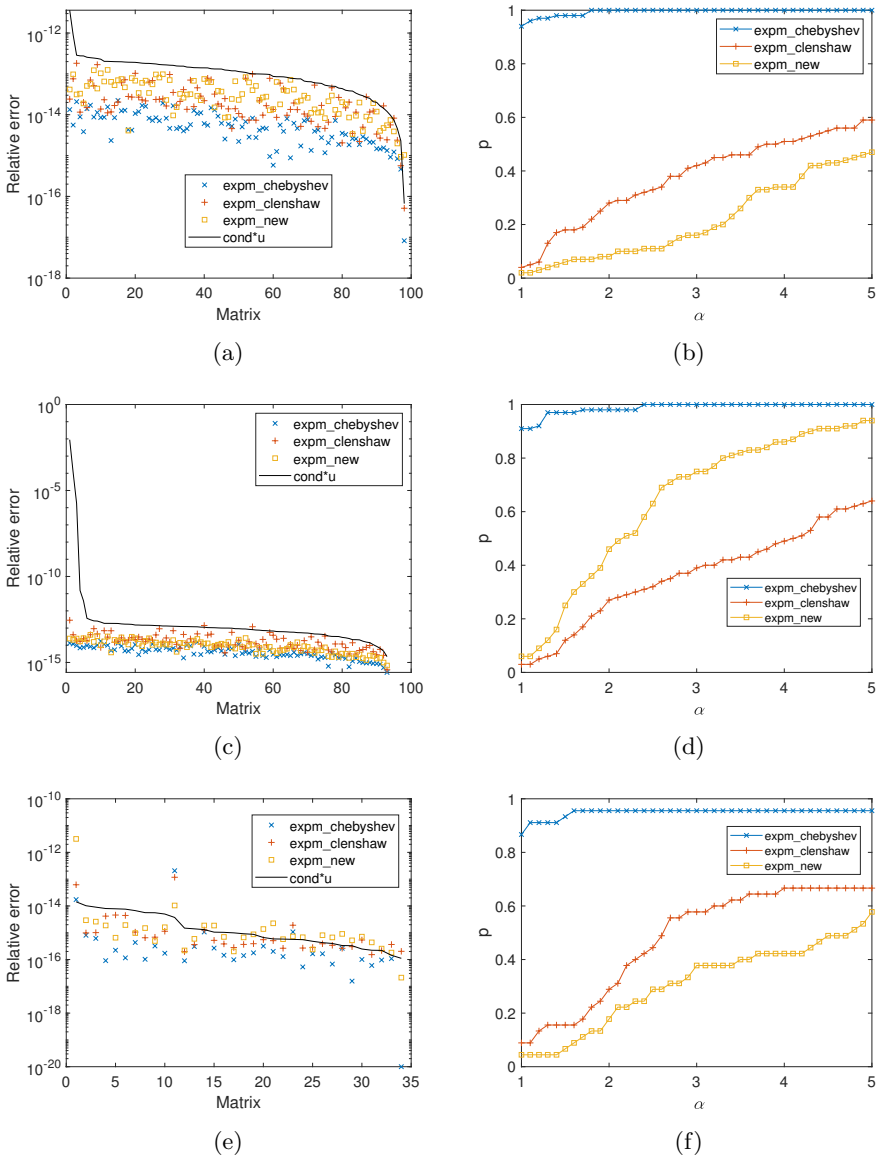


Fig. 1: Normwise relative errors for sets 1 (a), 2 (c), and 3 (e), and performance profiles for the same sets (b, d and f).

5 Conclusions

In general, the implementation based on Chebyshev polynomials is more accurate than the one based on Padé rational approximants, specially in the case of employing the Paterson-Stockmeyer technique to evaluate the matrix polynomial corresponding to the exponential function. Nevertheless, the computational cost may also be higher, particularly when using the Clenshaw algorithm.

Acknowledgements

This work has been supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21 and PAID-11-22).

References

1. Moler C.B. and Van Loan C.: Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*. SIAM Rev. 45, 3–49, 2003.
2. Higham N.J.: Functions of Matrices: Theory and Computation. SIAM, Philadelphia, PA, USA, 2008.
3. Jarlebring, E., Fasi, M., Ringh, E.: Computational graphs for matrix functions. ACM Trans. Math. Software 48 (4), 1–35, 2023.
4. Caglio, L., Stang, H., Brincker, R., Katsanos, E.: FE-aided Kalman filter for nonlinear state estimation with unknown input. Mech. Syst. Sig. Process. 200, 110513, 2023.
5. Caliari, M., Cassini, F.: Direction splitting of φ-functions in exponential integrators for d-dimensional problems in Kronecker form. arXiv preprint arXiv:2304.02327, 2023.
6. Albi, G., Caliari, M., Calzola, E., Cassini, F.: Exponential integrators for mean-field selective optimal control problems. arXiv preprint arXiv:2302.00127, 2023.
7. Arioli, M., Codenotti, B., Fassino, C.: The Padé method for computing the matrix exponential. Linear Algebra Appl. 240, 111–130, 1996.
8. Higham, N.J.: The scaling and squaring method for the matrix exponential revisited. SIAM J. Matrix Anal. Appl. 26 (4), 1179–1193, 2005.
9. Sastre, J., Ibáñez, J., Defez, E., Ruiz, P.: New scaling-squaring Taylor algorithms for computing the matrix exponential. SIAM J. Sci. Comput. 37 (1), A439–A455, 2015.
10. Sastre, J., Ibáñez, J., Defez, E.: Boosting the computation of the matrix exponential. Appl. Math. Comput. 340, 206–220, 2019.
11. Bader, P., Blanes, S., Casas, F.: Computing the matrix exponential with an optimized Taylor polynomial approximation. Mathematics 7 (12), 1174,(2019.
12. Sastre, J., Ibáñez, J., Defez, E., Ruiz, P.: Efficient orthogonal matrix polynomial based method for computing matrix exponential. Appl. Math. Comput. 217 (14), 6451–6463, 2011.
13. Defez, E., Ibáñez, J., Alonso-Jordá, P., Alonso, J.M., Peinado, J.: On Bernoulli matrix polynomials and matrix exponential approximation. J. Comput. Appl. Math. 404, 113207, 2022.
14. Alonso, J.M., Ibáñez, J., Defez, E., Alonso-Jordá, P.: Euler polynomials for the matrix exponential approximation. J. Comput. Appl. Math. 425, 115074, 2023.

15. Chihara, T.S.: An introduction to orthogonal polynomials. Courier Corporation, 2011.
16. Rivlin, T.J.: Chebyshev polynomials. Courier Dover Publications, 1974.
17. Rivlin, T.J.: Chebyshev polynomials. From approximation theory to algebra and number theory. John Wiley & Sons Limited, 1990.
18. Mason, J.C., Handscomb, D.C.: Chebyshev polynomials. Chapman and Hall/CRC, 2002.
19. Auckenthaler, T., Bader, M., Huckle, T., Spörl, A., Waldherr, K.: Matrix exponentials and parallel prefix computation in a quantum control problem. Parallel Comput. 36 (5-6), 359–369, 2010.
20. Tal-Ezer, H., Kosloff, R.: An accurate and efficient scheme for propagating the time dependent Schrödinger equation. J. Chem. Phys. 81 (9), 3967–3971, 1984.
21. Herb, K., Welter, P.: Parallel time integration using batched BLAS (Basic Linear Algebra Subprograms) routines. Comput. Phys. Commun. 270, 108181, 2022.
22. Schulte-Herbrüggen, T., Spörl, A., Waldherr, K., Gradl, T., Glaser, S.J., Huckle, T.: The HLRB cluster as quantum CISC compiler. Matrix methods and applications for advanced quantum control by gradient-flow algorithms on parallel clusters. In: High Performance Computing in Science and Engineering: Transactions of the Third Joint HLRB and KONWIHR Status and Result Workshop, Leibniz Supercomputing Centre, Germany, pp. 517–533. Springer, 2009.
23. Bader, P., Blanes, S., Casas, F., Seydaoğlu, M.: An efficient algorithm to compute the exponential of skew-Hermitian matrices for the time integration of the Schrödinger equation. Math. Comput. Simul. 194, 383–400, 2022.
24. Paterson, M.S. , Stockmeyer, L.J.: On the Number of nonscalar multiplications necessary to evaluate polynomials. SIAM J. Comput. 2 (1), 60–66, 1973.
25. Fasi, M.: Optimality of the Paterson–Stockmeyer method for evaluating matrix polynomials and rational matrix functions. Linear Algebra Appl. 574, 182–200, 2019.
26. Sastre, J.: Efficient evaluation of matrix polynomials. Linear Algebra Appl. 539, 229–250, 2018.
27. Jiang, H., Barrio, R., Li, H., Liao, X., Cheng, L., Su, F.: Accurate evaluation of a polynomial in Chebyshev form. Appl. Math. Comput. 217 (23), 9702–9716, 2011.
28. Barrio, R.: Stability of parallel algorithms to evaluate Chebyshev series. Comput. Math. Appl. 41 (10–11), 1365–1377, 2001.
29. Al-Mohy, A.H., Higham, N.J.: A new scaling and squaring algorithm for the matrix exponential. SIAM J. Matrix Anal. Appl. 31 (3), 970–989, 2010.
30. Higham, N.J.: The matrix computation toolbox, 2002, http://www.ma.man.ac.uk/~higham/mctoolbox.
31. Wright, T.G.: Eigtool, version 2.1, 2009, http://www.comlab.ox.ac.uk/pseudospectra/eigtool.

Advances on the Evaluation of Matrix Polynomials Beyond the Paterson–Stockmeyer Method

Jorge Sastre[1]

Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València,
emailjsastrem@upv.es,
http://hipersc.blogs.upv.es

Abstract. This paper presents recent advances on the methods introduced by the author for the evaluation of general matrix polynomials or matrix polynomial approximations more efficiently than the Paterson–Stockmeyer method. In general these methods are based on multiplications and sums of matrix polynomials. The computing cost of the matrix polynomial evaluation methods is given in terms of the number of matrix products evaluated, since they dominate asymptotically the overall cost. Formulas for the evaluation of matrix polynomial approximations of degree $2^5 = 32$ at cost 5 matrix products are given, allowing to obtain polynomial approximations reproducing the coefficients of the matrix powers of matrix polynomial approximations up to degree 24.
Recently, evaluation formulas for polynomials of degree 16 allowing to obtain polynomial approximations reproducing the coefficients of the matrix powers of matrix polynomial approximations up to degree 15 were given at cost 4 matrix products. On the contrary, in this paper we show that evaluating polynomials of degree 16 at cost 4 matrix products is not possible by using these methods.
Applications to the Taylor polynomial approximation of matrix functions are shown. Their efficiency is compared with that of the state-of-the-art evaluation methods for polynomial approximations, rational approximations and methods based on the mixed rational and polynomial approximations introduced by the author, providing a higher order of approximation for the same cost.

Keywords: matrix polynomial, efficient evaluation, approximation, matrix function

1 Introduction

Since the 70's the Paterson–Stockmeyer (PS) method [2] was considered the most efficient method for the evaluation of general matrix polynomials. This method intends to minimize the number of evaluations of matrix products. The rest of operations, i.e matrix sums and matrix multiplications by a constant,

have cost $O(n^2)$, meanwhile the matrix product has cost $O(n^3)$, dominating asymptotically the total computational cost. From now, the cost of one matrix product evaluation will be denoted on by M.

The multiplication of matrix polynomials to double the degree at a cost of $1M$ was considered in [2] to produce polynomials or degree 2^k, but this procedure was discarded since the number of parameters needed to evaluate polynomials of degree 2^k grows exponentially as 2^k+1, meanwhile the number of parameters obtained by multiplicating polynomials grows quadratically, making not possible to represent all polynomials of degree 2^k even for low values of k [2, Sec. 2]. However, [1, Prop. 1] and [3, Sec. 3.3] gave formulas to evaluate general matrix polynomials saving up to $1M$ and $2M$, respectively, with respect to the Paterson–Stockmeyer method, whenever the coefficient of the highest degree power of the polynomial is not null.

For cost $3M$ the evaluation of general matrix polynomials of the maximum possible degree $2^3 = 8$ is solved in [1, Ex. 3.1], whereas the maximum polynomial degree available at that cost with the Paterson–Stockmeyer method is 6.

For cost $4M$ the case of a Taylor based approximation of the matrix exponential of degree $2^4 = 16$ was shown in [1, Ex. 5.1]. However, in this example the evaluation formulas given, i.e. (57)–(59) from [1, Ex. 5.1], had only 16 parameters allowing to evaluate a Taylor polynomial approximation up to degree 15. From (60) from [1] the coefficient of the term of degree 16 had a relative error in absolute value 0.45 with respect to the respective matrix exponential Taylor coefficient 1/16!. Therefore, in [8, Sec. 3.2] the order of approximation of this Taylor based approximation was denoted by $m = 15+$ since it is more accurate than the Taylor approximation of order 15 but less than the Taylor approximation of order 16. The case for any kind of polynomial approximation of order 15+ is solved in general in [3, Prop. 1] by using the MATLAB Symbolic Toolbox `code fragments 4.1` and `4.2` from [3]. Using a similar MATLAB code it is possible to show that evaluating a polynomial of degree 16 with $4M$ is not possible despite all the parameters more than 16 are introduced in the evaluation formulas. For that purpose, following the nested procedure from (13) and (14) from [8, Sec. 3.2] and in a similar way to Proposition 1 from [3, Sec. 3.1], let

$$y_{02}(A) = \pm A_2(\sqrt{a_8}A_2 + a_7/(2\sqrt{a_8})A), \tag{1}$$

$$y_{12}(A) = \sum_{i=0}^{8} a_i A^i, \tag{2}$$

$$\begin{aligned} y_{22}(A) = {} & A(d_4 y_{12}(A) + d_3 y_{02}(A) + d_2 A^2 + d_1 A + d_0 I) \\ & \times (e_4 y_{12}(A) + e_3 y_{02}(A) + e_2 A^2 + e_1 A + e_0 I) \\ & f_3 y_{12}(A) + f_2 A^2 + f_1 A + f_0 I \\ = {} & P_{16}(A), \end{aligned} \tag{3}$$

where $P_m(A) = \sum_{i=0}^{m} b_i A^i$ is a polynomial of degree m and $y_{22}(A)$ uses all the possible combinations of polynomials of degree 16, 8, 4, 2, 1, 0 intending to evaluate polynomials of degree 16, $P_{16}(A)$, by using multiplications and additions of matrix polynomials. MATLAB code

http://personales.upv.es/jorsasma/Software/coeffspolm16NoSolution.m uses variable substitution with the MATLAB Symbolic Toolbox in a similar way to that in `MATLAB code fragment 4.1` from [3, Sec. 3.1], defining first the symbolic variables and equations (1)-(3). Then, the code equates the coefficients of matrix powers A^i from $y_{22}(A)$ with b_i from $P_{16}(A)$. Finally it solves the variables in the following order a_8, a_7, ..., a_1, f_4 showing that, in general, the system has no solution, see `coeffspolm16NoSolution.m`.

Moreover, in [3, Prop. 1] we solved general polynomial approximations of degree 15+ making the coefficient of the term in degree 16 dependent on the coefficients b_i, $i = 7, 8, \ldots, 15$, see line 35 and 36 of `MATLAB code fragment 4.1` [3]. However, other coefficient different from the coefficient in degree 16 can be selected which may have less relative error with respect to the corresponding Taylor coefficient or the corresponding polynomial approximation coefficient, or may have other interesting properties. For instance:

1. For the matrix exponential approximation of order 15+ said above the relative error in absolute value using the term of degree 8 is 0.06 instead of 0.45 obtained for the term of degree 16. Note that this kind of approximation with an error in the term of degree 8 is able to reproduce the term of degree 16.
2. For the Taylor approximation of $\cosh(A)$ the relative error for the term of degree 14 is 0.13 whereas the error using the term of degree 16 is 0.45.
3. For the matrix logarithm in [9] all the possible approximations of degree 15+ of the matrix logarithm were complex, whereas we have checked that selecting the coefficient of degree 15 to be dependent on the others instead of coefficient of degree 16, an approximation with real coefficients of order 14+ can be obtained, which also reproduces the Taylor approximation term of degree 16.

Finally, since the 70's rational approximations were considered more efficient than polynomial approximations [5], although this was not always true in the computation of matrix functions like the matrix exponential and the matrix cosine as shown in the last years [6, 7]. With the new polynomial evaluation method we showed that polynomial approximations now provide a higher order of approximation than the state-of-the-art computational methods for rational approximations for the same cost in terms of matrix products, and applications were given [1, 3, 8]. In this paper, we give evaluation formulas for matrix polynomials providing the maximum polynomial degree $2^5 = 32$ available for cost $5M$, with order of approximation 24+. These results improve the results from [3, Prop. 2] where approximations of degree only 24 and order 21+ for the same cost. Note that from [4, Tab. 3] the maximum order of approximations available for that cost using the Paterson–Stockmeyer method is 12. For Padé approximations the maximum available order of approximation is 12 at a higher cost of $5.33M$, see [1, Tab. 8], and for the mixed rational and polynomial approximations from [4] the maximum order available is 16 also at a higher cost $5.33M$, see [1, Tab. 8].

2 Evaluation of polynomial approximations of degree $2^5 = 32$ and approximation order 24+ at cost 5 matrix products

Following the nested procedure from (13) and (14) from [8, Sec. 3.2] and in a similar way to Proposition 1 from [3, Sec. 3.1], let

$$y_{02}(A) = \pm A_2(\sqrt{a_8}A_2 + a_7/(2\sqrt{a_8})A), \tag{4}$$

$$y_{12}(A) = \sum_{i=3}^{8} a_i A^i, \tag{5}$$

$$y_{22}(A) = \sum_{i=3}^{16} b_i A^i \tag{6}$$

$$\begin{aligned} y_{32}(A) = {} & A(y_{22}(A) + d_3 y_{02}(A) + d_2 A^2 + d_1 A) \\ & \times (y_{22}(A) + e_4 y_{12}(A) + e_3 y_{02}(A) + e_2 A^2 + e_1 A) \\ & f_4 y_{22}(A) + f_3 y_{12}(A) + f_2 A^2 + f_1 A + f_0 I, \end{aligned} \tag{7}$$

where:

1. Using variable substitution with the MATLAB Symbolic Toolbox in a similar way to that in `MATLAB code fragment 4.1` from [3, Sec. 3.1], the coefficients a_i, $i = 3, 4, \ldots, 8$, from (5) can be written in terms of coefficients b_i from $y_{22}(A)$ from (6). For instance,

$$a_8 = \sqrt{b_{16}}, \tag{8}$$

$$a_7 = \frac{b_{15}}{2\sqrt{b_{16}}}, \tag{9}$$

$$a_6 = \frac{4\,b_{14}\,b_{16} - {b_{15}}^2}{8\,{b_{16}}^{3/2}}, \tag{10}$$

$$a_5 = \frac{{b_{15}}^3 - 4\,b_{14}\,b_{15}\,b_{16} + 8\,b_{13}\,{b_{16}}^2}{16\,{b_{16}}^{5/2}}, \tag{11}$$

   where $a_8 = -\sqrt{b_{16}}$ could be selected to obtain another set of solutions for a_i, $i = 3, 4, \ldots, 8$. Note that the corresponding expressions for a_4 and a_3 are too long and have been omitted.
2. $y_{12}(A)$ from (5) can be computed at cost $3M$ as shown in [1, Ex. 3.1].
3. Using (4), (8) and (9) it follows that $y_{02}(A)$ from (4) can be expressed in terms of coefficients b_{16} and b_{15} of $y_{22}(A)$ from (6) as

$$y_{02}(A) = \pm A_2(\sqrt[4]{b_{16}}A_2 + b_{15}/(4\,\sqrt[4]{b_{16}}^3)A), \tag{12}$$

   or

$$y_{02}(A) = \pm iA_2(\sqrt[4]{b_{16}}A_2 + b_{15}/(4\,\sqrt[4]{b_{16}}^3)A). \tag{13}$$

4. $y_{22}(A)$ from (6) can be computed at cost $4M$ in a similar way to [3, Prop. 1] as

$$y_{22}(A) = A((y_{12}(A) + d_2'A^2 + d_1'A)(y_{12}(A) + e_3'y_{02}(A) + e_2'A^2 + e_1'A) + f_4'y_{12} + f_3'y_{02} - d_1'e_1'A^2, \tag{14}$$

   where using a similar code to `MATLAB code fragment 4.1` from [3, Sec. 3.1] it is easy to show that a_8 cannot be freely set and depends on b_i, $i = 7, 8, \ldots, 15$, see code line 35. Then, using (8) it is possible to show that b_{16} depends on the same variables as an equation given by a polynomial of several variables in b_i, $i = 7, 8, \ldots, 16$ equated to zero. This equation is far too long and has been omitted in this text.
5. From (4)-(7) it is easy to show that $y_{32}(A)$ can be written as a polynomial of degree 32

$$y_{32}(A) = \sum_{i=0}^{32} c_i A^i, \tag{15}$$

   with a total of 25 parameters b_i, $i = 3, 4, \ldots, 15$, d_1, d_2, d_3, e_1, e_2, e_3, e_4, f_0, f_1, f_2, f_3 and f_4, where b_{16} depends on b_i, $i = 7, 8, \ldots, 15$ as the polynomial of several variables b_i, $i = 7, 8, \ldots, 16$ equated to zero indicated in the previous item. This equation for b_{16} is added to the 25 equations that arise when equating the coefficients of each power of matrix A^i, $i = 0, 1, \ldots, 24$, from (7) with the coefficients c_i, $i = 0, 1, \ldots, 24$, of the desired polynomial approximation of degree 24.

Hence, in a similar way as the Taylor approximation of order 15+ of the matrix exponential from [8, Sec. 3.2] and [1, Ex. 5.1] it follows that $y_{32}(A)$ can be a polynomial approximation of order 24+ or a polynomial approximation that reproduces the coefficients of the matrix powers A^i, $i = 0, 1, \ldots, 24$ of a given polynomial approximation.

Using function `vpasolve` from the MATLAB symbolic toolbox Table 1 shows a real solution for the coefficients from $y_{22}(A)$ and $y_{23}(A)$ from (6) and (7) for a 24+ approximation of the Taylor series related to the matrix logarithm

$$-\log(I - A) = \sum_{i \geq 1} c_i A^i, \tag{16}$$

where $c_0 = 0$ and $c_i = 1/i$. For these approximation, the relative error in absolute value of the terms of degrees 25, 26 $\ldots, 32$, from (7) with respect to the corresponding Taylor coefficients $c_i = 1/i$ are 0.66, 1.11, 0.87, 0.82, 0.94, 0.86, 0.86 and 0.96 respectively. These solution was found giving an initial guess real solution to the MATLAB function `vpasolve`.

3 Conclusions

In this paper new advances on the matrix polynomial evaluation methods from [1] have been presented. Evaluation formulas for matrix polynomial approximations of degree $2^5 = 32$, the maximum available for 5 matrix products, are

c_3	$8.601783619558866\times10^{-1}$	c_{16}	$3.565553990239603\times10^{-2}$
c_4	$9.944287436604792\times10^{-1}$	d_1	$7.375609514477606\times10^{-2}$
c_5	$-3.074615389296343\times10^{-2}$	d_2	$3.575370392498894\times10^{-2}$
c_6	$1.592445908502036\times10^{-1}$	d_3	$-2.621067300901111\times10^{0}$
c_7	$-1.011053421644045\times10^{-2}$	e_1	$-1.049585293286937\times10^{0}$
c_8	$1.508589505550175\times10^{-2}$	e_2	$-7.839707251483478\times10^{-2}$
c_9	$1.603095162964840\times10^{-1}$	e_3	$-1.263266986928881\times10^{-2}$
c_{10}	$1.150108059224550\times10^{-2}$	e_4	$2.551827558314212\times10^{0}$
c_{11}	$-4.977397388243390\times10^{-2}$	f_0	0
c_{12}	$6.56775568034843\times10^{-2}$	f_1	1
c_{13}	$1.329804039052917\times10^{-2}$	f_2	$0.5 - d_1e_1$
c_{14}	$9.285680783553884\times10^{-3}$	f_3	$6.054408747641660\times10^{-1}$
c_{15}	$6.302431936136797\times10^{-2}$	f_4	$5.888361184781625\times10^{-1}$

Table 1: Coefficients of y_{22} and y_{32} from (6) and (7) for computing the matrix logarithm Taylor approximation from (16) of order $m = 24+$.

given, allowing to reproduce the terms of degrees up to 24 of those polynomial approximations. Their efficiency is compared to the Paterson–Stockmeyer methods, Padé rational methods, and the mixed polynomial and rational methods from [4].

It is also shown that it is not possible to evaluate a general polynomial of degree 16 with evaluation formulas (1)-(3) using multiplications and additions of polynomials of degrees 16, 8, 4, 2, 1, 0. Despite the evaluation formulas have more than the 17 needed parameters, they are not independent. Moreover, the option to give evaluation formulas of matrix polynomial approximations of degree 16 reproducing the polynomial approximation coefficients of the matrix powers A^i for $i \neq 16$ is proposed, and examples of its advantages for the matrix functions $\exp(A)$, $\cosh(A)$ and the matrix logarithm are given.

4 Acknowledgements

This research was supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21 and PAID-11-22)

References

1. Sastre, J., Efficient evaluation of matrix polynomials, *Linear Algebra Appl.*, 538:229–250 (2018). https://doi.org/10.1016/j.laa.2017.11.010

2. Paterson, M.S., Stockmeyer, L.J., On the number of nonscalar multiplications necessary to evaluate polynomials, *SIAM J. Comput.* 2(1):60–66, 1973.
3. Sastre, J.; Ibáñez, J., Evaluation of Matrix Polynomials beyond the Paterson-Stockmeyer Method, *Mathematics*, 9, 1600 (2021). https://doi.org/10.3390/math9141600
4. Sastre, J., Efficient mixed rational and polynomial approximation of matrix functions, *Appl. Math. Comput.* 218(24):11938–11946 (2012). http://dx.doi.org/10.1016/j.amc.2012.05.064
5. Higham, N.J., Functions of Matrices: Theory and Computation. Philadelphia, PA, Society for Industrial and Applied Mathematics, USA, 2008.
6. Ruiz, P., Sastre, J., Ibáñez, J., Defez, E., High performance computing of the matrix exponential, *J. Comput. Appl. Math.* 291:370–379 (2016). http://dx.doi.org/10.1016/j.cam.2015.04.001
7. Sastre, J., Ibáñez, J., Alonso, P., Peinado, J., Defez, E., Two algorithms for computing the matrix cosine function, *Appl. Math. Comput.* 312:66–77 (2017). http://dx.doi.org/10.1016/j.amc.2017.05.019
8. Sastre, J.; Ibáñez, J., E. Defez, Boosting the computation of the matrix exponential. Appl. Math. Comput., 340, 206–220 (2019). https://doi.org/10.1016/j.amc.2018.08.017
9. Ibáñez, J.; Sastre, J.; Ruiz, P.; Alonso, J.M.; Defez, E. An Improved Taylor Algorithm for Computing the Matrix Logarithm. Mathematics, 9, 2018 (2021). https://doi.org/10.3390/math9172018

Part X

Student Project's

Machine Learning-based Graph Size Reduction for Electric Vehicle Routing Problems

Yusef Ahsini[1], Pablo Díaz-Masa[1], Belén Inglés[1], and J. Alberto Conejero[2]

[1] Escuela Técnica Superior de Ingeniería Informática,
Universitat Politècnica de València,
Valencia, 46022, Spain
{yahsoua, pdiamas, binggra}@etsinf.upv.es
[2] Instituto Universitario Matemática Pura y Aplicada,
Universitat Politècnica de València,
Valencia, 46022, Spain
aconejero@mat.upv.es

Abstract. In light of the growing demand for online shopping and home delivery services, the importance of optimizing electric delivery vehicle routes in urban areas must be addressed. Such a measure is critical to significantly reducing environmental pollution and improving operational efficiency. To address this pressing issue, we propose a novel approach for solving routing problems specific to electric vehicles. The proposed method combines XGBoost with the utilization of synthetic data, introducing a rapid graph reduction technique to achieve efficient results.

Keywords: electric vehicles, graph theory, routing problems

1 Introduction

The shift from combustion engine vehicles to electric vehicles (EV) in the automotive industry is leading to a positive environmental impact [7]. This transformation represents a crucial step towards a cleaner and more sustainable transportation system. The introduction of electric cars significantly impacts several Sustainable Development Goals (SDGs). It contributes to Goal 13 (*Climate Action*) by reducing greenhouse gas emissions, Goal 7 (*Affordable and Clean Energy*) by promoting renewable energy use, and Goal 11 (*Sustainable Cities and Communities*) by fostering cleaner and more efficient transportation systems, leading to improved urban air quality and healthier living environments.

For that reason, developing efficient routing solutions for EVs is crucial due to their unique characteristics, such as sensitivity to slopes and the possibility of routes with negative consumption. Optimizing routes considering these factors maximizes driving range and energy efficiency. In delivery services, where problems like the Traveling Salesman Problem (TSP) arise, tailored EV routing becomes even more vital. Fast and real-time routing algorithms are essential for timely deliveries and optimal battery utilization, making EV routing indispensable for sustainable and efficient transportation solutions.

In this work, we combine a machine learning model with a heuristic algorithm for finding a solution to routing problems like the Travelling Salesman Problem for an electric vehicle, where elevation data is considered for estimating energy consumption. As we will see, this innovative approach significantly accelerates the solution process while maintaining accuracy, allowing us to reduce the use of more computationally demanding search algorithms in favor of faster machine learning models. The resulting routing solutions efficiently and effectively address complex city routing problems with driverless EVs, which can be applied to transport, delivery, or car-rental business [5, 6].

In Section 2, we explain our methodology. In Section 3, we explain how we generate synthetic data to train the ML predictive model indicated in Section 2. Section 4 covers the model training process, and in Section 5, we showcase the methodology validation.

2 Methodology

A city can be modeled with a directed graph where the nodes represent crossroads or intersections, and the edges represent streets and roads. The TSP consists of giving a list of nodes on a connected graph and the distances (weights) between them, finding the shortest possible path that visits each node just once, and returning to the first node [10]. In this work, we propose to solve the TSP on a reduced graph whose nodes are the ones in the statement of the TSP, and the edge weights are estimated through machine learning. Figure 1 provides a visual representation of this new graph for a routing problem involving 4 nodes $\{3, 7, 15, 19\}$. After solving the problem in this reduced graph, we obtain the order to visit the nodes for solving the TSP. Now, between each pair of nodes, the real route is obtained in the whole graph through a shortest path algorithm like Dijkstra or Bellman-Ford. This approach streamlines the process and optimizes the solution while maintaining accuracy.

In the context of EVs, we prefer to estimate the edge weight relying on the energy consumption instead of the distance, which lets us consider the energy recovery of an EV. As in [11], we propose considering the energy consumption between a pair of nodes by a machine learning (ML) predictive model instead of assigning weights to the edges representing the energy consumption and using classical search algorithms like Bellman-Ford [2, 8].

We generate a training dataset of random routes to train our ML model. Each route consists of an initial and final node. For each node, we consider its coordinates and its elevation. To estimate the energy required to go from one node to another, we proceed as follows. We consider the whole city graph, where each edge weight is estimated as the energy consumption of going from one node to another. Once this graph is computed, we have applied Bellman-Ford algorithm to compute the less energy-consuming path between the origin and the final destination of each route. It is advisable to use Bellman-Ford instead of Dijkstra as edges with negative weights may appear, as this is the case when the street is going down. We will provide further details in Section 3.

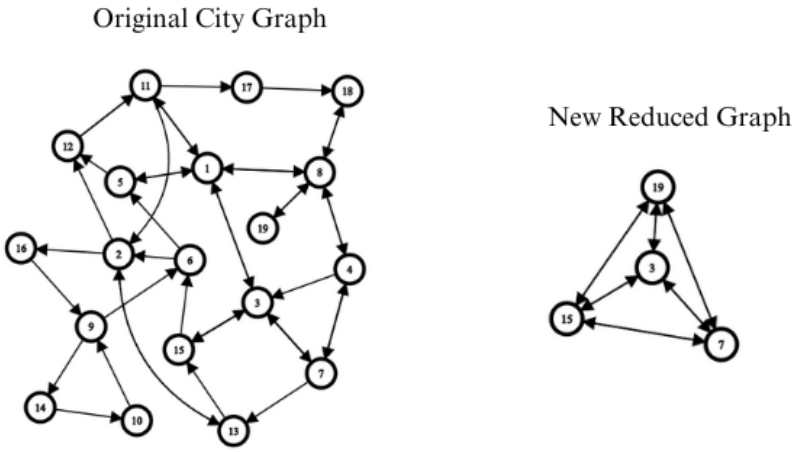


Fig. 1: Visual representation of the graph reduction for an instance of the TSP problem involving nodes $\{3, 7, 15, 19\}$.

Suppose that the ML predictive model has been trained. Given an instance of the TSP problem with n-stops, we create a new, simple complete, directed weighted graph of n nodes representing the stops in the routing problem. This graph has $(n-1)^n$ edges, where the edge weights are computed through the ML predictive model. This reduced graph, considerably smaller than the original city graph, allows a quicker computation of solutions. Once the solution is obtained, we reconstruct the routes using the search algorithm to obtain the whole route in the original graph.

3 Syntetic Data generation

We have generated random routes of synthetic data from Madrid to validate our methodology. As we have said, they consist of the initial and final locations' coordinates and elevation and the energy consumption required to go from the initial to the final location by taking the less energy-consuming route obtained using Bellman-Ford. We have obtained the city graph of Madrid through the OSMmx library [3]. To enhance accuracy, elevation data from Spain's National Geographic Information Center [14] was integrated, and energy estimation followed the methodology proposed by Graser et al. [9]. This energy estimation can be improved if we add historical traffic data. For the case of Madrid, such data can be obtained from Madrid Open Data [12]. If this is added, the initial time the route starts has to be added as a characteristic of the consumption for traveling between the initial and final locations.

Unlike generating data under static graph conditions, the study incorporated real traffic data from October and November 2022. For each hour during this period, a corresponding graph was generated, changing the edge's maximum speed

to the traffic's velocity. This allows the models to extract energy consumption variations due to the effect of traffic congestion.

The data generation process involved simulating 244,000 random routes within the graph and computing the estimated energy consumption for each route using the Bellman-Fords algorithm. For every route, the latitude, longitude, and altitude of both origin and destination points and the hour and day of travel were recorded. Table 1 illustrates an example dataset generated for the city of Madrid in October 2022.

Table 1: Examples of data generated for October, 2022. For each route we have the longitude (lon), latitude (lat), and elevation (elv) of the origin (org) and the destination (dest), the day and hour, and the energy required.

Day	Hour	Org_lon	Org_lat	Org_elv	Dest_lon	Dest_lat	Dest_elv	Energy_used
10	0	−3.59920	40.36398	641.917	−3.73172	40.39640	621.366	2047.03
10	0	−3.71903	40.38655	605.261	−3.62685	40.40193	677.495	1739.99
10	0	−3.67403	40.39392	588.506	−3.70828	40.42033	652.923	996.85
10	0	−3.66224	40.42910	661.864	−3.66718	40.43910	685.921	295.73
10	0	−3.70515	40.57854	748.271	−3.70782	40.45727	715.779	2882.84
10	0	−3.77492	40.45370	621.177	−3.60805	40.40292	685.249	3340.12

Prior to model training, it is essential to conduct a brief data preprocessing step. This involves converting categorical variables into numerical ones and extracting weekday information, including whether the day is a working day. Following the preprocessing step, the resulting dataset comprises 39 variables. Among these variables are new binary representations for identifying working days, encoding for hours (24 variables), and weekdays (7 variables).

4 Model Trainning

The model utilized to estimate the energy consumption is XGBoost, an machine learning algorithm based on gradient boosting and tree-based ensemble learning [4]. XGBoost and Scikit-Learn Python's libraries were employed to train and test the model parameters using cross-validation.

In XGBoost, the learning rate, max depth, and number of estimators are important hyperparameters. The learning rate controls how fast the model learns from data during training, while the max depth limits the complexity of individual trees, affecting their ability to capture patterns in the data. The number of estimators determines the total trees in the ensemble, impacting model accuracy and computation time. Properly tuning these hyperparameters is crucial

to strike the right balance between model performance and generalization. To ensure the correct selection of these hyperparameters, the dataset was divided into 5 blocks to carry on a 5-fold cross-validation. Each configuration involved training 5 models, using 4 blocks for training and 1 for testing. The average root mean squared error (RMSE) of the 5 models generated for each configuration was then computed. The errors of the various trained models can be found in Table 2.

Table 2: Cross-validation results for the XGBoost model.

Learning Rate	**Max Depth**	**Estimators**	**Average RMSE**
0.1	3	200	235.69
0.1	3	500	186.98
0.1	3	800	168.18
0.1	5	200	151.63
0.1	5	500	110.58
0.1	5	800	96.6
0.1	7	200	101.46
0.1	**7**	**500**	**80.87**

Finally, we train a last model using the best configuration hyperparameters {learning rate = 0.1, max deph = 7, number of estimators = 500}. This model was trained with 80% of the synthetic data, and the remaining 20% was used for validation.

5 Validation

The proposed methodology's validation uses the Traveling Salesman Problem (TSP) as a benchmark. To solve the TSP, the 2-OPT algorithm with a Nearest Neighbours initialization is employed, as described by Nuraiman [13]. We compare the TSP's solution using Bellman-Ford with the solution provided by our hybrid method combining XGBoost with Bellman-Ford for reconstructing the whole route.

As we can see, the difference in thersm of the energy estimation is of the order of 10^{-4}%, while the execution time is reduced in an almost 90%.

Table 3: Comparison between node filtering with Bellman-Ford and XGBoost.

	Bellman-Ford	**XGBoost**	**Difference**
Average energy consumption (kWh)	10.966	10.970	0.00036%
Average execution time (s)	19.25	2.19	-88.62%

6 Conclusion

The suggested approach has demonstrated its effectiveness in simplifying routing problems for EVs. This leads to quicker and more efficient solutions for routing challenges, paving the way for innovative, customized routing options that cater to the distinctive features of EVs and are suitable for real-time applications. Moreover, implementing this methodology introduces the potential to incorporate city traffic considerations, ensuring dependable solutions that always remain reliable.

Furthermore, the positive impact of this paper extends to developing solutions that promote the broader adoption of EVs in general. Significant barriers that may have hindered the transition to a more sustainable fleet are overcome by offering more efficient and effective approaches to electric vehicle route planning.

7 Acknowledgements

J.A.C. acknowledges funding from Grant No.PID2022-138860NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe," by the "European Union."

References

1. Ahsini, Y., Díaz-Masa, P., Inglés, B., Rubio, A., Martínez, A., Magraner, A., and Conejero, J. A. (2023). The electric vehicle travelling salesman problem on digital elevation models for traffic-aware urban logistics. Algorithms [Preprint].
2. Bellman, R. (1958). On a routing problem. Q. Appl. Math. 16(1) 87–90 . https://doi.org/10.1090/qam/102435
3. Boeing, G. (2017). OSMNX: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput. Environ. Urban Syst. 65, 126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004.
4. Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 784-795. https://www.doi.org/10.1145/2939672.2939785
5. Conejero, J.A., Jordán, C., and E. Sanabria-Codesal. An iterative algorithm for the management of an electric car-rental service.J. Appl. Math. 483734 (2014). https://doi.org/10.1155/2014/483734.

6. Conejero, J.A., Jordán, C., and Sanabria-Codesal, E. (2016). An algorithm for self-organization of driverless vehicles of a car-rental service. Nonlinear Dyn. 84, 107-114. https://doi.org/10.1007/s11071-015-2237-4
7. Ferrero, E., Alessandrinia, S., and Balanzino, A. (2016). Impact of the electric vehicles on the air pollution from a highway. Appl. Energy. 169, 450-459. https://doi.org/10.1016/j.apenergy.2016.01.098.
8. L.R. Ford Jr. (1956). Network flow theory. Technical report, Rand Corp Santa Monica Ca.
9. Graser, A., Asamer, J., and Ponweiser, W. (2015). The elevation factor: Digital elevation model quality and sampling impacts on electric vehicle energy estimation errors. In 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 81–86. https://doi.org/10.1109/MTITS.2015.7223240
10. Gross, J.L., Yellen, J., and Anderson, M. (2018). Graph Theory and its Applications. Chapman and Hall/CRC.
11. Jindal, I. and Chen, X. and Nokleby, M. and Ye, J. and et al. A unified neural network approach for estimating travel time and distance for a taxi trip. arXiv preprint arXiv:1710.04350, 1–9 (2018). https://arxiv.org/abs/1710.04350
12. Open Data Madrid. Madrid's historic traffic data. https://datos.madrid.es/ Last access on July 31st, 2023.
13. Nuraiman, D., Ilahi, F., Dewi, Y., and Hamidi, E.A.Z. (2018). A new hybrid method based on nearest neighbor algorithm and 2-Opt algorithm for Traveling Salesman Problem. 4th International Conference on Wireless and Telematics (ICWT), Nusa Dua, Bali, Indonesia,1-4, https://doi.org/10.1109/ICWT.2018.8527878.
14. Spanish National Geographic Information Center (CNIG). Download center https://centrodedescargas.cnig.es. Last access on July 31st, 2023.

Misinformation Detection Pipeline

Hugo Albert Bonet, Iván Arcos Gabaldón David Borregón Sacristán, Diana Haj, Kexin Jiang Chen, and José Francisco Olivert Iserte

Valencia Polytechnic University, Valencia 46022, Spain,
misinformationdetectionupv@gmail.com
*all authors contributed equally

Abstract. Misinformation during elections has become a significant concern, as it can influence public opinion and undermine democratic processes. In the context of Nigeria's 2023 Presidential Election, UNICC posed the challenge to build an automatic system capable of effectively detecting false claims, hoaxes, and other forms of misinformation spread on Twitter. The system consists of a fact-checking algorithm and a machine learning model trained on Nigerian political tweets using advanced Natural Language Processing (NLP) techniques like BERT-style transformers. By incorporating information on toxicities and emotions, achieving high F1-Scores of around 0.9. Also, a fact-checking database and reusable resources were created, showcasing the effectiveness of NLP techniques in detecting misinformation. The project offers additional value by establishing an extensive fact-checking database and providing resources for future projects. Monitoring the system's performance in real-time during upcoming elections can contribute to ongoing research on misinformation detection and its impact on democratic processes.

Keywords: misinformation, elections, NLP, transformers, fact-checking, machine learning.

1 Introduction

Misinformation refers to false information, regardless of whether or not it is intended to mislead or deceive people. With the constant evolution of digital platforms and technologies like social media, misinformation spreads farther, faster, and deeper than truthful information where the most common issues are related to immigration, gender, politics, equality and vaccination that can cause real-world consequences like deterioration of the trust in journalism and science. It is crucial to address this problem in order to create safe digital spaces. Misinformation detection comes with several challenges, such as models becoming outdated due topic/vocabulary mismatch between the new social media post and the training data used to build models. Additionally, the availability of annotated data is limited as it takes time and effort to compile an up-to-date annotated dataset.

Our main focus will be tackling misinformation in Nigerian elections, where the spread of false information can have serious consequences. Therefore our

main goal is to help effectively detect and prevent false claims, hoaxes, and other forms of misinformation from spreading on social media. To accomplish this, we will develop a misinformation detection pipeline (an automatic system capable of distinguishing fake news) consisting of two main components: a machine learning module and a fact checking module. This pipeline solves the issues of misinformation detection mentioned above by re-training the model and annotating data semi-automatically. Previous research on automated fact-checking models such as those by Kotonya and Toni [1], Nakov et al [2], provided valuable insights about the process followed to automate the fact-checking tasks, and the challenges faced in the process. We found that BERT-style transformers are the most commonly used models by researchers.

The goals and values of this project align with the importance of combating misinformation and promoting the use of accurate and reliable information in society, which can have a positive impact on individuals, organizations, and society as a whole. We hope that with this misinformation detection pipeline, we can help individuals and organizations quickly and accurately identify and correct false information, thereby preventing it from spreading further on social media platforms, and strengthen trust between the users.

2 Methods

2.1 Data

Data extraction We collected data related to Nigerian elections from two main sources; fact-checking pages and news outlets using Python's Requests package [3] and Selenium [4]. The fact-checking pages included iVerify Zambia [5], Politifact [6], Africa Check [7], Dubawa [8], AFP [9], Lead Stories [10], The Dispatch [11] and News Verifier Africa [12], while the news outlets were The Guardian [13], Punch Nigeria [14], Daily Post [15], Sahara Reporters [16], The Nation Online Nigeria [17], and Vanguard Nigeria [18].

Due to the disparity of the labels used in the different pages, we unified the labels to False and True. Resulting in a total of 673 rows with 546 False claims and 127 True claims. To address the imbalance of classes, we created an Evidence Database of 22567 observations from the fact-checking pages and Wikipedia. This is a new dataset with the evidence from fact-checking pages separated by sentences and general information related to Nigeria extracted from Wikipedia. We also labeled 2000 tweets using the GPT 3.5 Turbo API [19], but due to a high number of False labels, we generated synthetic tweets to balance the classes. We supplied five randomly selected sentences from the Evidence Database and generated a tweet related to the information provided. The final dataset had a total of 3004 observations, with 623 True and 1381 False labels. This data will be used to build the Fact-Checking Module.

Data preparation Data preparation process involved merging individual CSV files from different fact-checking pages into a consolidated dataset, extracting evidences from it and creating an index for easy reference and information retrieval.

Text data is converted to lowercase to ensure consistency, and inconsistencies or missing labels are addressed by unifying classes and removing duplicates. Additional features, like toxicity, sentiment, and subjectivity, are extracted using libraries such as NLTK, Emo-RoBERTa, unbiased-toxic-roberta, and TextBlob to gain deeper insights from the data.

2.2 Model building

Machine learning module The machine learning module consists of a machine learning model trained specifically to classify a claim according to the presence of misinformation in it. The process can be divided into different steps which we will explain in detail below:

1. Embedding and Similarity Computation. The text of the claim and each retrieved document is converted into numerical vectors. This facilitates the computation of similarity between the claim and the documents, We then compute the cosine similarity between the claim embedding and each document embedding, selecting the top 5 most relevant documents (evidence) for further processing.

2. Preprocessing, Feature Extraction and Dataset Creation. We use a pre-trained SBERT model [20], 'mitra-mir/setfitmodel-Feb11-Misinformation-on-Media- Traditional-Social', designed for misinformation detection on traditional and social media, to classify the claim based on the retrieved documents. The data is preprocessed by concatenating the claim and the top 5 most relevant evidence sentences, and then encoded using the SBERT model to obtain embeddings. With that we create a structured dataset that includes the claim, the top 5 evidence sentences, and the label. To enrich the dataset and improve classification performance, we incorporate predictions from various models:

- Emotion detection using the 'arpanghoshal/EmoRoBERTa' model [21] to capture the emotional context of claims. This model detects 28 different types of emotions in the text. We have included the probability scores for each emotion in the dataset as this information can help in understanding the emotional context of the claims.

- Toxicity detection using the 'unitary/unbiased-toxic-roberta' model [22] to identify and filter out toxic or biased claims such as general toxicity, insults, identity attacks, etc. By incorporating this information, we can potentially identify and filter out toxic or biased claims from the dataset.

3. Prediction and explanation. The best-performing model is selected based on its accuracy on the validation set. This model is then used to classify the claim as true or false based on the retrieved evidence. We then built models to provide an explanation for the classification of the claim. We made use of Explainable AI and GPT3 to do so, which are explained in detail below:

3.1. Explainable AI. In the context of Nigerian elections, Explainable AI (XAI) is crucial to determine which features the model is using to determine whether a claim is true or false.

3.2. GPT3. Since we wanted to obtain a comprehensive explanation for the label assigned to a claim, we opted for GPT-3 in the end. We provided GPT-3 with the claim, the top 5 pieces of evidence and the predicted label obtained from the classification models.

Fact-checking module We built a fact-checking module, which consists of a rule-based module that emulates the behavior of a journalist during the process of fact-checking [23]. To build the fact-checking module, we explored many different approaches which we discuss below:

- **Emotion Analysis.** This approach involves analyzing the sentiment of the text to determine if it contains misinformation. Sentiment analysis can be used to identify the emotional tone of the text and determine if it is consistent with the facts presented in the article.

- **Decision Tree Analysis.** We created a decision tree by analyzing our database of known true and false articles and identifying the features that distinguish them. The decision tree is built by selecting a feature that it believes is important in distinguishing true and false news articles. It then splits the data into two groups based on this feature and repeats the process for each group until we have a tree that accurately classifies the data. We then repeat the process for each group until we have a tree that accurately classifies the data. Once we have created the decision tree, we can explore it backwards to see the rule that it applies in each split. This can help us to understand the features that are most important in distinguishing true and false news articles and to identify patterns of misinformation.

With these ideas and using the data previously labeled, we built our no-machine-learning algorithm.

Evaluation The dataset is split randomly into training and validation sets with a 80-20 split ratio. Various classification algorithms were trained on the training set using a 10-fold cross validation, and the model's hyperparameters were optimized using a parameters grid and repeating the process of training with a of these parameters. The performance of these models is then evaluated on the validation set using accuracy as the performance metric. However, to gain a deeper understanding of the model's performance and potential areas of improvement, we also considered other metrics such as F1-score, that is used to give more importance to costly misclassifications. Different models have been compared, and they are described in the table of section 3.2 where results are discussed.

3 Results and Discussion

3.1 Machine Learning Module

We conducted a comprehensive evaluation of different approaches. We compared the performance of the following models:

- Baseline Model: The 'roberta-fake-news' model in its original configuration.
- 'roberta-fake-news' Embeddings generated by 'roberta-fake-news' model. + Classical Classifiers such as SVM, MLP, and RF.
- Fine-tuning 'roberta-fake-news' + CNN with Evidence, Sentiment, and Toxicity, We fine-tuned the 'roberta-fake-news' model using a 1D Convolutional Neural Network (CNN) to incorporate evidence, sentiment, and toxicity information. We evaluate the impact of considering different numbers of evidence sentences (k = 0, 1, 2, 3, 4, 5) on the model's performance, as shown in Fig.1.

The models have been trained on a dataset consisting of 2835 claims, of which 1577 are labeled as true claims and 1258 as false claims. These claims were used to train the models and teach them patterns and relationships between tokens to make predictions on new, unseen data. After training, we evaluated the models' performance on a separate test set containing 315 claims. Among these claims, 171 were labeled as true claims and 144 as false claims.

Attention Between Claims and Evidences is All You Need: Tackling Misinformation in Nigeria with Transformers and 1D Convolutional Networks

We will perform fine-tuning on the "roberta-fake-news" model, which has 125 million trainable parameters and 12 attention heads. Our goal is to adapt it specifically to the task of misinformation detection in Nigeria. The input to the transformer model will consist of tokens representing the claim we want to verify, followed by a special separation token "SEP", and tokens representing the top "k" (initially 5) evidence sentences that are closest to the claim. To determine these closest evidence sentences, we will calculate the cosine similarity between the embeddings using the "all-MiniLM-L6-v2" model.

Instead of using the "[CLS]" token for classification, we propose applying one-dimensional convolutional operations with a specific number of filters. Subsequently, we will apply one-dimensional max pooling layers successively until we obtain a flattened vector with the final outputs.

This approach aims to capture relationships between the claim tokens and the evidence tokens to predict whether it is misinformation or not. By leveraging

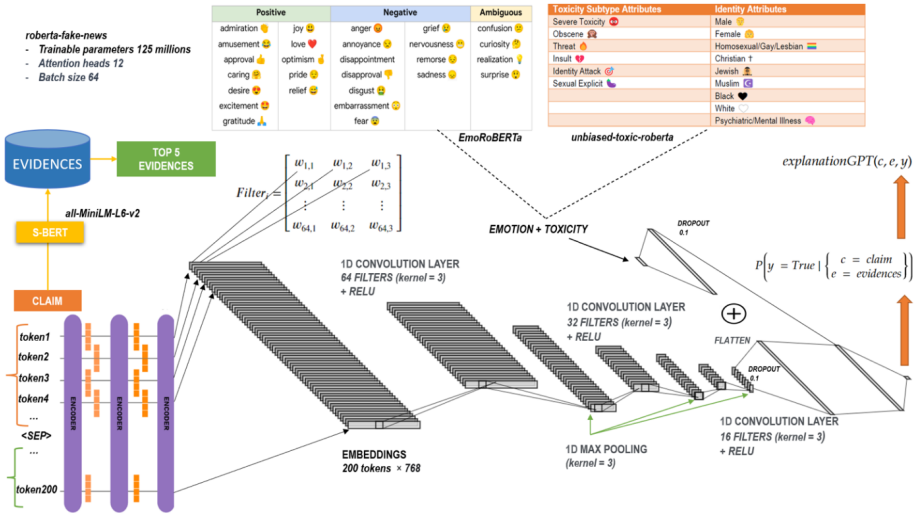


Fig. 1: Fusion of Transformer, 1D Convolutional Networks, and Variable Concatenation for Misinformation Detection in Nigeria

the attention capabilities of the transformer and utilizing one-dimensional convolutional operations, we aim to capture relevant features that can help discern between accurate and misleading information.

After obtaining the representation from the convolution process, we concatenate it with a vector associated with the outputs of a dense layer applied to other types of variables that we have considered. These variables include emotions extracted from the claim using the EmoRoBERTa transformer. EmoRoBERTa is trained on the GoEmotions dataset, which consists of 58,000 Reddit comments labeled for 27 emotion categories, including Neutral.

Additionally, we have incorporated various toxicity variables using the unbiased-toxic-roberta transformer. This model is trained on The Civil Comments dataset, which comprises approximately 2 million public comments from the now-closed Civil Comments platform. The dataset is annotated for toxicity and other attributes, including identity labels, to facilitate research on improving online conversations. Moreover, the model takes into account different identity labels.

By incorporating these emotion and toxicity variables into the model architecture, we aim to capture additional contextual information and potential correlations with misinformation detection in Nigeria.

3.2 Fact-checking module

In order to build the fact-ckecking module we have tried different approaches besides the one explained in Section 2.2.2.

The first approximation was to clean the text removing hashtags and stopwords and vectorize it using a TDF-IDF representation and using a decision

tree to classify whether the claim is True or False decided to try with a decision tree and classify. Surprisingly, we obtained a F1-score of around 0.85, which is a pretty good result keeping in mind that we only use text as the input with no other features.

After, we thought of including sentiment analysis using the NLTK library in Python (sentiment intensity analyzer) and from the TextBlob library we used the sentiment.subjectivity module and the sentiment.polarity, to be able to classify the tweet as positive, negative or neutral. We combined both data frames and ran a Grid Search to find the best hyperparameters for the tree, hoping for better results. Unfortunately, results were a bit below the approach where we used just text (getting around 0.82 of F1-score), but we decided to represent the tree because it can give us more information.

We also created a model using just sentiments, and even though the predictive capacity of the model is a bit better than randomness, we obtained interesting conclusions from the polarity variable. We observed that positive tweets are classified as True and negative ones as False. Figure 8 shows the tree generated by the model including text and sentiment.

Using a Decision Tree with a maximum depth of 4 we get a F1-score of 0.824. We can observe that the first and most important splits are word based using words like "elections" or "Nigeria elections". Sentiments come after in the splits of the tree and we can deduce that they are less important for the classification.

3.3 Results

The combination of embeddings and classical classifiers, such as SVM, MLP, XGBoost, and RandomForest, proved to be much more effective. These approaches achieved F1-Scores close to 0.9, indicating good performance in detecting misinformation.

There was a beneficial effect when adding information on toxicities and emotions to the embeddings and classical classifiers models.

The fine-tuning technique using a combination of CNN and k evidences, where k represents the number of evidences used, also proved to be effective. The obtained F1-Scores were around 0.9. It was observed that adding evidence to the claims significantly improved the model's performance, increasing the F1-Score from 0.81 (without evidence) to almost 0.9 (with the closest evidence based on cosine similarity).

As the number of evidences (k) increased, the model's performance continued to improve, reaching an F1-Score of 0.907. This indicates that providing the model with more supporting information enhances its ability to detect misinformation more accurately.

3.4 Deployment

We deployed our misinformation detection pipeline through a prototype application, which aims to address the objectives of the project. Our prototype

Approach	Model	F1-Score	Precision	Recall	Cost
Baseline	*roberta-fake-news*	0.537	0.460	0.643	1717
Embeddings + Classical Classifiers					
	SVM	0.907	0.907	0.912	265
	MLP	0.870	0.858	0.883	390
	XGBoost	0.884	0.879	0.889	343
	RandomForest	0.910	**0.938**	0.883	240
Embeddings + Classical Classifiers with emotions and toxicities					
	SVM	0.907	0.907	0.912	265
	MLP	0.897	0.900	0.895	296
	XGBoost	0.904	0.897	0.912	285
	RandomForest	**0.924**	0.924	**0.924**	**221**
Fine-tuning + CNN + k Evidences with emotions and toxicities					
	k=0	0.807	0.912	0.725	444
	k=1	0.898	0.920	0.877	308
	k=2	0.886	0.908	0.865	325
	k=3	0.897	0.900	0.895	297
	k=4	0.905	0.916	0.895	273
	k=5	0.907	0.927	0.889	276

Table 1: Performance Evaluation of Misinformation Detection Approaches

application consists of a user-friendly interface, similar to a chatbot, that allows users to input a claim and receive a detailed analysis of the veracity, including the sentiment and toxicity analysis to explain why our model is predicting that label. It is created to help human fact checkers to reduce their work time in searching and finding evidences to put a label on each claim, giving them some reliable model to lean on. The goal of this is not to replace human labor in this kind of task. A human team is needed to supervise the model.

By automating the detection process and providing detailed explanations for the assigned labels, our prototype addresses the challenges of outdated models and the availability of annotated data, which is the main value that we wanted our project to have.

4 Conclusion

First of all we got to mention the easy, fast and efficient way GPT-3 has given us to label the unlabelled data. When data is scarce or the labeling task is so tedious, it is a pretty good option. It could work even better with the supervision of human experts. We observed that state-of-the-art transformers and vector representations like TF·IDF with simple models perform really well. Sentiment analysis can be very useful to improve the explicability of our models and help the final user understand the decision taken for each instance by certain models. Definitely, misinformation detection is a very challenging issue, but with a good understanding of the problem, hard work, fresh ideas, initiative and a proactive team, everything is possible.

References

1. N. Kotonya and F. Toni, "Explainable Automated Fact-Checking for Public Health Claims," Oct. 2020, arXiv:2010.09926 [cs]. [Online]. http://arxiv.org/abs/2010.09926
2. P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. D. S. Martino, "Automated Fact-Checking for Assisting Human Fact-Checkers," Mar. 2021, arXiv:2103.07769 [cs] version: 1. [Online]. http://arxiv.org/abs/2103.07769
3. Requests: HTTP for Humans™ — Requests 2.30.0 documentation https://requests.readthedocs.io/en/latest/
4. Selenium with Python — Selenium Python Bindings 2 documentation https://selenium-python.readthedocs.io/
5. Home - iVerify Zambia https://iverify.org.zm/
6. Politifact https://www.politifact.com/
7. Africa Check hhttps://africacheck.org/fact-checks
8. Dubawa https://dubawa.org/
9. African Fact Check https://factcheck.afp.com/
10. Lead Stories https://leadstories.com/
11. Dispatch https://thedispatch.com/category/fact-check/
12. News Verifier Africa https://newsverifierafrica.com/

13. The Guardian Nigeria News - Nigeria and World News https://guardian.ng/
14. Punch Newspapers - Breaking News, Nigerian News, Entertainment, Sport, Business and Politics https://punchng.com/
15. Daily Post Nigeria - Nigeria News, Nigerian Newspapers https://dailypost.ng/
16. Sahara Reporters https://saharareporters.com/
17. The Nation Online Nigeria Newspaper https://thenationonlineng.net/
18. Vanguard Nigeria News https://www.vanguardngr.com/
19. OpenAI Product https://openai.com/product
20. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Dec. 2017, arXiv:1706.03762 [cs] version: 5. [Online]. http://arxiv.org/abs/1706.03762
21. arpanghoshal/EmoRoBERTa · Hugging Face, Jan. 2023 https://huggingface.co/arpanghoshal/EmoRoBERTa
22. unitary/unbiased-toxic-roberta · Hugging Face https://huggingface.co/unitary/unbiased-toxic-roberta
23. Reuters Fact Check https://www.reuters.com/fact-check/about

May Maths Be With You

Damian Oussa Vañó Fernández[1]

Double Degree in Mathematics and Civil Engineering
Universitat Politècnica de València, Camí de Vera s/n Valencia, 46022, Spain
dovaofer@cam.upv.es

Abstract. We are the resistance, we are the future. In a universe reigned by caos where the Empire rules with tyranny, we fight for the weakest. It has come to our knowledge that they have created a weapon, which has the power to destroy entire worlds and turn them into dust, named the Death Star. However we have discovered that we do have a chance to counterattack and destroy it. Our mission is to infiltrate the computer system and destroy it from within, but to do so we must solve math problems of great difficulty. This is the reason we have an ace up our sleeve: a numerical method. Will we be able to destroy the Death Star and save the universe?

With this preposition, from the subject of numerical resolution of linear and no linear systems we will study the convergence and behavior of a family of numerical methods and discover the vast universe behind it.

Keywords: numerical analysis, fixed point, critical point, error equation, parameter plane, dynamic plane and rational function

1 Study of The Family of Methods

Given an initial estimate x_0, the iterative expression is

$$y_k = x_k - \frac{f(x_k)}{f'(x_k)}$$

$$x_{k+1} = y_k - \frac{f(y_k)}{(1-\alpha)f[x_k, y_k] + \alpha f'(x_k)}$$

where α is a free parameter (real or complex)

1.1 Extension to systems of equations

Is the assigned method directly extensible to systems of equations?
Analyzing the expression of our family for equations, we observe that in the denominator of the second equation, when we pass to systems we will obtain a sum of matrices since we will have the divided difference operator and a Jacobian matrix and we will be able to perform its inverse directly because we will not have vectors. That is why we can directly extend the method to a system of equations. This is the result:

$$F(x) = 0$$
$$y^{(k)} = x^{(k)} - [F'(x^{(k)})]^{-1} F(x^{(k)})$$
$$x^{(k+1)} = y^{(k)} - [(1-\alpha)[x^{(k)}, y^{(k)}; F] + \alpha F'(x^{(k)})]^{-1} F(y^{(k)})$$

1.2 Family Error Equation and Order of Convergence

We know that in equations the order of a method is at most 2^{d-1} where d are their functional evaluations, in this case for our method $d = 3$ since we have $f(x), f'(x)$ and $f(y)$, then the order of our family will be at most 4, this order at moving to systems may be lower or remain the same.
The development carried out in Mathematica provides us with the following results:
General error equation: $e^{k+1} = (1+\alpha)C_2^2 e_k^3 + (-((3+5\alpha+\alpha^2)C_2^3) + (3+4\alpha)C_2C_3)e_k^4 + O(e_k^5)$
With this expression of the result we can see a very important thing and that is for $\alpha = -1$ the coefficient that is with e_k^3 will be canceled and the method will be order 4.
With $\alpha = -1$ we obtain this error equation: $e^{k+1} = (C_2^3 - C_2C_3)e_k^4 + O(e_k^5)$
Order of convergence:

- $\alpha \neq -1 \longrightarrow$ Order of convergence 3
- $\alpha = -1 \longrightarrow$

  Order of convergence 4

With the demonstration done by hand taking into account the non-commutativity of the operators we obtain this error equation: $e^{(k+1)} = (1+\alpha)C_2^2 e_k^3 + ((2+2\alpha)C_2C_3 + (1+2\alpha)C_3C_2 + (-3-5\alpha-\alpha^2)C_2^3 - 3C_2C_3 + 3C_3C_2)e_k^4 + O(e_k^5)$

With the results obtained we can conclude two things: the order of convergence obtained in Mathematica is preserved and as we had seen previously, the order of convergence is not the same for all values of alpha.
$\alpha \neq -1 \longrightarrow$ Order of convergence 3
$\alpha = -1 \longrightarrow$ Order of convergence 4 and the method with that value will be an optimal method.

1.3 Rational Function

Which is the rational function associated with our method resulting from applying it to $p(x) = (x-a)\cdot(x-b)$? And after applying the Möbius transformation? What we have done is find the polynomial of second degree and its derivative and thus calculate the rational function associated with our method, where we write the expression of our family with the polynomial. First we calculated the first step which was, $y_k = x_k - \frac{f(x_k)}{f'(x_k)}$, next we calculate the finite differences factor,

which by definition is $dd = \frac{f(x_k)-f(y_k)}{x-y}$, to be able to calculate the operator of our family at the end x_{k+1}.

```
In[1]:= p[x_] = (x - a) (x - b);
        dp[x_] = D[p[x], x]
                 |deriva
Out[2]= -a - b + 2 x

In[9]:= y = Factor[x - p[x] / dp[x]]
          |factoriza
        dd[x_] := ((p[x] - p[y]) / ((p[x]) / (dp[x])))
        RC[x_, α_] = y - p[y] / ((1 - α) * dd[x] + (α * dp[x]));
Out[9]= (-a b + x^2) / (-a - b + 2 x)
```

Rational Function Obtained: $RC = \frac{-ab+x^2}{-a-b+2x}$.
Now we apply the Möbius transformation to what is obtained as a rational function (RC). This is used in order to obtain the rational operator without variables a and b. As we see, we have obtained what we expected, our rational function only depends on x and $alpha$.

```
In[6]:= M[u_] = (u - a) / (u - b);
        iM[u_] = (u * b - a) / (u - 1);
        R[x_, α_] = Simplify[M[RC[iM[x], α]]]
                    |simplifica
Out[8]= x^3 (1 + x + α) / (1 + x + x α)
```

Cayley Test: An iterative method is said to satisfy the Cayley test if its associated rational function, after the Möbius transformation, is z^n, $n \in \mathbb{N}$.
Why have we announced this result? Because when we apply the Möbius transformation to $\alpha = -1$ the Cayley test is fulfilled since we have x^4, which assures us furthermore, the method for $\alpha = -1$ has convergence order 4. It is also very good for the stability of the method since we will not have free critical points. For $\alpha \neq -1$ we see that in the numerator the rational operator has a factor x^3, which ensures that for all $\alpha \neq -1$ the order of convergence of our family is 3.

So in conclusion the order of convergence for $\alpha = -1$, is 4 and for $\alpha \neq -1$, is 3.

We also corroborate what was previously obtained with Mathematica, in the

demonstration of the order. Finally we check with Traub's Theorem the order of convergence. As we see, the first and second partial derivatives with respect to any α are null, and the third is different from 0, therefore order 3. The first, second and third partial derivatives are zero and the fourth is different from 0, therefore for $\alpha = -1$, the order is 4.

1.4 Fixed Points

Are all the fixed points directly related to the roots of the polynomial? Are there strange fixed points? We are going to calculate and analyze the stability of all fixed points, clearly indicating the reasons why they are said to be attractors, repellers or parabolics.
To obtain the fixed points we must equate the rational function R obtained in the part before to x and solve the equation.

```
In[35]:= Solve[R[x, α] == x, x]
         resuelve
Out[35]= {{x → 0}, {x → 1}, {x → 1/2 (-2 - α - √α √(4 + α))}, {x → 1/2 (-2 - α + √α √(4 + α))}}
```

We have to see if infinity is also a fixed point. For this we define the inverse of rational operator and if 0 appears, in this case it will be infinity.

```
In[38]:= In7[x_, α_] = 1 / R[1 / x, α] // Simplify;
                                        simplifica
         In7[0, α] // Simplify
                      simplifica
Out[39]= 0
```

Not all fixed points are directly related to the roots of the polynomial. The fixed points that are related are 0 and infinity, which are roots of the polynomial second grade. There are three strange fixed points, they are 1, ex1 and ex2.

$$\texttt{ex1}[\alpha\_] = \frac{1}{2}\left(-2 - \alpha - \sqrt{\alpha}\,\sqrt{4+\alpha}\right);$$
$$\texttt{ex2}[\alpha\_] = \frac{1}{2}\left(-2 - \alpha + \sqrt{\alpha}\,\sqrt{4+\alpha}\right);$$

We observe that the points ex1 and ex2 are conjugate. The strange fixed point z = 1 comes from the divergence of the original method, which when applying Möbius becomes a fixed point. Now we are going to analyze the stability of the fixed points. To do this we define the derivative of the rational operator.

For $z = 0$, the operator at the fixed point $z = 0$ is canceled for any value of α so it will be super attractor, this corroborates what was concluded in the previous section, where we had said that zero was the root of the polynomial, that is why its derivative is 0.

For $z = \infty$, to analyze this fixed point, we define the derivative operator of the inverse of the rational operator. For any value of α the operator is canceled, so the fixed point $z = \infty$ is a super attractor, this corroborates what was concluded in the previous section, where we had said that infinity was the root of the polynomial, that is why its derivative is 0.

For $z = 1$, what is this fixed point? To do this, we must evaluate for which values of α the fixed point $z = 1$ is a super attractor, attractor, repulsor or neutral or parabolic. We will carry out the demonstration to know when it is an attractor, but with the final result we can apply it directly to the rest of cases.

Demonstration

$$|2 + \frac{2}{2+\alpha}| < 1$$
$$|\frac{2(2+\alpha)+2}{2+\alpha}| < 1$$
$$|2(2+\alpha)+2| < |2+\alpha|$$

Now we replace $\alpha = c + id$ $c, d \in \mathbb{R}$

$$|2(2+(c+id)+2| < |2+(c+id)|$$
$$\sqrt{(6+2c)^2+(2d)^2} < \sqrt{(2+c)^2+(d)^2}$$
$$32 + 20c + 3c^2 + 3d^2 < 0$$
$$\frac{32}{3} + \frac{20}{3}c + c^2 + d^2 < 0$$
$$-\frac{4}{9} + c^2 + (\frac{10}{3})^2 + \frac{20}{3}c + d^2 < 0$$
$$(c + \frac{10}{3})^2 + d^2 < \frac{4}{9}$$
$$(c + \frac{10}{3})^2 + d^2 < (\frac{2}{3})^2$$
$$|\alpha + \frac{10}{3}| < \frac{2}{3}\square$$

From this demonstration we obtain very interesting information: for $\alpha = -2$, $z = 1$ is not a fixed point.
The section where we can see the stability of the fixed point will be in the

circumference $(a+\frac{10}{3})^2+b^2=(\frac{2}{3})^2$, in the center circle $x=(-\frac{10}{3},0)$ and radius $r=\frac{2}{3}$.
Let's see for which value of α $z=1$ is a super attractor fixed point.

$$|2+\frac{2}{2+\alpha}|=0$$

$$|\frac{2(2+\alpha)+2}{2+\alpha}|=0$$

$$|2(2+\alpha)+2|=0$$

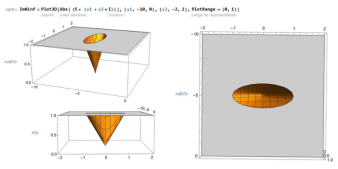
$$|6+2\alpha|=0 \longleftrightarrow \alpha=-3$$

Therefore, we can now classify the stability of the fixed point $z=1$ according to the α values:
If $|\alpha+\frac{10}{3}|<\frac{2}{3}$, the point $z=1$ is a fixed attractor point and if $\alpha=-3$ then it will be a super attractor fixed point.
If $|\alpha+\frac{10}{3}|>\frac{2}{3}$, the point $z=1$ is a repulsor fixed point.
If $|\alpha+\frac{10}{3}|=\frac{2}{3}$, the point $z=1$ is a neutral or parabolic fixed point. In the graphical representation we have a circumference of center $x=(-\frac{10}{3},0)$ and ratius $r=\frac{2}{3}$. The gray area is where the $z=1$ fixed point is repulsor. In the same circumference is when it is neutral or parabolic and in the orange zone it is when it is an attractor and at the tip when the derivative operator is worth 0, that is, it is a super attractor, specifically at $\alpha=-3$. We will be interested in selecting values of α for which $z=1$ falls in the gray zone and is repulsive and thus the only ones attractor fixed points are 0 and infinity.

For $z=ex1 \wedge z=ex2$ we have obtained directly the expression by which we can classify the stability of the fixed points ex1 and ex2.
Let's see when the point is super atractor.

$$|\alpha+5|=0 \longleftrightarrow \alpha=-5$$

If $|\alpha+5|<1$, $z=ex1$ y $z=ex2$ are fixed attractor points and if $\alpha=-5$ then the strange fixed points ex1 and ex2 are superattractors.
If $|\alpha+5|>1$ the points $z=ex1$ and $z=ex2$ are repulsor fixed points.
If $|\alpha+5|=1$ the points $z=ex1$ and $z=ex2$ are neutral or parabolic fixed points.
We can see everything studied about these strange fixed points represented in the following graph.



We can observe the circle with center $x = (-5, 0)$ and radius $r = 1$. If we fall in the gray area the points ex1 and ex2 will be repulsor. If we fall on the circle, then they will be neutral or prabolic and if we fall in the orange zone it will be an attractor, and if we fall at the peak, $\alpha = -5$, they will be super attractors. As we have commented previously, we will be interested in falling into the gray zone so that they are repulsor and thus have only 0 and infinity as fixed attractor points.

Now we will be interested in finding the alpha values where the strange fixed points they match.
We observe that ex1 and ex2 coincide for the values of $\alpha = -4 \wedge 0$. Whith $\alpha = -4$ take the value of $z = 1$, it is a neutral or parabolic point and with $\alpha = 0$ take the value of $z = -1$ which is a repulsor point. We are going to see if with some value of α the strange fixed points values may coincide with 0 or infinity and we found that they will never coincide, so that is very good for our method.

So we can conclude:
For $\alpha = -4$ the strange fixed points ex1 and ex2 coincide with the fixed point $z = 1$ and it is a neutral or parabolic point.
For $\alpha = 0$ the strange fixed points ex1 and ex2 are the point $z = -1$ which is a repulsor point.
The strange fixed points ex1 and ex2 never coincide with the fixed point $z = 0$ nor $z = \infty$
To finish the question we are going to see what happens with our $\alpha = -1$ where the method was of order 4. For $\alpha = -1$ the strange fixed points ex1 and ex2 are repulsive and for $z = 1$ it is also repulsor, so $\alpha = -1$ is a very good value for our family since the only fixed points will be 0 and infinity.

1.5 Critical Points

We are going to calculate the critical points of the operator R. Are all the critical points directly related to the roots of the polynomial p(z) or are there free critical points?
To see the critical points we must set the derivative operator of the rational function equal to zero and solve the equation.
In Mathematica we have solved the derivative operator equal to zero and we have obtained as critical points $z = 0$, and two free critical points.

$$\text{pce1}[\alpha\_] = \frac{-3 - 2\alpha - \alpha^2 - \sqrt{-6\alpha + \alpha^2 + 4\alpha^3 + \alpha^4}}{3(1+\alpha)}$$

$$\text{pce2}[\alpha\_] = \frac{-3 - 2\alpha - \alpha^2 + \sqrt{-6\alpha + \alpha^2 + 4\alpha^3 + \alpha^4}}{3(1+\alpha)}$$

We also solve the operator equation derived from the inverse of the rational function equal to zero where we obtain that $z = \infty$ is also a critical point.

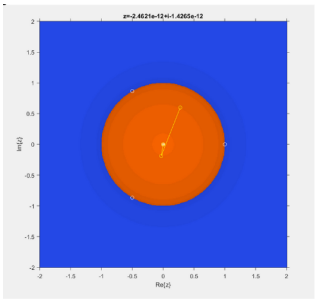
The answer to the question proposed is no, since the only critical points directly related to the roots of the polynomial are 0 and infinity, and if that we have free critical points that are pce1 and pce2, which are also conjugate. But it should be emphasized that for $\alpha = -1$ there are no free critical points since the denominator of these is cancelled. This corroborates what was obtained before where we had obtained that $\alpha = -1$ satisfied the Cayley test and therefore would not have free critical points.

Is important that pce1 and pce2 will have the value of 1 with $\alpha = -3$ and $\alpha = -2$, and the value of -1 with $\alpha = 0$ and $\alpha = 1$. We also checked to see if there were more alpha values so these points coincide but there are none. Also, the critical point $z = -1$ will be a pre-periodic point of 1.

With $\alpha = -3$ the free critical points coincide with the fixed point $z = 1$ and for this value exactly the point will be super attractor.
With $\alpha = -2$ the free critical points coincide with $z = 1$ but exactly at this case was the value for which $z = 1$ was not a fixed point.
With $\alpha = 0$ and $\alpha = 1$ we obtain the critical point $z = -1$ which is always pre-image of the strange fixed point $z = 1$ where will be a point.
Also, any value of α of the free critical points coincide with the 0 or infinity.

1.6 Analisis Of Dynamic Planes Of Different α Values

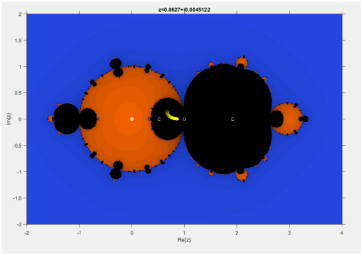
For $\alpha = -1$: Op = $x \wedge 4$, Fixed Points = 0, 1, -0.5-0.8660i, -0.5 + 0.8660i, Critical Points = 0, 0, 0 and Attractor Points = 0.



With the Op we corroborated what was previously studied where we saw that for this value of α the method fulfilled the Cayley test and we observed that there is no free critical point. We clearly observe that there are no black areas in the dynamic plane. The critical points that are 0 and infinity are in the basins of attraction of 0 and infinity.

For $\alpha = -4$: Op = $-(x \wedge 3 * (x - 3))/(3 * x - 1)$, Fixed Points = 0, 1, 1,

1, Critical Points = 0, 0.5195, 1.9250, Attractor Points = 0.



For this value of α the strange fixed points ex1 and ex2 coincide with the point fixed $z = 1$ and this point was neutral or parabolic.
In this dynamic plane we do have black regions. The black areas correspond to a basin where there may be strange fixed points attractors or periodic attractor orbits. The critical point $z = 0$ is in the corresponding orange basin of attraction to the supertractor fixed point $z = 0$. The critical point $z = \infty$ is in the blue basin of attraction that corresponds to the fixed point super attractor $z = \infty$. The free critical points $z_1 = 1.925$ and $z_2 = 0.5195$ are in the black region and it is something negative for the stability of the method since in the black areas there may be attractor orbits or strange attractor fixed points but it should be noted that in dynamic sense provides you more interest.

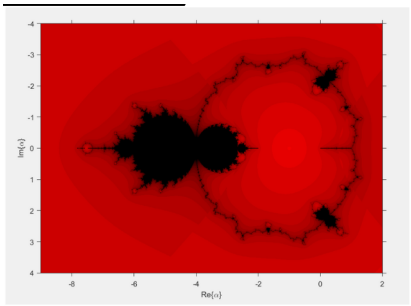
1.7 Parameter Plane

The existence of the red regions is directly related to the values of the parameter α, in this case if it is red it means that the free critical point with that The value of α will converge to 0 or infinity and therefore in the dynamic planes we will find stable schemes. The wider the red areas the better since we will have more values of α for which the family will present stable schemes.
The existence of black regions is also directly related to the values of the parameter α, in this case if it is black it means that the critical point free with that value of α will end up at a strange fixed point or attractor or an attractor orbit and therefore in the dynamic planes we will find unstable schemes.

We are going to generate the parameter planes of our free critical points.

In this case we only have two and since they are conjugated they will have the same parameter plane so we will only have to generate a parameter plane.
We have to identify the regions of the parameter plane that correspond with the stability of the strange fixed points, we look for the center circle x = (-5, 0) and radius 1 that we had obtained before.

Now we are going to search for the period 3 orbit. We are going to search



in the black areas, since as we have commented is where unstable schemes will appear in dynamic planes and therefore black areas will surely appear that will be where our periodic orbits will be.
We have obtained an orbit of period 3. The dynamic plane is for α = - 5.603736128236745 - 1.007422934648581i. The dinamic plane gave us one node of the orbit, $z = 3.7153 + 1.1343i$. Therefore by Sharkovsky's Theorem, the existence of period 3 orbits guarantees the existence of orbits of any period. This something negative for the stability of the method since it may be that our method does not converge but that in the dynamic sense gives it more interest.

1.8 Conclusion

We have been able to observe that the divided difference can add a certain dynamic interest to the family since it allows us to obtain a great variety of different methods and not only varieties that start from the Traub method. Numerically speaking it has much interest because with the study carried out in general the method is very stable, where generally, we obtain the solution and everything within optimal conditions, that is, with correct behavior of the convergence rates and the ACOC. It is also worth noting that there are several values of α that satisfy the Cayley test which is not very common and is also very good for our family since we are talking about optimal methods.

Also, thanks to our family of numerical methods we have been able to destroy the Death Star and save the universe.

References

1. A. Cordero, J. García-Maimó, Juan R. Torregrosa, P. Vindel, Chaos in King's iterative family, AML 26 (2013) 842–848.
2. I.F. Traub, Iterative Methods for the Solution of Equations. Prentice-Hall, Englewood Cliffs, New York, 1964.
3. J. Milnor. Dynamics in one complex variable. Stony Brook IMS prepint (1990)

Drawing fractals with Matlab: Parameter planes and dynamical planes for families of iterative methods.

Jorge Rico[1]

Double Degree in Mathematics and Civil Engineering,
Universitat Politècnica de València, Camino de Vera, s/n, 46022-Valencia, Spain,
jricest@cam.upv.es

Abstract. Motivated by the second course of numerical analysis in the mathematics degree, this work presents a Matlab® App Designer application that facilitates the interpretation of Parameter Planes and Dynamical Planes based on some families of iterative methods. These visual representations play a crucial role in the correct analysis of the dynamical behavior of these families. One of the purposes of this resource is its use in teaching.

Keywords: Parameter planes, dynamical planes, non-linear equations, families of iterative methods, dynamical behavior, Matlab App Designer

1 Introduction

In the first and second courses of numerical analysis in mathematics, we, the students, learn the importance of using iterative methods to solve non-linear equations, $f(z) = 0$, with $f : \mathbb{C} \to \mathbb{C}$.

First we are taught Newton's well-known method. We soon become experts in multi-step methods such as Traub's, Chebyshev's, Halley's and others. Not very late we are taught to study families of iterative methods by introducing complex parameters, to analyse in depth the order of convergence, the efficiency index, the approximate computational order of convergence, etc... Finally, we study in detail rational functions derived from these families, whose dynamics are not well known.

In the literature, interesting dynamic planes are created to study specific families that contain some periodic behaviour and other anomalies. However, unlike the dynamic planes, the parameter planes associated with a family of methods allow us to understand the behaviour of the different members of the family of methods and help us to choose a particular one.

In teaching, Matlab is used to program these families and draw the dynamic and parameter schemes. This programme has an exponential learning curve that enables students to go beyond their initial knowledge. Therefore, the aim of the present work is to create a resource in Matlab App Designer that facilitates the interpretation and drawing of the dynamic diagrams.

The rest of the paper is structured as follows. Section 2 is dedicated to the presentation of basic concepts of stability of the methods introduced by complex dynamics. In this sense, the rational function resulting from the application of families to quadratic polynomials is studied and the fractals generated in the planes are explained. In Section 3, a flow chart is presented, as well as the families used and the different functions that have the computational resource. Finally, Section 4 summarises the conclusions and the objectives achieved and to be achieved.

2 General complex dynamics features

Some basic concepts of complex dynamics are introduced in this section. First, a quadratic polynomial is applied to the families of iterative methods for solving non-linear equation. A rational operator is obtained and its dynamic behaviour is analysed as a function of the initial estimation. Finally, the parameter and dynamic planes and the procedure for generating them are defined.

2.1 Rational Operators in Quadratic Polynomials

Given a family for solving non-linear equations depending on the parameter $\alpha \in \hat{\mathbb{C}}$. Let $p(z) = (z-a)(z-b)$ be a polynomial defined on $\hat{\mathbb{C}}$ and, $a, b \in \hat{\mathbb{C}}$ the roots of the polynomial. Therefore, $R_{\alpha,a,b} : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ is defined as the rational function, defined on the Riemann sphere, obtained by applying to the family the polynomial. Let us recall that it depends on the parameter of the family and on the roots of the polynomial.

To obtain a one-parametric operator, we apply the Möbius transformation.

$$h(u) = \frac{u-a}{u-b},\ [h(u)]^{-1} = \frac{ub-a}{u-1} \tag{1}$$

It is obtained a rational operator associated with the family that does not depend on a and b, only on the parameter, $R_\alpha(z) = (h \circ R_{\alpha,a,b} \circ h^{-1})(z)$. Then the dynamics of the family on every quadratic polynomials can be studied by analysing it. In addition, the Möbius transformation h maps its roots a and b to $z = 0$ and $z = \infty$, respectively.

2.2 Basics on Complex Dynamics

Based in [2] we will examine some fundamental concepts of complex dynamics, which will be subsequently defined and applied for further use. Let $R_\alpha : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ be the rational function defined on the Riemann sphere. The **orbit of a point** $z_0 \in \hat{\mathbb{C}}$ is given by the set of its images by R_α as follows,

$$\mathscr{O}(z_0) = \left\{ z_0, R_\alpha(z_0), R_\alpha^2(z_0), \ldots, R_\alpha^n(z_0), \ldots \right\} \tag{2}$$

A point $z_f \in \hat{\mathbb{C}}$ is a **fixed point** if $R_\alpha(z_f) = z_f$. Note that in this case, 0 and ∞ are fixed points of the operator R_α. Fixed points that do not agree

with a root of $p(x) = 0$ are **strange fixed points**, these can be classified as follows. A fixed point z_f is called an **attractor** if $|R(z_f)| < 1$, **superattractor** if $|R(z_f)| = 1$, **repulsor** if $|R(z0)| > 1$ and **parabolic** if $|R(z0)| = 1$. Many works in the literature, (see [5], [6], [7], [8]) which analyse specific families, try to find methods where the strange fixed points are repulsors.

For an attracting fixed point z_f of the rational function R_α its **basin of attraction** is defined as the set of its pre-images of any order such that

$$\mathscr{B}(z_f) = \left\{ z \in \hat{\mathbb{C}} : \lim_{n\to\infty} R_\alpha^n(z) = z_f \right\} \tag{3}$$

The set of points in the Riemann sphere whose orbits, $\mathscr{O}(z_0)$, tends to an attracting fixed point z_f is defined as the Fatou set, $\mathscr{F}(R_\alpha)$. The complementary set in $\hat{\mathbb{C}}$, the Julia set $\mathscr{J}(R_\alpha)$.

A **periodic point** z_p of period $p > 1$ is a point such that $R_\alpha^p(z_p) = z_p$ and $R_\alpha^k(z_p) \neq z_p$, for $k < p$. A **pre-periodic point** is a point z_{pp} that is not periodic but there exists a $k > 0$ such that $R_\alpha^k(z_{pp})$ is periodic. An important result, Sharkovsky's theorem (see [3]), states that if a periodic orbit of period 3 exist then periodic orbits of arbitrary periodicity can appear.

A **critical point** z_c is a point where the derivative of the rational function vanishes, $R'_\alpha(z0) = 0$. In addition, a **free critical point** is the critical point that does not match the roots of the polynomial. Another classical result, [4], establishes that there is at least one critical point associated with each immediate invariant Fatou set.

2.3 Dynamical Planes

The process of generating a dynamical plane is straightforward, aimed at gaining a visual understanding of a family method's behaviour.

Let be the rational operator that associates any function with an iterative method, the dynamical plane illustrates the basins of attraction of the operator. In this case, the function is the quadratic polynomial and the method is one of the family given by a specific parameter, and then applying the Möbius transformation, the operator $R_\alpha(\alpha$ known) is obtained. A map of the complex plane is defined and the orbit of every point in the dynamical map is studied. The dynamical plane can be visualised whether the method converges to a fixed point, or not, and its speed of convergence. The different points of the orbit can also be represented visually. Normally, each basin of attraction is drawn with a different color, using bright colours for the basins of attraction of attracting fixed points and black for the basins of attraction of attracting periodic points. In addition, the brightness of the colour indicates the number of iterations needed to reach the fixed point.

In Figure 1 we have drawn, first, a dynamical plane with a periodic orbit of period 3 and, second, a plane that has two attracting strange fixed points. Both maps belong to different parameters of the well-known King's family. The dynamics of the King's family are studied here [5].

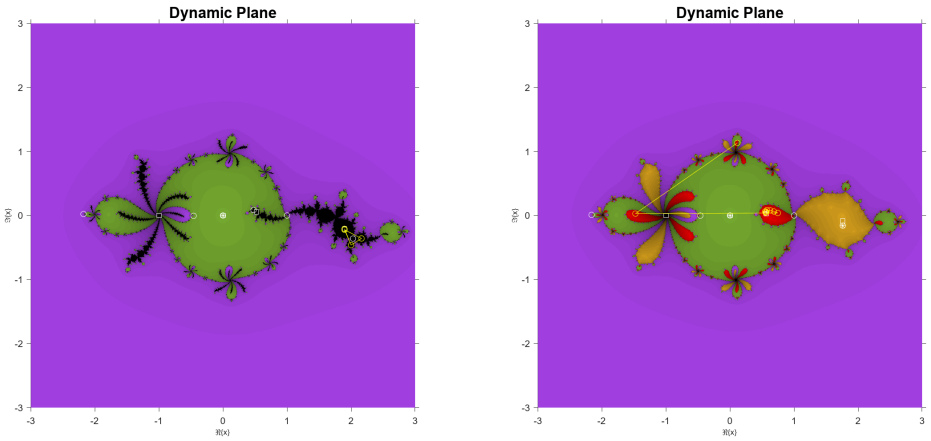


Fig. 1: Some dynamical planes

2.4 Parameter Planes

As well as dynamical planes, parameter planes are also relatively easy to draw. The main idea behind the creation of these maps is the result discussed above, which states that in each invariant Fatou component is associated with at least one critical point. Therefore, the behaviour of a free critical point will be studied for each member of the family, to see if there is any basin of attraction different from the roots.

First, the expression of the free critical point is taken, which depends on the parameter. The dynamical behavior of operator R_α depends on the values of the parameter α so a map of the complex plane representing α parameters is defined, each point representing a different method of the family. The method corresponding to each point is iterated over the free critical point and, we paint this point of the complex plane in a specific colour if the method converges to any of the roots, 0 and ∞, and they are black in other cases. Also, the brightness of the colour indicates the number of iterations needed to reach the root. It is known that every value of α belonging to the same connected component of the parameter space gives rise to subsets of schemes of the family with similar dynamical behavior. So, it is interesting to find regions of the parameter plane as stable as possible, because these values of α will give us the best members of the family in terms of numerical stability.

In Figure 2 we have drawn two different parameter planes associated to two free critical points of the same family (studied here [7]). Both are closely connected because for the same parameter we have 2 different free critical points and it is necessary to study both cases to find the stability.

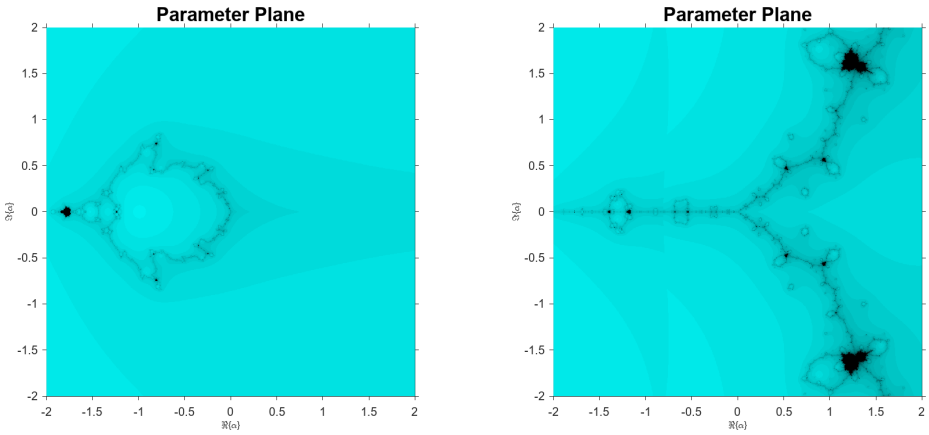


Fig. 2: Some parameter planes

3 Computational resource

The computational resource was developed in Matlab App Designer, a development environment that provides layout and code views, a fully integrated version of the Matlab editor, and a large set of interactive components that help to create user-friendly programmes. In [8] Chicharro et al. described the Matlab codes to draw the fractal images of the dynamical and parameter planes, which forms the core of this work, as well as lecture notes. To start drawing fractals we need to define R_α, i.e. define the families with which we will create the planes. Afterwards, the input parameters for drawing the maps have to be defined. And finally, create the features to interact with them. The application is already published in Matlab File Exchange, see in [10].

3.1 Families included in work

To begin with, it has been decided to take six families that have simple expressions of free critical points, such as the following, to being used in the application:

King's family of iterative methods studied here [5]:

$$\begin{aligned} y_k &= z_k - \frac{f(z_k)}{f'(z_k)},\ k = 0, 1, \ldots \\ z_{k+1} &= y_k - \frac{f(z_k) + (2+\alpha) f(y_k)}{f(z_k) + \alpha f(y_k)} \frac{f(y_k)}{f'(z_k)} \end{aligned} \tag{4}$$

It has an order of convergence 4 for every value of α. Its rational function and critical points are described by Cordero et al.

PM family of iterative methods studied here [6]:

$$\begin{aligned} y_k &= z_k - \alpha \frac{f(z_k)}{f'(z_k)},\ k = 0, 1, \ldots \\ z_{k+1} &= z_k - \frac{f(z_k)^2}{b f(z_k)^2 + c f(y_k)^2} \frac{f(z_k)}{f'(z_k)} \end{aligned} \tag{5}$$

Where $b = \frac{1-\alpha+2\alpha^2}{2\alpha^2}$ and $c = \frac{1}{2\alpha^2(\alpha-1)}$,
where $\alpha \neq 0$ and $\alpha \neq 1$. It has an order of convergence 3 for every value of α. Its rational function and critical points are described by the authors.

KLAM family of iterative methods studied here [7]:

$$\begin{aligned} y_k &= z_k - \frac{f(z_k)}{f'(z_k)},\ k = 0, 1, \ldots \\ z_{k+1} &= y_k - \frac{1+2\mu_k}{1+\alpha\mu_k^2}\frac{f(y_k)}{f'(z_k)} \end{aligned} \tag{6}$$

Where $\mu_k = \frac{f(y_k)}{f(z_k)}$. It has an order of convergence 4 for every value of α. Its rational function and critical points are described by the authors.

Kim's family of iterative methods studied here [8]:

$$\begin{aligned} y_k &= z_k - \frac{f(z_k)}{f'(z_k)},\ k = 0, 1, \ldots \\ z_{k+1} &= y_k - \frac{1+\beta u+\alpha u^2}{1+(\beta-2)u+\mu u^2}\frac{f(y_k)}{f'(z_k)} \end{aligned} \tag{7}$$

Where $u = \frac{f(y_k)}{f(z_k)}$, and it is supposed $\beta = \mu = 0$. It has an order of convergence 4 for every value of α. Its rational function and critical points are described by the authors.

The following families were proposed to be studied in the second course of numerical analysis in the mathematics degree, the below mentioned students decided to name them.

Rico's family of iterative methods studied by R. Gomez, M. Molés and J. Rico:

$$\begin{aligned} y_k &= z_k - \alpha\frac{f(z_k)}{f'(z_k)},\ k = 0, 1, \ldots \\ z_{k+1} &= z_k - \frac{f(z_k)+\gamma f(y_k)+\delta\frac{(f(y_k))^2}{f(z_k)}}{f'(z_k)} \end{aligned} \tag{8}$$

Where $\alpha \neq 0$ and $\alpha \neq 1$, $\gamma = -\frac{1}{\alpha^2}$ and $\delta = \frac{1}{\alpha^2(1-\alpha)}$. It has an order of convergence 3 for every value of α. The rational operator is

$$\frac{z^3\left(\alpha^2+2\alpha+\alpha z^3-z^3+4\alpha z^2-4z^2+5\alpha z-5z-2\right)}{\alpha+\alpha^2 z^3+2\alpha z^3-2z^3+5\alpha z^2-5z^2+4\alpha z-4x-1} \tag{9}$$

and the free critical point is

$$\frac{-\sqrt{5}\sqrt{-\alpha^2(\alpha(\alpha+12)-12)}+2(\alpha-3)\alpha+6}{3\alpha(\alpha+2)-6} \tag{10}$$

Bambi's family of iterative methods studied by A. Damià and A. Toledo:

$$\begin{aligned} y_k &= z_k - \frac{f(z_k)}{f'(z_k)},\ k = 0, 1, \ldots \\ z_{k+1} &= y_k - \frac{u_k}{1-2u_k+\alpha u_k^2}\frac{f(z_k)}{f'(z_k)} \end{aligned} \tag{11}$$

Where $u_k = -\frac{f(y_k)}{f(x_k)}$. It has an order of convergence 4 for every value of α. The rational operator is

$$\frac{z^4\left(\alpha + z^2 + 2z + 1\right)}{\alpha z^2 + z^2 + 2z + 1} \tag{12}$$

and the free critical points are

$$\begin{array}{l} \frac{1}{8}\left(-\frac{5}{\alpha+1} - \frac{A}{B} - \sqrt{\frac{16AB - 2\alpha(\alpha(\alpha(4\alpha+23)+59) - 3AB + 40)}{(\alpha+1)^3}} - 3\right) \\ \frac{1}{8}\left(-\frac{5}{\alpha+1} + \frac{A}{B} - \sqrt{2}\sqrt{-\frac{\alpha(\alpha(\alpha(4\alpha+23)+59) + 3AB + 40) + 8AB}{(\alpha+1)^3}} - 3\right) \end{array} \tag{13}$$

Where $A = \sqrt{(17 - 8\alpha)\alpha^2}$ and $B = \sqrt{(\alpha+1)^2}$

3.2 Features of the work

Figure 3, Figure 4 and Figure 5 are three screenshots to show all the components and explain how the application works.

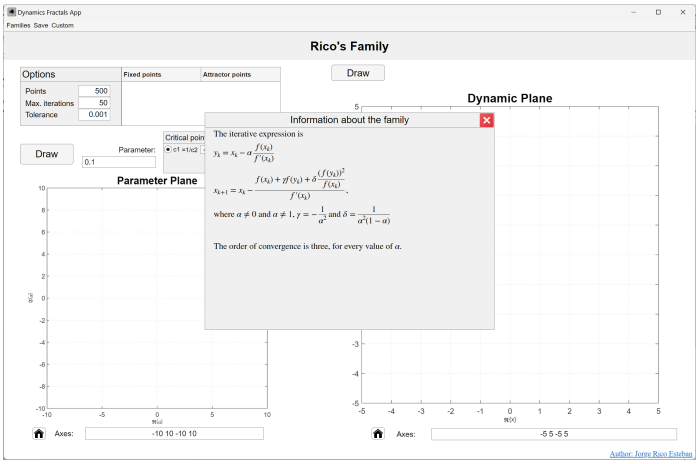


Fig. 3: Screenshot after choosing a family

The first is the pop-up window that appears after selecting a family from the top left menu. It indicates information to be known about the selected family. In the same menu where you select from one of the existing families there is an option to add 3 new families to the resource. To add them you only have to enter the name, rational operator and one free critical point to generate the parameter map. Then in the same menu you can edit it, and delete it.

The second image shows a first session with Bambi's family. In it, both maps have been drawn without modifying the default options. First, the parameter

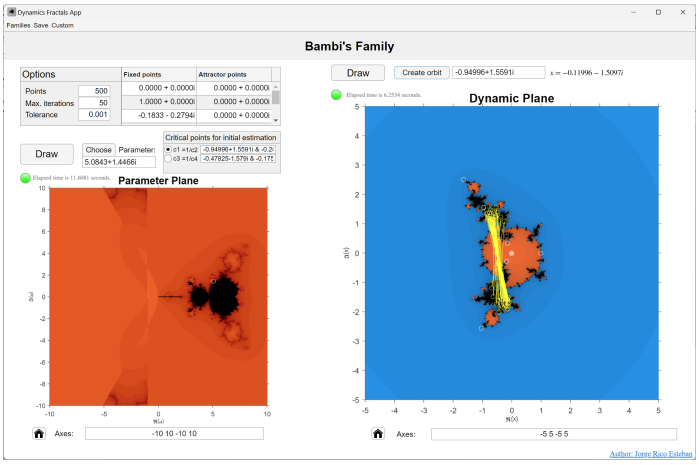


Fig. 4: Screenshot of an example of the application

map is drawn by pressing the "Draw" button and then, to select a parameter on it, the "Choose" button is clicked form the map. With the chosen parameter, the dynamic plane is drawn. You can also write the parameter to be used. In both drawing processes, the green indicator changes to yellow while the iterative process occurs and turns green when the drawing is ready. It turns black if any options have been modified and the active map does not correspond to those options. You can generate orbits in the dynamic plane in two ways, on the one hand by typing the seed of the orbit and clicking on "Create orbit", on the other hand by leaving the seed input box empty and selecting it on the map. The first case is very useful because you can copy and paste the critical point that has been used in the parameter plane and see their behaviour. As the orbit is generated in the upper right corner, the points of the orbit appear one after the other.

Figure 5 presents, the options available for generating the new plans have been modified. First, the colours used in the "Custom" menu have been changed. Additionally, the parameters used to generate the meshes of the planes have been modified, specifically the number of points and the axes. The axes can also be modified by zooming in on the generated map (one of the main objectives of the work) and returning to the default values by clicking on the house symbol. Also, the characteristic parameters of the iterative process that generates the planes have been changed, i.e. the maximum number of iterations to finish the loop, and the tolerance that is searched for when reaching a result. An important change is in the free critical point that is used to generate the parameter plane. This family has 2, so being able to switch from one to the other is very interesting and also one of the objectives of the work. In addition, the numerical value of the critical points of the selected parameter is added so you can use it as the seed of the orbit. Last but not least, the fixed points of the method and the attracting

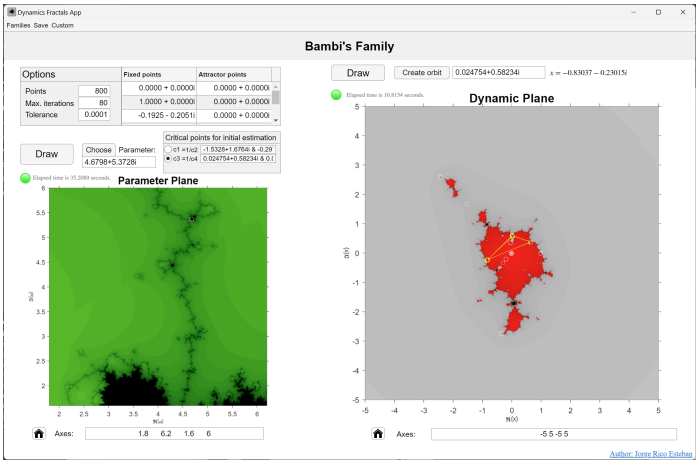


Fig. 5: Screenshot of an example of the application

are shown numerically in a table and also drawn in the dynamic plane (circles are the fixed points, asterisks are the attracting ones and squares are the critical points).

Figure 6 depicts a general flow chart of how the programme works.

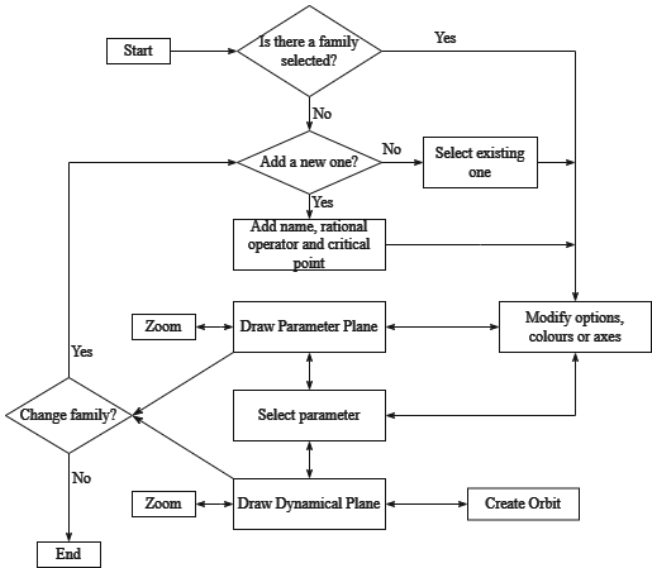


Fig. 6: Flow chart of a basic programme execution

4 Conclusions

This work aims to improve teaching and increase knowledge and interest in dynamic planes, for all the public fascinated by fractals and their beauty. A MATLAB® App Designer tool has been developed based on the lecture notes and the motivation to learn new learning tools. No specific conclusions were drawn for each family, but these could be drawn as in the following work [5] [6] [7] [8].

The interface has been designed in such a way that the user can select the desired family and even add new ones. The personalisation and manipulation of the planes has been made easily without having to modify the code. The main objective of linking the two maps by selecting parameters has also been achieved. Finally, the facility to zoom in on a plane without losing the axis reference and then redraw it has been successfully achieved.

In future developments, we will try to add more families and improve the speed of drawing the planes by modifying the code. Alternatively, a similar application will be programmed where for each method and equation to be solved, the dynamic plane will be drawn for the specific basins of attraction.

References

1. P. B. Chand, F. I. Chicharro, N. Garrido, P. Jain, Design and complex dynamics of potra–pták-type optimal methods for solving nonlinear equations and its applications, Mathematics 7 (10) (2019).
2. P. Blanchard, The dynamics of newton's method, Proc. Sympos. Appl. Math. 49 (1994) 139–154.
3. R.L. Devaney, The Mandelbrot set, the Farey tree and the Fibonacci sequence, Amer. Math. Monthly 106 (4) (1999) 289–302
4. Robert L. Devaney, Chaotic dynamical systems: An introduction. Addison-Wesley (1989).
5. A. Cordero, J. García-Maimó, J. R. Torregrosa, M. P. Vassileva, P. Vindel, Chaos in king's iterative family, Applied Mathematics Letters 26 (8) (2013) 842–848.
6. Alicia Cordero, J. R. Torregrosa, P. Vindel, Dynamical analysis to explain the numerical anomalies in the family of ermakov-kalitlin type methods, Mathematical Modelling and Analysis (2019).
7. A. Cordero, R. V. Rojas-Hiciano, J. R. Torregrosa, M. P. Vassileva, Fractal complexity of a new biparametric family of fourth optimal order based on the ermakovn-dash;kalitkin scheme, Fractal and Fractional 7 (6) (2023).
8. Francisco I. Chicharro, Alicia Cordero, J. R. Torregrosa, Drawing dynamical and parameters planes of iterative families and methods, The Scientific World Journal 2013 (2013) 11.
9. MathWorks, App Designer Matlab and Simulink Mathworks (2017) https://es.mathworks.com/help/matlab/app-designer.html
10. J. Rico, Dynamics Fractals App (2023) https://www.mathworks.com/matlabcentral/fileexchange/132398-dynamics-fractals-app, MATLAB Central File Exchange.

Introduction to solving systems of non-linear equations with iterative methods

Belén Perelló García[1], Blanca Tordera Amorós[1], and Lucía López Ribera[1]

Double Degree in Mathematics and Engineering in
Telecommunication Technologies and Services
Escuela Técnica Superior de Ingeniería de Telecomunicación,
Universitat Politècnica de València.
Camí de Vera s/n, València, Spain.
bpergar@teleco.upv.es, btoramo@teleco.upv.es, lloprib@teleco.upv.es

Abstract. A large group of mathematics is devoted to the implementation of iterative methods for solving systems of non-linear equations. This work is crucial, as these equations can model numerous physical, biological and even social phenomena that can finally be studied and even solved thanks to advances in this field.

This presentation is part of the assessment of a second year mathematics course and focuses on the analysis of a family of iterative methods. It will analyse their convergence, study their dynamical and parameter planes as well as a detailed analysis of their stability depending on certain parameters. Finally, it will be observed how it is possible to implement this family to solve specific systems of equations and thus verify the practicality of these studies for the resolution of real problems.

1 Introduction

In this project we are going to make an introduction to iterative methods for solving systems of non-linear equations, based on the analysis of their efficiency, stability and quality.
To do so, we will start with an example of one of them.

Given an initial estimate x_0, the iterative expression is:

$$y_k = x_k - \frac{2}{3}\frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots$$

$$x_{k+1} = x_k - \left(1 - \frac{3}{4}\frac{u_k \cdot (1 + \beta u_k)}{1 + u_k \cdot (\beta + \frac{3}{2})}\right)\frac{f(x_k)}{f'(x_k)},$$

We will explain the basis of the most important parameters defining an iterative method and look at these parameters in the example at hand. Thus, the aim is to give a clear and simplified view of a very important area of Mathematics such as numerical methods.

2 Convergence analysis

The first step in any analysis of an iterative method is to check whether it is extensible from simple equations to systems of equations.

The fundamental requirement for this is that there are no vectors in the denominator when making the change, since a denominator with a matrix can be substituted by the inverse of the matrix, but a denominator with a vector cannot be substituted at all.

In our case:

$$y^{(k)} = x^{(k)} - \frac{2}{3}[F'(x^{(k)})]^{-1}[F(x^{(k)})], \quad k = 0, 1, \ldots$$

$$x^{(k+1)} = x^{(k)} - \left(I - \frac{3}{4} \cdot \left[I + U^k(\beta + \frac{3}{2}) \right]^{-1} \cdot \left[U^k \cdot (I + \beta \cdot U^k) \right] \right) [F'(x^{(k)})]^{-1}[F(x^{(k)})]$$

where we denote $U^k = [F'(x^{(k)})]^{-1}[F'(y^{(k)}) - F'(x^{(k)})]$, which, being the product of two matrices, is a matrix. Bearing this in mind, the proposed transformation to systems is valid because, carrying out the products, we are left with the vector $x^{(k)}$ by subtracting the product matrix by vector (which is a vector), which corresponds to the form that will have $x^{(k+1)}$.
Then, the transformation to apply our method to systems is direct.

- Scalar convergence order with Mathematica.

For scalar convergence analysis, Taylor developments are used to express f(x) and its derivative as error expressions. Using a simple algorithm in Mathematica, obtenemos:

$$e_{k+1} = \left(c_2^3 - \frac{8\beta c_2^3}{3} - c_2 c_3 + \frac{c4}{9} \right) e_k^4 + O(e_k^5)$$

Whereby the family has order 4.

- Development of the vector convergence order. Taking the system transformation from the previous section, we will write vector Taylor developments up to fourth order, based on the scalar order deduction:

$$F(x^{(k)}) = F'(\bar{x}) \left[e_k + c_2 e_k^2 + c_3 e_k^3 + c_4 e_k^4 \right] + O(e_k^5) \tag{1}$$

$$F'(x^{(k)}) = F'(\bar{x}) \left[I + 2c_2 e_k + 3c_3 e_k^2 + 4c_4 e_k^3 \right] + O(e_k^4) \tag{2}$$

$$[F'(x^{(k)})]^{-1} = \left[I - 2c_2 e_k + (4c_2^2 - c_3)e_k^2 + (6c_2 c_3 - 8c_2^3 + 6c_3 c_2 - 4c_4)e_k^3 \right] [F'(\bar{x})]^{-1} + O(e_k^4) \tag{3}$$

Error in iteration $y^{(k)}$

We calculate the error in the first step, equation (1):

$$y^{(k)} = x^{(k)} - \frac{2}{3}[F'(x^{(k)})]^{-1}[F(x^{(k)})], \quad k = 0, 1, \ldots$$

We substitute into the above expressions, operating and grouping terms. Substracting for the solution $\bar{x}$ in both sides, we obtain:

$$y^{(k)}-\bar{x}=e_k-\frac{2}{3}z=\frac{1}{3}e_k+\frac{2}{3}c_2e_k^2-\frac{4}{3}(-c_3+c_2^2)e_k^3-\frac{2}{3}(-3c_4+4c_2c_3-4c_2^3+3c_3c_2)e_k^4 \tag{4}$$

Development of U^k

We have $U^k=[F'(x^{(k)})]^{-1}[F'(y^{(k)})-F'(x^{(k)})]$. To be able to write it, the term $F'(y^{(k)})$ is needed. Taking the expression of $F'(x^{(k)})$ and substituting the error terms e_k by $y^{(k)}-\bar{x}$ (because it is a derivative, we will develop up to third order):

$$F'(y^{(k)})=F'(\bar{x})\left[I+\frac{2}{3}c_2e_k+\left(\frac{4}{3}c_2^2+\frac{1}{3}c_3\right)e_k^2+\left(-\frac{8}{3}(-c_2c_3+c_2^3)+\frac{4}{3}c_2c_3+\frac{4}{27}c_4\right)e_k^3\right]+O(e_k^4)$$

Calculating now $F'(y^{(k)})-F'(x^{(k)})$ we obtain the development of U^k:

$$U^k=-\frac{4}{3}c_2e_k+\left(4c_2^2-\frac{8}{3}c_3\right)e_k^2+\left(\frac{28}{3}c_2c_3-\frac{32}{3}c_2^3+4c_3c_2-\frac{104}{27}c_4\right)e_k^3+O(e_k^4)$$

Expression of the error

$$x^{(k+1)}=x^{(k)}-\left(I-\frac{3}{4}\cdot\left[I+U^k(\beta+\frac{3}{2})\right]^{-1}\cdot\left[U^k\cdot(I+\beta\cdot U^k)\right]\right)[F'(x^{(k)})]^{-1}[F(x^{(k)})]$$

The term that complicates the development of the demonstration is the inverse of $[I+U^k(\beta+3/2)]$. Operating with the elements inside the parenthesis, we will propose a development for its inverse by posing a system of equations. We assume: $\left[I+U^k(\beta+\frac{3}{2})\right]^{-1}=[I+x_1e_k+x_2e_k^2+x_3e_k^3]$

And knowing that a matrix by its inverse has to be the identity; we multiply, equalise the terms of the same order and solve:

$$x_1=\frac{4}{3}\beta c_2+2c_2 \qquad x_2=-2c_2^2+\frac{4}{3}\beta^2c_2^2+\frac{8}{3}\beta c_3+4c_3$$

$$x_3=\frac{4}{3}\beta c_2c_3-\frac{16}{3}c_2^3\beta+\frac{20}{3}\beta c_3c_2+\frac{104}{27}\beta c_4-6c_2c_3+2c_3c_2+\frac{52}{9}c_4+\frac{32}{9}\beta^2c_2c_3+\frac{64}{27}\beta^3c_2^3+\frac{32}{9}\beta^2c_3c_2$$

And so, we are left with:

$$\left[I + U^k(\beta + \frac{3}{2})\right]^{-1} = [I + x_1 e_k + x_2 e_k^2 + x_3 e_k^3] = I + \left(\frac{4}{3}\beta c_2 + 2c_2\right) e_k + \left(-2c_2^2 + \frac{4}{3}\beta^2 c_2^2 + \frac{8}{3}\beta c_3 + 4c_3\right) e_k^2 \left(\frac{4}{3}\beta c_2 c_3 - \frac{16}{3}c_2^3\beta + \frac{20}{3}\beta c_3 c_2 + \frac{104}{27}\beta c_4 - 6c_2 c_3 + 2c_3 c_2 + \frac{52}{9}c_4 + \frac{32}{9}\beta^2 c_2 c_3 + \frac{64}{27}\beta^3 c_2^3 + \frac{32}{9}\beta^2 c_3 c_2\right) e_k^3$$

Returning to the iterative expression of the second step, simplifying and subtracting by the solution on both sides:

$$\boxed{e_{k+1} = \left(c_2 c_3 + c_2^3 - 2c_3 c_2 + \frac{1}{9}c_4 - \frac{8}{9}\beta c_2^3\right) e_k^4 + O(e_k^5)} \tag{5}$$

Thus, order four is demonstrated for systems.
We can conclude that the order generalises for all members of the family, since, given any β, the order continues to be four.

3 Efficiency analysis

Another crucial step in method development is the analysis of the efficiency of our method, because if we are working with a method that is not very efficient, we may not reach the objective we are looking for and it may be preferable to modify some aspect to improve efficiency.

For a good analysis, it is useful to compare the efficiency of our method with that of another existing method to see how it performs.

Family G2

We are going to calculate the number of functional evaluations per iteration (d) and the number of products/quotients per iteration (op), which are the two values we need for the calculation of the efficiency.
To calculate d, we note that our method works with two different Jacobian matrices and with the vector F(x), so d will be:

$$d = 2 \cdot n^2 + n$$

Next, to calculate op, we will have to develop the expression in such a way that we can group from right to left resolutions of systems or matrix-vector products in order to achieve a more efficient programming of the method, since if we program it without developing it, directly as it is written, the computational cost will be very high. Furthermore, we must bear in mind that when we find

ourselves with a matrix-matrix product, we must treat it as the resolution of n linear systems with the same matrix of coefficients.

Based on these expressions, we develop:

$$y^{(k)} = x^{(k)} - \frac{2}{3}[F'(x^{(k)})]^{-1}[F(x^{(k)})], \quad k = 0, 1, \dots$$

$$x^{(k+1)} = x^{(k)} - \left(I - \frac{3}{4} \cdot \left[I + U^k(\beta + \frac{3}{2}) \right]^{-1} \cdot \left[U^k \cdot (I + \beta \cdot U^k) \right] \right) [F'(x^{(k)})]^{-1}[F(x^{(k)})]$$

where we denote $U^k = [F'(x^{(k)})]^{-1}[F'(y^{(k)}) - F'(x^{(k)})]$.

We obtain:

- (n+3) linear systems with the same matrix of coefficients $\rightarrow (1/3) \cdot n^3 + (n+3) \cdot n^2 - (1/3) \cdot n$
- A linear system with a different coefficient matrix $\rightarrow (1/3) \cdot n^3 + n^2 - (1/3) \cdot n$
- Two matrix-vector products $\rightarrow 2 \cdot n^2$

In total, adding both terms d and op:

$$op + d = \frac{5}{3}n^3 + 8 \cdot n^2 + \frac{1}{3}n$$

With this sum, we will calculate the computational efficiency and Ostrowski indices, which are what will help us to estimate numerically how our method works.

Ostrowski

$$I = p^{\frac{1}{d}} = 4^{\frac{1}{2n^2+n}}$$

Computational

$$IC = p^{\frac{1}{d+op}} = 4^{\frac{1}{\frac{5}{3}n^3+8\cdot n^2+\frac{1}{3}n}}$$

In both cases, we can observe that the order tends to 1 when n tends to infinity, which indicates that the method (like all iterative methods for solving systems) loses efficiency for very large systems.

Newton's Method

Using the same process for Newton's method, we obtain the following indices:

Ostrowski

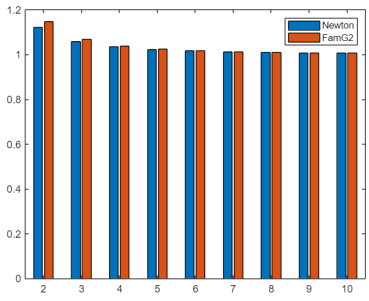
$$I_N = p^{\frac{1}{d}} = 2^{\frac{1}{n^2+n}}$$

Computational

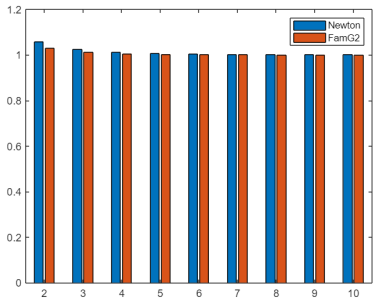
$$IC_N = p^{\frac{1}{d+op}} = 2^{\frac{1}{n^2+n+\frac{1}{3}n^3+n^2-\frac{1}{3}n}} = 2^{\frac{1}{\frac{1}{3}n^3+2n^2+\frac{2}{3}n}}$$

Comparison of methods

The most visual comparison of the two methods is done by means of graphs, in which the values of the indices for different n and both methods are plotted.
We can observe that for small values of n the Ostrowski efficiency index is higher for the G2 Family than for Newton, while the computational efficiency index works the other way around. However, for larger values of n, there is almost no difference in the value of either index, indicating that their efficiency is quite similar.



(a) Ostrowski's Index



(b) Computational Index

4 Implementation

The implementation of the Family G2 will be programmed using Matlab. First of all, we have to define the input parameters: an initial estimate of the solution, the parameter β, the tolerance, which defines the maximum desired error of the solution, and a maximum number of iterations.
The central structure of the program consist of an iterative loop in which, in this case, will be used as a stopping criterion $||x^{(k+1)} - x^{(k)}|| + ||F(x^{(k+1)})||$, or that reaches the maximum number of iterations. Inside the loop, the $n+4$ linear systems will be solved by means of the backslash, since the computation of the inverse of a matrix is very unstable (n of the systems are solved all at once using the backslash between two matrix). Moreover, with each iteration the values of $x^{(k)}$ and $x^{(k+1)}$ will be updated, as well as the increase.
Once the loop is exited we will either have the solution to the system, or the loop will have ended because the maximum number of iterations has been reached. In addition, it is advisable to calculate the convergence rates and the ACOC (the parameter that calculates the order of convergence).
With all this, the iterative loop of the program is as follows:

```
while incre + incre2>tol && iter<maxiter
    v1 = dFx\Fx;
    y = x0 - (2/3)*v1;
    [~, dFy] = feval(funcion, y);
    M1 = dFy - dFx;
    v2 = M1*v1;
    v3 = dFx\v2;
    v4 = v1 + beta*v3;
    v5 = M1*v4;
    v6 = dFx\v5;
    M2 = M1*(beta +(3/2));
    b = dFx\M2;
    B = eye(n,n) + b;
    v7 = B\v6;
    x1 = x0 - v1 + (3/4)*v7;


    A(:, iter + 1) = x1;
    [Fx, dFx] = feval(funcion, x1);
    incre = norm(x1 - x0);
    I = [I, incre];
    incre2 = norm(Fx);
    iter = iter + 1;
    x0 = x1;
    I2 = [I2 incre+incre2];
end
```

5 Stability and dynamics of the method

5.1 Rational function and Möbius transformation

The study on the stability and dynamics of the method will be performed on a rational function obtained by applying on $p(x) = (x-a)\cdot(x-b)$, an arbitrary polynomial of second degree, the expression of the family to be studied.

For the case of the Family G2, using Mathematica,we obtain the following rational function.

```
p[x_] = (x - a) * (x - b);

y = Factor[x - (2 / 3) * p[x] / D[p[x], x]];
    factoriza                   deriva
uk = (p'[y] - p'[x]) / p'[x];

Op[x_, β_] = FullSimplify[x - (1 - 3/4 * uk * (1 + β * uk) / (1 + uk (β + 3/2))) * p[x] / D[p[x], x]]
             simplifica completamente
```

$$Op(x,\beta) = x + \frac{(x-a)(x-b)\left(\frac{(a-x)(x-b)\left(4\beta(a-x)(b-x)-3(a+b-2x)^2\right)}{(a+b-2x)^2\left(3(a^2-2x(a+b)+b^2+2x^2)+4\beta(a-x)(x-b)\right)}+1\right)}{a+b-2x}$$

The function obtained depends on the parameters a and b. To achieve the independence of these parameters we apply the Möbius operator, $(h \circ Op \circ h^{-1})(x)$, and in this way, we obtain a rational function that only depends on x and β. Performing the calculations with Mathematica we find the following rational function.

```
M[u_] = (u - a) / (u - b);
iM[u_] = (u * b - a) / (u - 1);
Oper[x_, β_] =
 Simplify[M[Op[iM[x], β]]]
 simplifica
```

$$R\,(x,\,\beta) = \frac{x^4\left(8\beta - 3x^2 + (4\beta - 6)x - 3\right)}{(8\beta - 3)x^2 + (4\beta - 6)x - 3}$$

From this expression, deriving the function and evaluating it at 0, we can check the order of the method for the second degree polynomials.

$$R(0) = 0,\ R'(0) = 0,\ R''(0) = 0,\ R'''(0) = 0,\ R^{IV} = 8 \cdot (3 - 8\beta) \neq 0$$

In fact, this method has order 4, except for $\beta = \frac{3}{8}$ which also cancels R^{IV}, and therefore, for this value of β has order 5.

5.2 Fixed points

The fixed points of a function are those that return the solution $f(x) = x$. There are two types of fixed points: those related to the roots of the polynomial and the odd points.
The usefulness of calculating these points is that by analysing the behaviour of the method when working with these points, we can estimate the reliability and stability of the solutions it provides in general.

We have the rational function:

$$R(x, \beta) = \frac{x^4\left(8\beta - 3x^2 + (4\beta - 6)x - 3\right)}{(8\beta - 3)x^2 + (4\beta - 6)x - 3}$$

To calculate the fixed points, we equal our variable x, from where we get:

$$\begin{cases} pf1 = 0 \\ pf2 = 1 \\ pf3/pf6 = \pm\frac{1}{12}\left(-\sqrt{16\beta^2 + 72\beta + 9} - \sqrt{32\beta^2 - 8\sqrt{16\beta^2 + 72\beta + 9}\beta + 18\left(\sqrt{16\beta^2 + 72\beta + 9} - 3\right)} + 4\beta - 9\right) \\ pf4/pf5 = \pm\frac{1}{12}\left(-\sqrt{16\beta^2 + 72\beta + 9} + \sqrt{32\beta^2 - 8\sqrt{16\beta^2 + 72\beta + 9}\beta + 18\left(\sqrt{16\beta^2 + 72\beta + 9} - 3\right)} + 4\beta - 9\right) \end{cases}$$

In addition, we must check whether the ∞ is a fixed point by calculating the inverse operator. With Mathematica:

```
InOper[x_] = FullSimplify[1 / (-(x^4 (3 + 6 x + 3 x^2 - 8 β - 4 x β)) / (-3 - 6 x - 3 x^2 + 4 x β + 8 x^2 β)) /. x → 1 / x];
```

```
InOper[0]
```

```
0
```

The only fixed points that come from roots of the polynomial are pf1=0 and the ∞. Then pf2=1, pf3, pf4, pf5 and pf6 are odd fixed points.

Next, we will analyse the stability of 4 fixed points, two strange ones (1 and pf3) and the ones related to the roots. We must check whether the points are super attractors, attractors, parabolics or repulsors depending on the value of β we choose.

Focusing on zero and infinity, these points should be super-attractors. To check this, we derive the operator and evaluate it at both values:

```
In[•]:= Dop[x_, β_] = Factor[D[Oper[x, β], x]]
```

$$\text{Out}[\bullet]= -\frac{12\,x^3\,(1+x)^2\,\left(-3-6\,x-3\,x^2+8\,\beta+4\,x\,\beta+8\,x^2\,\beta-8\,x\,\beta^2\right)}{\left(-3-6\,x-3\,x^2+4\,x\,\beta+8\,x^2\,\beta\right)^2}$$

```
In[•]:= Dop[0, β]
Out[•]= 0

In[•]:= Limit[1 / Dop[1 / x, β], x → 0]
Out[•]= 0
```

Indeed, since both evaluations have returned zero, we can say that zero and infinity are superattractors.
Now we proceed with the strange fixed points

z=1

We evaluate the derivative of the rational function at z=1:

```
dop1[β_] = Simplify[Dop[1, β]]
```

From where: $dop1(\beta) = \frac{12-8\beta}{3-3\beta}$

We see that for $\beta = \frac{3}{2}$, z=1 is superattractor, since this value cancels the numerator.
For $\beta = 1$, z = 1 is not a fixed point, since the denominator of the rational function is cancelled.

Now let's look at the values of β, where z=1 is repulsor, attractor or parabolic: We take $\beta = a + ib$.

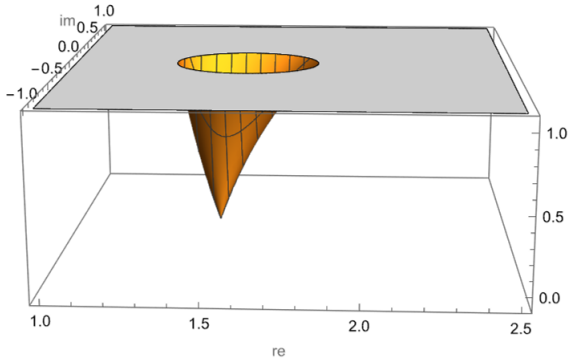
$$\left|\frac{12 - 8(a+ib)}{3 - 3(a+ib)}\right| = 1$$

and developing using the method of perfect squares we arrive at the equation of a circumference:

$$\left(a - \frac{87}{55}\right)^2 + b^2 = \frac{144}{3025}$$

Then for the values of β inside this circumference this strange fixed point is an attractor; at the boundary of the circle it is parabolic; and for all other values it is a repulsor.

This can be checked by means of a graphical representation with Mathematica, where the orange area indicates the values of β at which this strange fixed point is an attractor (and superattractor), and the grey area the values at which it is a repulsor:



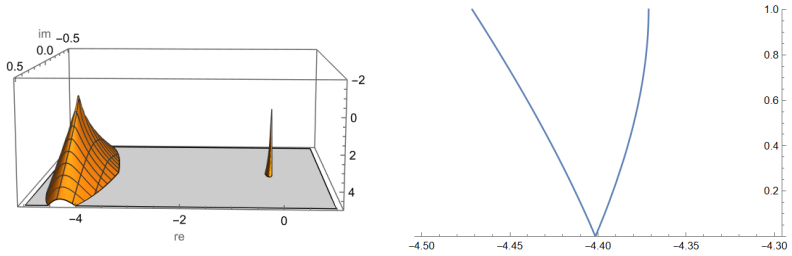
z=pf3

We evaluate the derivative of the rational function at z=pf3:

```
In[•]:= dop2[β_] = Simplify[Dop[pf3[β], β]];
                   simplifica
        NSolve[dop2[β] == 0, β]
        resuelve numéricamente
Out[•]= {{β → -4.40149}, {β → -0.126061}}
```

We can observe that for the values $\beta = -4.40149$ y $\beta = -0.126061$ pf3 will be a super attractor. In a region around both points the point will be an attractor. To calculate these regions, we solve $|dop2| = 1$.

Since trying to solve this equation with Mathematica yields inconclusive results, we estimate the value of the boundaries of the region using two graphs in 3D and 2D.



Then the fixed point pf3 will be:
- Attractor for $-0.12868 < \beta < -0.12062$ and for an approximate region of $-4.47 < beta < -4.37$.
- Parabolic for $\beta = -4.47$, $\beta = -4.37$, $\beta = -0.12868$, $\beta = -0.12062$.
- Repulsor on all other real values.

The rest of the analysis for the other points would be done in the same way, taking into account that the conjugate fixed points behave the same in the same areas in terms of attraction.

5.3 Critical points

Critical points override the derivative operator:

$$Dop = \frac{dOper}{dx} = \frac{d}{dx}\left(-\frac{x^4\left(-8\beta+3x^2-4\beta x+6x+3\right)}{8\beta x^2-3x^2+4\beta x-6x-3}\right) = 0$$

Solving with Mathematica, we obtain the following points: x=0, x=-1, and two conjugate roots (we will denote them as cr3, cr4):

$$\mathrm{cr}(\beta_{-}) = \frac{-3\beta^2 \pm \sqrt{3}\sqrt{3\beta^4+28\beta^3+84\beta^2+80\beta}-14\beta-20}{4(2\beta+5)}$$

We prove that x=-1 is a pre-image of 1 and that infinity is too a critical point:

```
Oper[-1, β]                Limit[Simplify[1 / Dop[1 / x, β]], x → 0]
                           límite  simplifica
1                          0
```

Within the critical points, we are interested in the free critical points: those that do not come from roots of the polynomial and can create basins of attraction that lead to strange fixed points. In this case, we have: -1, cr3 and cr4.

Before the analysis, we check from which we determine that, if β=0, numerator and denominator are simplified. Let us analyse cr3, cr4 which depend on the parameterβ.

cr3[β], cr4[β]

Our goal is to locate free critics to determine the behaviour of a large part of the methods in our family. We are interested in finding a value of β that causes these critics to be non-free. Matching:

```
In[•]:= Lista = {0, 1, pf3[β], pf4[β], pf5[β], pf6[β]};
        Table[Solve[Lista[[k]] == cr3[β], β], {k, 1, 6}] // N
        tabla  resuelve                                  vi
Out[•]= {{}, {{β → 1.}, {β → 1.5}}, {{β → 0.}},
         {{β → 0.}, {β → -0.126061 + 4.44089 × 10^-16 i}, {β → -4.40149 - 2.22045 × 10^-16 i}}, {{β → 2.02755 - 2.96059 × 10^-16 i}}, {}}
```

As expected, we get the same results, due to both critics being conjugated. For these values, the only free critical point will be x=-1.

By solving, we see that β=0 makes cr3, cr4 non-free critical points. Moreover, this value simplified our operator. Let us see what happens for this value: β=0:

```
Oper0[x_] = Simplify[M[Op[iM[x], 0]]]
            simplifica                                  x^4
Dop0 = D[Oper0[x], x]
       deriva                                           4 x^3
Solve[Oper0[x] == x, x]
resuelve                                                {{x → 0}, {x → 1}, {x → -(-1)^1/3}, {x → (-1)^2/3}}
Solve[4 * x^3 == 0, x] // Simplify
resuelve              simplifica                        {{x → 0}, {x → 0}, {x → 0}}
```

Then, β=0 causes the Möbius transform on our family to satisfy Cayley's test. In this case, x=-1 it is neither fixed nor critical, and there are no free critics, therefore the dynamic plane for this value of β shall consists os two basins of attraction, corresponding to 0 and ∞.

We also analyse those values that cancel the denominator of cr3, cr4:$4(2\beta + 5) = 0 \rightarrow \beta = 3/8$. If we go through the same steps again for this value, we get that the only free critical point is x=-1.

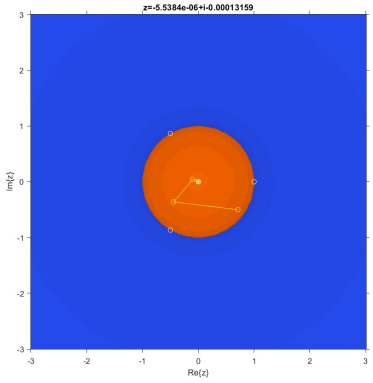
It is to be expected that, when β does not take any of the values listed above, the free critical points will be -1, cr3[β] y cr4[β].

5.4 Dynamic planes

A dynamic plane is a graphical representation that allows to observe the basins of attraction of a certain parameter β. Each point on the plane symbolizes an

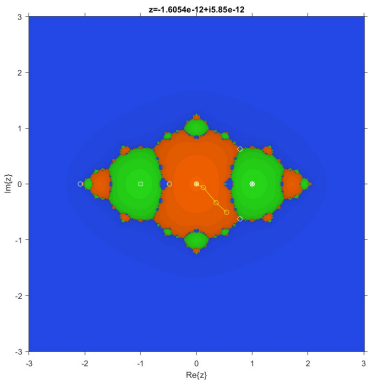
initial estimate (the x-axis represents the real part and the y-axis the imaginary one) which will be painted in a color if it converges to a critical point for that estimation, and in black if it converges to an odd fixed point or does not converge at all.
We will analyze the dynamic planes for the Family G2 using two different values of β in order to see the different behaviors.

$\underline{\beta = 0}$



For this value of β the rational function simplifies to $R(x) = x^4$ satisfying the Cayley test. This fact explains why the only two basins of attraction are the 0 (orange) and the ∞ (blue), which are the two critical superattractor points. Furthermore, we observed 3 strange fixed points (1, -0.5-0.866i, -0.5+0.866i) which are repulsors since they are found in the Julia set, and we do not have any free critical points.
As we can see, the only black region of the plan corresponds to the Julia set.

$\underline{\beta = 3/2}$



For this β value there are 7 fixed points (0, 1, -0.4805, 0.7808-0.6248i, 0.7808+0.6248i, -2.081, ∞). There are 4 critical points; (0, 1, ∞) are superattractor points and each one has its own basin of attraction. The 0 basin of attraction is orange, the 1 basin is green, and the ∞ basin is blue. On the other hand, the -1 is a free critical point but is pre-image of the 1 so it falls in the basin of attraction of the 1. This has a negative influence on the stability since z=1 is a fixed point coming from the divergence of the original method.
The other 4 strange fixed points are on Julia set and therefore they are repulsors. Once again, the Julia set is the only black area in the plane.

5.5 Parameter planes

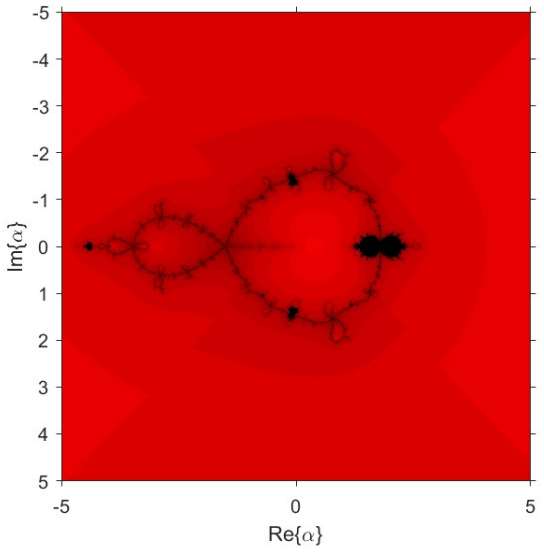
The parameter planes are a graphical representation that helps us to analyze the stability of the family for different values of the parameter β. For its representation, we must use a free critical point that depends on the parameter. Each point of the plane corresponds to a value of β, which will be colored red if the critical point we are working with converges to zero or infinity, for that value of the parameter and, on the other hand, the point will be painted black if the critical point converges to another thing, usually to an odd fixed point. In general, a method will be stable if all its critical points converge to a fixed point that is not strange, that is to say, to zero or to infinity.

This means that the existence of red regions denotes a strong stability of the method for the parameter values that fall on them. As the red gets darker, the method associated with this value of β will become more unstable, since this will indicate that more iterations have been needed to converge, until reaching black, which are regions that denote a high instability of the method for the parameters over them.

In the case of the Family G2, there are two free critical points that depend on β, but conjugate with each other. For this reason, only one plane will be represented, given that in both cases they will be the same.

$$\mathrm{cr}(\beta) = \frac{3 - 2\beta + 4\beta^2 \pm 2\sqrt{9\beta - 9\beta^2 - 4\beta^3 + 4\beta^4}}{8\beta - 3}$$

We obtain the following parameter plane.



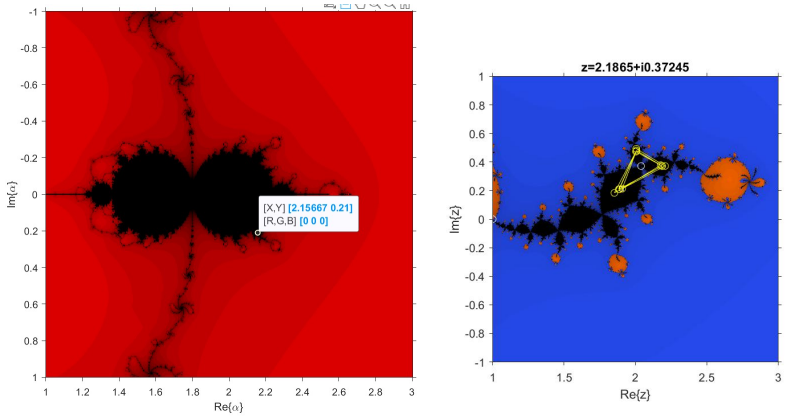
A large red zone can be observed abounding in the plane, indicating that this family is quite stable. However, there are black areas that will cause those values of the parameter not to converge to the roots.

5.6 Orbits

It is important when working with our method to check whether or not it has a large number of orbits, as orbits can negatively influence the stability of the method. When an iteration method falls into one of these orbits, it will go through each of the points that form the orbit over and over again and will not find the solution.

Sharkovsky's theorem tells us that if the family of iterative methods has an orbit of period three, then it will have orbits of any period.

To search for the period three orbit we resorted to trial and error, focusing on searching for them on the black areas of the parameter plane, in particular on the Mandelbort sets.



(a) β chosen (b) Orbit of period 3

We note that the method that arises from taking $\beta = 2.15667 + 0.2i$ gives us an orbit of period three.

Matlab shows us the last point it has reached by iterating a given maximum number of times, in this case z= 2.1865 + 0.37245i. Knowing this point, we can then check that it is indeed the orbit we are looking for, by evaluating in Mathematica the operator of our family on this point 3 times, and obtaining again the same point:

```
In[91]:= OpOrb[x] = Oper[x, β] /. β → (2.15667 + 0.2 I)

Out[91]= x^4 ((14.2534 + 1.6 i) + (2.62668 + 0.8 i) x - 3 x^2) / (-3 + (2.62668 + 0.8 i) x + (14.2534 + 1.6 i) x^2)

In[94]:= OpOrb[x] /. x → 2.1865 + 0.37245 I

Out[94]= 1.88617 + 0.208918 i

In[95]:= OpOrb[x] /. x → 1.886170897871448` + 0.20891772927176133` i

Out[95]= 2.0069 + 0.479777 i

In[96]:= OpOrb[x] /. x → 2.0068953007791612` + 0.47977669514770527` i

Out[96]= 2.18652 + 0.372442 i
```

We can therefore conclude that our method, by having an orbit of period three, will have orbits of any period.

6 Conclusions

By making a numerical study, comparing the black and red zones, we can conclude:

When taking parameters from the black area, it was observed that the results were irregular, especially when viewing the errors made in each iteration. In addition, values such as $\beta = 2$, $\beta = 1.5$ lead to unstable methods, since it was necessary to take an initial estimate close enough to the solution, and in practical cases it is not possible to know the solution in advance.

It is also interesting to note the non-convergence to a uniform solution when working with different values of the parameter. This is explained by the fact that each beta has a different dynamic plane associated with it, with different convergence basins, so that the same initial estimates converge to different solutions in each case.

References

1. Subject Resolución de sistemas de ecuaciones lineales y no lineales, taught by J.R.Torregrosa and A.Cordero, 2023
2. I.F. Traub, Iterative Methods for the Solution of Equations. Prentice-Hall, Englewood Cliffs, New York, 1964.
3. J. Milnor. Dynamics in one complex variable. Stony Brook IMS prepint (1990).