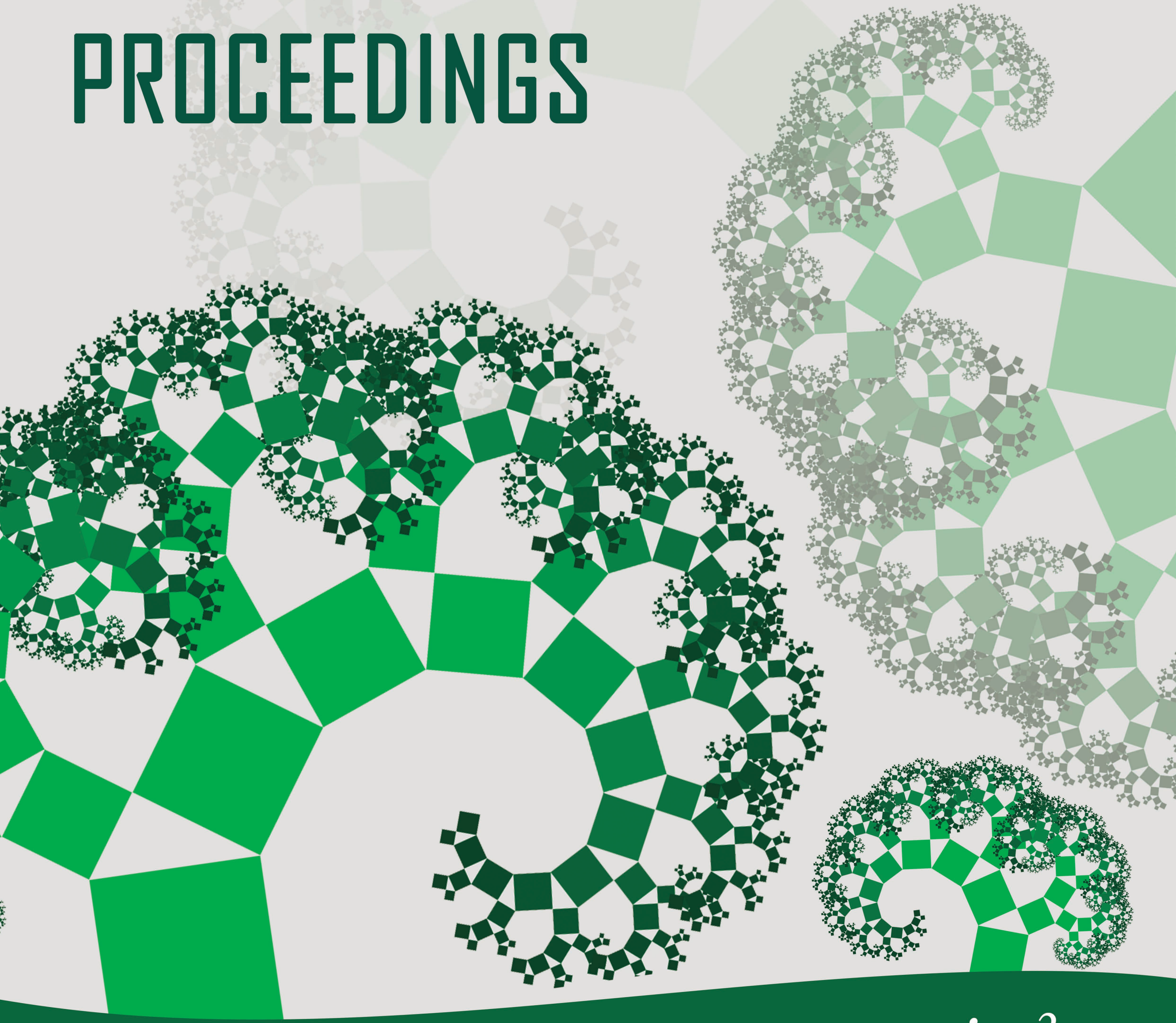
MODELLING FOR ENGINEERING & HUMAN BEHAVIOUR 2022

PROCEEDINGS



Edited by

Juan Ramón Torregrosa
Juan Carlos Cortés
Antonio Hervás
Antoni Vidal
Elena López-Navarro

im²
Instituto Universitario
de Matemática Multidisciplinar

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Modelling for Engineering & Human Behaviour 2022

València, July 14th-16th, 2022

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

I.S.B.N.: 978-84-09-47037-2

November 30th, 2022
Report any problems with this document to imm@imm.upv.es.

Edited by: I.U. de Matemàtica Multidisciplinar, Universitat Politècnica de València.
J.R. Torregrosa, J-C. Cortés, A. Hervás, A. Vidal-Ferràndiz and E. López-Navarro

im²

Instituto Universitario
de Matemática Multidisciplinar

Contents

Developable surface patches bounded by NURBS curves
L. Fernández-Jambrina 1

Impact of antibiotic consumption on the dynamic evolution of antibiotic resistance: the colistin-resistant *Acinetobacter baumannii* case
Carlos Andreu-Vilarroig, Juan-Carlos Cortés, Rafael-Jacinto Villanueva 7

Relative research contributions towards the application and materialization of Wenner four-point method on concrete curing
L. Andrés, J. H. Alcañiz, T. P. Real, P. Suárez 12

Application of mathematical models and multicriteria analysis to establish an optimized priorization for the maintenance of bridges in a network
L. Andrés, F. Ribes, E. Fernández, J. Maldonado 18

Probabilistic analysis of scalar random differential equations with state-dependent impulsive terms via probability density functions
V. Bevia, J.C. Cortés, M. Jornet and R.J. Villanueva 24

Combined and reduced combined matrices
R. Bru, M.T. Gassó and M. Santana 30

Random Fractional Hermite Differential Equation: A full study un mean square sense
C. Burgos, T. Caraballo, J. C. Cortés and R. J. Villanueva 36

Mathematical modelling of frailty and dependency in basic activities of daily living in general population aged 70 years
S. Camacho Torregrosa, C. Santamaría Navarro and X. Albert Ros 41

Jordan structures of an upper block echelon matrix
B. Cantó, R. Cantó and A.M. Urbano 47

Modeling inflammation in wound healing
A. Patrick and B. Chen-Charpentier 53

The geometry behind PageRank rankings
Gonzalo Contreras-Aso, Regino Criado and Miguel Romance 59

Surveillance model of the evolution of the plant mass affected by Xylella fastidiosa in Alicante (Spain)
José Juan Cortés Plana, María Teresa Signes Pont, Joan Boters Pitarch and Higinio Mora Mora 64

New tools for linguistic pattern analysis and specialized text translation: hypergraphs and its derivatives
A. Criado-Alonso, D. Aleja, M. Romance and R. Criado 70

On an accurate method to compute the matrix logarithm
E. Defez, J.J. Ibáñez, J. M. Alonso and J.R. Herráiz 76

A distribution rule for allocation problems with priority agents using least-squares method
J.C. Macías Ponce, A.E. Giles Flores, S.E. Delgadillo Alemán, R.A. Kú Carrillo and L.J.R. Esparza 85

A reaction-diffusion equation to model the population of Candida Auris in an Intensive Care Unit
C. Pérez-Diukina, J.-C. Cortés López and R.J. Villanueva Micó 91

Relative research contribution towards railways superstructure quality determination from the vehicles inertial response
E. Gómez, J. H. Alcañiz, G. Alandí and F. E. Arriaga 97

Computational Tools in Cosmology
Màrius Josep Fullana i Alfonso and Josep Vicent Arnau i Córdoba 102

Dynamical analysis of a family of Traub-type iterative methods for solving nonlinear problems
F.I. Chicharro, A. Cordero, N. Garrido and J.R. Torregrosa 107

Multidimensional extension of conformable fractional iterative methods for solving nonlinear problems
Giro Candelario, Alicia Cordero, Juan R. Torregrosa and María P. Vassileva 113

Application of Data Envelopment Analysis to the evaluation of biotechnological companies
B. Latorre-Scilingo, S. González-de-Julián and I. Barrachina-Martínez 119

An algorithm for solving Feedback Nash stochastic differential games with an application to the Psychology of love
Jorge Herrera de la Cruz and José-Manuel Rey 125

Detection of border communities using convolution techniques
José Miguel Montañana, Antonio Hervás, Samuel Morillas and Alejandro Méndez 133

Optimizing rehabilitation alternatives for large intermittent water distribution systems
Bruno Brentan, Silvia Carpitella, Ariele Zanfei, Rui Gabriel Souza, Andrea Menapace, Gustavo Meirelles and Joaquín Izquierdo 139

Performance analysis of the constructive optimization of railway stiffness transition zones by means of vibration studies
M. Labrado, J. del Pozo, R. Cabezas and A. Arias 143

Can any side effects be detected as a result of the COVID-19 pandemic? A study based on social media posts from a Spanish Northwestern-region
A. Larrañaga, G. Vilar, J. Martínez and I. Ocarranza 149

Probabilistic analysis of a cantilever beam with load modelled via Brownian motion
J.-C. Cortés, E. López-Navarro, J.I. Real Herráiz, J.-V. Romero and M.-D. Roselló . . . 155

Pivoting in ISM factorizations
J. Mas and J. Marín 160

Relative research contributions towards the characterization of scour in bridge piers based on operational modal analysis techniques
S. Mateo, J. H. Alcañiz, J. I. Real, E. A. Colomer 166

Short-term happiness dynamics as a consequence of an alcohol or caffeine intake
Salvador Amigó, Antonio Caselles, Joan C. Micó and David Soler 172

First Order Hamiltonian Systems
Joan C. Micó 177

Dynamical analysis of a new sixth-order parametric family for solving nonlinear systems of equations
Marlon Moscoso-Martínez, Alicia Cordero, Juan R. Torregrosa and F. I. Chicharro 185

Higher order numerical methods for addressing an embedded steel constitutive model
J.J. Padilla, A. Cordero♭, A.M. Hernández-Díaz and J.R. Torregrosa 191

A generalization of subdirect sums of matrices
F. Pedroche 198

A new iterative inverse display model
M.J. Pérez-Peñalver, S.-W.Lee, C. Jordán, E. Sanabria-Codesal and S. Morillas 202

Application of polynomial algebras to non-linear equation solvers
J. Canela and D. Pérez-Palau 208

A Linear Quadratic Tracking Problem for Impulsive Controlled Stochastic Systems. The Infinite Horizon Time Case
V. Drğan, I-L. Popa and I.Ivanov 213

A new model for the spread of cyber-epidemics
E. Primo, D. Aleja, G. Contreras-Aso, K. Alfaro-Bittner, M. Romance and R. Criado 218

Bifurcation analysis in dryland vegetation models with discrete and distributed delays
I. Medjahdi, F.Z. Lachachi, M.A. Castro and F. Rodríguez 224

Relative research contribution towards railways superstructure quality determination from the vehicles inertial response
J. R. Sánchez, F. J. Vea, G. Muinelo and G. Mateo 230

Probabilistic analysis of the pseudo-n order adsorption kinetic model
C. Andreu-Vilarroig, J.-C. Cortés, A. Navarro-Quiles and S.-M. Sferle 236

Approximating Fixed Points by New and Fast Iterative Schemes
Puneet Sharma, Vinay Kanwar, Ramandeep Behl and Mithil Rajput 242

Higher-order multiplicative derivative iterative scheme to solve the nonlinear problems
G. Singh, S. Bhalla and R. Behl 247

A Seventh Order Steffensen type Iterative Method for Solving Systems of Nonlinear Equations and Applications
Sana Sultana and Fiza Zafar 251

Modifying Kurchatov's method to find multiple roots
A. Cordero, N. Garrido, J.R. Torregrosa and P. Triguero-Navarro 257

The Relativistic Anharmonic Oscillator within a Double-Well Potential within a Double-Well Potential
Michael M. Tung and Frederic Rapp 263

Real-valued preconditioners for complex linear systems arising from the nuclear reactor noise equation
A. Vidal-Ferràndiz, A. Carreño, D. Ginestar and G. Verdú 268

Mathematical model for heat transfer and stabilization of LED lamps for measurements in a laboratory
Carlos Velásquez, M. Ángeles Castro, Francisco Rodríguez and Francisco Espín 274

A Seventh Order Family of Jarratt Type Iterative Methods for Solving Systems of Nonlinear Equations and Applications
Saima Yaseen and Fiza Zafar 279

A Seventh Order Jarratt type Iterative Method for Solving Systems of Nonlinear Equations and Applications
Fiza Zafar, Alicia Cordero, Husna Maryam and Juan R. Torregrosa 285

A Multiplicative calculus approach to solve nonlinear equations
S. Bhalla and R. Behl 291

Basins of Convergence in a Modified CR3BP
D. Villalibre, A. Herrero, J.A. Moraño and S. Moll 294

SPECIAL SESSION: STUDENT'S PROJECTS 301

Predicting PDA battery health using machine learning methods
Alberto Bono Monreal, Irene Cánovas Vidal, Eduardo Gómez Fernández, Rubén Marco Cabanes, Andrea Pérez López, Alejandra Sánchez Torres and Marta Valero Buj 303

Solving a Crime with Graph Theory
Elsa Blasco Novell and Lucía López Ribera 313

Harry Potter and the relics of the graphs
D. Vañó Fernández and R. Fornas Sáez 318

Evaluating the sudden change in flight cancellations in Paris during November 2015
D. Romero, C. Gallego, R. Gironés, J. Grau, L. Lillo and A. Losa 325

1

Developable surface patches bounded by NURBS curves

L. Fernández-Jambrina[♭][1]

(♭) ETSI Navales,
Universidad Politécnica de Madrid
Avenida de la Memoria 4, 28040-Madrid, Spain.

1 Introduction

In this talk we review the problem of constructing a developable surface patch bounded by two rational or NURBS (Non-Uniform Rational B-spline) curves.

NURBS curves are curves which are piecewise rational. That is, they are a generalisation of spline curves, which are piecewise polynomial curves. Similarly we define NURBS surfaces and solids. NURBS curves have been the standard in Computer Aided Design [1] for a long time through IGES and STEP specifications. They are described by a list of points called control polygon and two list of numbers: the list of weights and the list of knots.

Developable surfaces are ruled surfaces with null Gaussian curvature [2]. This implies that they can be constructed from planar surfaces by just cutting, rolling and folding, so that metric properties such as lengths and angles between curves and areas are preserved. These geometric properties are of great interest for steel and textile industry, since these are pieces designed in the plane and then combed into space.

For instance, in naval architecture sheets of steel are adapted to fit into the hull of a ship [3–5]. If these sheets are combed just with a folding machine, the costs are lower than if they require the use of heat. In textile industry cloth is planar and is cut and sewn to produce garments and the quality is improved if it is not stretched [6]. They have also been used for designing facades in architecture [7] and in automobile industry [8].

In geometric design the standard relies on the use of rational B-spline curves and surfaces (NURBS), which are described by control polygons or nets and lists of weights and knots. In the case of ruled surfaces, the control net is formed by just the control polygons of the bounding curves, since segments can be described by just their endpoints, which form their control polygon.

This problem has been addressed in several ways [9], but the key drawback is that when we require the developable surface to be NURBS and bounded by NURBS curves, the possibilities are restricted [10, 11].

In some cases, the bounding curves are planar and lie on parallel planes [12,13] and this simplifies the problem.

For instance, one can obtain general solutions for developable NURBS surfaces bounded by NURBS curves [15–19].

[1] leonardo.fernandez@upm.es

Considering the dual space in projective geometry has been also profitable, since developable surfaces may be seen as envelopes of lines of planes [20, 21].

For this reason our proposal is to consider developable surface patches which are not NURBS, though bounded by NURBS curves [22]. In fact, we are able to obtain every possible solution, good or bad, to this problem in our framework.

Our approach is based on performing a reparametrisation on one of the bounding curves of the surface patch and imposing that the resulting surface be developable [22]. One would expect a differential equation for the reparametrisation function. However, the condition happens to be algebraic. Another approach [23] reduces the null Gaussian curvature condition to quadratic equations.

Moreover, if the bounding curves are (piecewise) polynomial or rational of degree n, the developability condition is an algebraic equation of degree $2n-2$. The degree may be lowered to $n-1$ if the curves are (piecewise) polynomial and lie on parallel planes.

2 Methods

We start with a ruled surface parametrised by $b(t,v)$ and bounded by two parametrised curves, $c(t)$, $d(t)$,

$$b(t,v) = (1-v)c(t) + vd(t), \quad t, v \in [0,1].$$

For given $c(t)$ and $d(t)$, this ruled surface will not be developable in general, but, if it is developable, this feature shall not depend on the chosen parametrisation.

The Gaussian curvature $K(t,v)$ is the quotient of the determinants of the second, $B(t,v)$, and first, $G(t,v)$, fundamental forms of the surface patch parametrised by $b(t,v)$,

$$K(t,v) = \frac{\det B(t,v)}{\det G(t,v)}.$$

The first fundamental form $G(t,v)$ is the usual scalar product restricted to tangent vectors to the surface at the point $b(t,v)$ and is used, for instance, for calculating areas,

$$\text{Area}(S) = \int_D \sqrt{\det G(t,v)}\, dtdv,$$

but, since its determinant is positive and appears at the denominator of $K(u,v)$, we need not pay attention to it, though due to Gauss' Theorema Egregium [2], the Gaussian curvature is an intrinsic property and may be written in terms of just the first fundamental form.

On the other hand, the second fundamental form $B(t,v)$ is the extrinsic curvature of the surface at the point $b(t,v)$ with ν as unitary normal to the surface and can be constructed with the projections of the second derivatives of the parametrisation along the unitary normal $\nu(t,v)$ to the surface at $b(t,v)$,

$$B(t,v) = \begin{pmatrix} b_{tt}\cdot\nu & b_{tv}\cdot\nu \\ b_{vt}\cdot\nu & b_{vv}\cdot\nu \end{pmatrix}_{(t,v)}.$$

In this sense, the second fundamental form is used to compute the normal curvature at points on the surface.

For a ruled surface parametrised as in (1),

$$\det B(t,v) = \left| \begin{matrix} ((1-v)c''(t) + vd''(t))\cdot\nu & (d'(t)-c'(t))\cdot\nu \\ (d'(t)-c'(t))\cdot\nu & 0 \end{matrix} \right|_{(t,v)} = -\left((d'(t)-c'(t))\cdot\nu\right)^2,$$

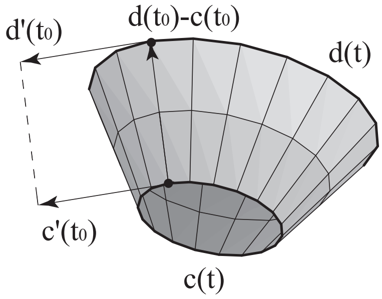


Figure 1: Developability is equivalent to having the same tangent plane along each ruling of a ruled surface

we see that the Gaussian curvature is always non-positive and hence there are no elliptic points on a ruled surface, since $\det B(t,v)$ is always positive.

Moreover, vanishing Gaussian curvature is equivalent to vanishing $(d'(t)-c'(t))\cdot\nu$ on the points of the ruled surface.

Since at a point $b(t,v)$ on the surface we have $b_t(u,v)=(1-v)c'(t)+d'(t)$, $b_v(t,v)=d(t)-c(t)$ as two tangent vectors to the surface, we may construct a normal vector to the surface at $b(t,v)$ as $N=b_t\times b_v$, and then having a vanishing Gaussian curvature is equivalent to having a vanishing triple product,

$$c'(t)\cdot d'(t)\times(d(t)-c(t))=\det\left(c'(t),d'(t),d(t)-c(t)\right)=0,\qquad t\in[0,1] \tag{1}$$

at all points of the surface patch.

As we see in Fig 1, this is equivalent to requiring that the three vectors $c'(t)$, $d'(t)$ and $d(t)-c(t)$ lie on the same plane for all values of t. This also implies that the normal to the surface is the same for all points on the segment (ruling) linking $c(t)$ and $d(t)$.

Our contribution to deal with this problem is based on reparametrisation of one of the bounding curves by a function $T(t)$,

$$\tilde{b}(t,v)=(1-v)c(t)+vd(T(t)) \tag{2}$$

and require $\tilde{b}(t,v)$ to satisfy the null Gaussian curvature condition.

3 Results

The developability condition (1) applied to parametrisations such as (2) can be seen to be algebraic in $T(t)$, since the dependence on the derivative $T'(t)$ is factored out by the determinant,

$$\det\left(c'(t),\dot{d}(T),d(T)-c(t)\right)=0, \tag{3}$$

where the dot stands for derivation with respect to T.

In the case of (piecewise) polynomial or rational curves $c(t)$, $d(t)$, further consequences may be derived:

Theorem 1: Let $c(t)$, $d(T)$, $t,T\in[0,1]$ be rational curves of degree n. The parameterized ruled surface,

$$b(t,v)=(1-v)c(t)+vd(T(t)),\qquad t,v\in[0,1],$$

is a developable surface if the reparameterization function $T(t)$ satisfies the algebraic equation

$$\det\left(c'(t),\dot{d}(T),d(T)-c(t)\right)_{T=T(t)}=0,$$

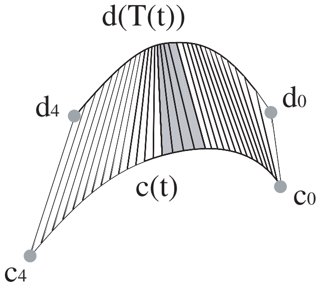


Figure 2: Developable patch bounded by two cubic spline curves

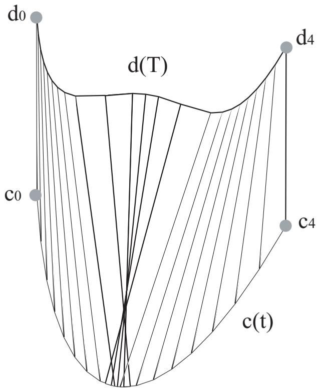


Figure 3: Developable surface with a regression area

and is a real monotonically increasing function of t.

This equation is of degree $2n-2$ at most. If both curves are (piecewise) polynomial and lie on parallel planes, the equation is of degree $n-1$ at most. We may see an example in Fig 2.

The price to pay is that solutions of this algebraic equation will not be rational or polynomial in general and $\tilde{b}(t,v)$ will no longer be NURBS.

Since the condition on the reparametrisation is algebraic, the number of possible solutions is finite, but not all of them are geometrically acceptable.

For being a reparametrisation, $T(t)$ must be a monotonically increasing function. Otherwise, we would have unpleasant regression areas with more than on ruling through some points of the curves (See Fig 3). This can be checked with the help of

$$T'(t) = \left.\frac{\det\left(c''(t), \dot{d}(T), d(T)-c(t)\right)}{\det\left(\ddot{d}(T), c'(t), d(T)-c(t)\right)}\right|_{T=T(t)},$$

which we derive from the null Gaussian curvature condition.

This implies that monotonicity is granted if

$$\operatorname{sgn}\left(c''(t)\cdot\nu(t)\right) = \operatorname{sgn}\left.\left(\ddot{d}(T)\cdot\nu(t)\right)\right|_{T=T(t)},$$

where $\nu(t)$ is the unitary normal to the ruled surface along the segment at t. This means that the normal curvatures of both curves must have the same sign for each value of t.

Hence, acceptable solutions just appear if both curves are qualitatively similar regarding their curvature.

Theorem 2: Let $c(t)$, $d(T)$, $t, T\in[0,1]$ be parameterized curves. Let $T(t)$ be a reparameterization function so that

$$b(t,v) = (1-v)c(t) + v d(T(t)), \qquad t, v \in [0,1],$$

is a developable surface. $T(t)$ is a monotonically increasing function if and only if for all t,

$$\operatorname{sgn}\left(c''(t)\cdot\nu(t)\right)=\operatorname{sgn}\left(\ddot{d}(T)\cdot\nu(t)\right)\Big|_{T=T(t)},$$

where $\nu(t)$ is the unitary normal to the surface along the ruling at t.

Or equivalently, for the normal curvatures $k_{n,c}$, $k_{n,d}$ of both curves

$$\operatorname{sgn}\left(k_{n,c}(t)\right)=\operatorname{sgn}\left(k_{n,d}(T(t)\right),$$

for all values of t.

In the case of parameterizations of class C^k of differentiability, $T(t)$ is of class C^{k-1}.

4 Conclusions

We have produced a new approach for dealing with the problem of constructing a developable surface patch between two parametrised curves $c(t)$ and $d(t)$.

This approach is grounded on performing a reparametrisation of one of the curves and the developability condition is not a differential equation, but an algebraic equation, and provides all possible solutions to the problem.

In the case of (piecewise) polynomial or rational curves of degree n, the developability condition is an algebraic equation of degree $2n-2$. Since the most usual degree in Computer Aided Design is three, this means we are dealing with a fourth degree equation, which can be handled either by numerical or analytical methods.

For (piecewise) polynomial curves of degree n, lying on parallel planes, the algebraic equation is of degree $n-1$.

Requiring that the reparametrisation function be a monotonically increasing function, in order to avoid regression areas on the developable surface patch, is achieved if the sign of the normal curvatures of both bounding curves is the same at the endpoints of each ruling.

With this approach, it is easy to control the final class of differentiability of the surface in the piecewise case.

References

[1] Farin, G, Curves and surfaces for CAGD: a practical guide, 5th Edition. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 2002.

[2] Struik, D.J., Lectures on classical differential geometry, 2nd Edition, New York, Dover Publications Inc., 1988.

[3] Kilgore, U., Developable hull surfaces, in: Fishing Boats of the World, Vol. 3, Fishing News Books Ltd., Surrey, 425–431, 1967.

[4] Chalfant, J.S., Maekawa, T. Design for manufacturing using B-spline developable surfaces., *J. Ship Research* 42: 207–215, 1998.

[5] Pérez, F., Suárez, J, Quasi-developable B-spline surfaces in ship hull design, *Computer-Aided Design* 39: 853–862, 2007.

[6] Rose, K., Sheffer, A., Wither, J., Cani, M.P., Thibert, B., Developable surfaces from arbitrary sketched boundaries, in: A. Belyaev, M. Garland (Eds.), Eurographics Symposium on Geometry Processing (2007), The Eurographics Association, 163–172, 2007.

[7] Pottmann, H., Asperl, A., Hofer, M., Kilian, A., Architectural geometry, Exton, Bentley Institute Press, 2007.

[8] Frey, W.H., Bindschadler, D., Computer aided design of a class of developable Bézier surfaces, Tech. rep., GM Research Publication R&D-8057, 1993.

[9] Pottmann, H., Wallner, J., Computational line geometry, Mathematics and Visualization, Berlin, Springer-Verlag, Berlin, 2001.

[10] Fernández-Jambrina, L, Bézier developable surfaces, *Comput. Aided Geom. Design* 55: 15–28, 2017.

[11] Fernández-Jambrina, L, Characterisation of rational and NURBS developable surfaces in Computer Aided Design, *Journal of Computational Mathematics* 39: 550–568, 2021.

[12] Aumann, G., Interpolation with developable Bézier patches, *Comput. Aided Geom. Design* 8: 409–420, 1991.

[13] Maekawa, T., Design and tessellation of B-spline developable surfaces, *ASME Transactions Journal of Mechanical Design* 120: 453–461, 1998.

[14] Chu, C.H., Séquin, C.H., Developable Bézier patches: properties and design, *Computer Aided Design* 34: 511–527 (2002).

[15] Aumann, G., A simple algorithm for designing developable Bézier surfaces, *Comput. Aided Geom. Design* 20: 601–619, 2003

[16] Aumann, G., Degree elevation and developable Bézier surfaces, *Comput. Aided Geom. Design* 21: 661–670, 2004.

[17] Fernández-Jambrina, L, B-spline control nets for developable surfaces, *Comput. Aided Geom. Design* 24: 189–199, 2007.

[18] Cantón, A., Fernández-Jambrina, L., Interpolation of a spline developable surface between a curve and two rulings, F*rontiers of Information Technology & Electronic Engineering* 16:173–190 (2015).

[19] Cantón, A., Fernández-Jambrina, L, Non-degenerate developable triangular Bézier patches, in: J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, L. Schumaker (Eds.), Curves and Surfaces, Vol. 6920 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 207–219, 2012.

[20] Bodduluri, R.M.C, Ravani, B., Design of developable surfaces using duality between plane and point geometries, *Computer Aided Design* 25: 621–632, 1993.

[21] H. Pottmann, G. Farin, Developable rational Bézier and B-spline surfaces, Comput. Aided Geom. Design 12 (5) (1995) 513–531.

[22] Fernández-Jambrina, L, Pérez-Arribas, F, Developable surfaces bounded by NURBS curves, *Journal of Computational Mathematics* 38: 693–709, 2020.

[23] Tang, C., Bo, Wallner, J., Pottmann, H., Interactive design of developable surfaces, *ACM Trans. Graph.* 35: 1–12, 2016.

2

Impact of antibiotic consumption on the dynamic evolution of antibiotic resistance: the colistin-resistant *Acinetobacter baumannii* case

Carlos Andreu-Vilarroig ♭,[1] Juan-Carlos Cortés ♭
Rafael-Jacinto Villanueva ♭

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

One of the major public health threats today is antibiotic resistance, i.e. the ability of certain bacteria to genetically adapt and resist antibiotic treatment [12, 14]. The evolution of antibiotic resistance has been particularly worrying in recent years, and the latest forecasts estimate that by the year 2050, resistance will cause around 10 million deaths annually and an annual cost of 100 billion USD [7]. In the face of this challenge, mathematical modeling can provide great value in predicting, monitoring and aiding public health policy decisions.

On the evolution of the dynamics of antibiotic resistance, is well-known that antibiotic consumption increases bacterial resistance [3]. This evidence is based on the fact that, in an environment with the presence of antibiotic, resistance is a competitive advantage for the bacteria and, consequently, resistant bacteria survive in greater proportion than sensitive bacteria [9]. However, in the absence of antibiotic, there is no clear consensus on whether maintaining resistance is a disadvantage for the resistant bacteria and, then, whether antibiotic resistance can be reversed [2, 10].

In this work, a general random mathematical model that describes the evolution over time of the proportion of a resistant microorganism against an antibiotic have been proposed. The equation depends on the antibiotic consumption. Using real available data on antibiotic consumption and the proportion of resistance, the random model parameters have been calibrated using the Multi-Objective Particle Swarm Optimization (MOPSO) bio-inspired evolutionary algorithm [5] and the Monte Carlo method. In this way, the probability distributions of the parameters that best capture the uncertainty of the time series of data have been obtained.

[1] caranvi1@upv.es

2 Methods

2.1 Model description

The proposed random mathematical model describes the evolution of bacterial resistance to an antibiotic in a population over time. This equation has been derived from a compartmental microbiological model with three species (subpopulations): resistant bacteria, sensitive bacteria and free space left by dying bacteria [11]. The random equation for the proportion of resistant bacteria in a population is given by:

$$p(t) = \frac{e^{\tau A(t) - \Delta\kappa t + \theta_0}}{e^{\tau A(t) - \Delta\kappa t + \theta_0} + 1} \in (0, 1), \tag{1}$$

where $\tau \in \mathbb{R}^+$ is the antibiotic kill rate, $\Delta\kappa \in \mathbb{R}^+$ is the difference between resistant and sensitive bacterial mortality rates, and $\theta_0 \in \mathbb{R}$ is the ratio $\theta(t)$, which is defined as $\theta(t) = \log\left(\frac{p(t)}{1-p(t)}\right)$, at time $t = 0$. The three parameters are considered in this model as random variables. The cumulative antibiotic consumption function, defined as $A(t) = \int_0^t a(u)du \in \mathbb{R}^+$ (where $a(t)$ is the antibiotic consumption registered at time t), is considered as a known deterministic function.

2.2 Data

The random model has been applied to study the resistance evolution of the *Acinetobacter baumannii* (or *A. baumannii*) bacterium, which is one of the most problematic microorganisms in hospitals. As antibiotic, we have chosen colistin because is one of the last efective antibiotics against *A. baumannii* [4, 8]. For our application, the data series of interest have been *(i)* the proportion $\hat{\mathbf{p}} = \{\hat{p}_t : \ t = 0, 1, \dots, 107\}$ of *A. baumannii* colistin-resistant monthly cases in the 2012-2020 period [13], and *(ii)* the yearly colistin consumption $\hat{\mathbf{a}} = \{\hat{a}_t : \ t = 47, \dots, 107\}$ [1] in the 2016-2020 period. As consumption data between 2012 and 2015 are not registered in antibiotic consumption database, and a linear growth have been observed, the 2012-2015 period monthly consumption data have been interpolated by a linear regression model, obtaining an estimation function $\hat{a}(t) = \beta_1 t + \beta_0 = 0.000525t + 0.031625$ for the 2012-2020 entire period.

2.3 Model calibration

The first step in the calibration process is to specify the family distribution of the model parameters. From the data we do not have any information about the effect of decreased resistance to a reduction in consumption. Consequently, we assume the worst possible scenario: resistance is not a competitive disadvantage in the absence of antibiotic, or, $\Delta\kappa = 0$. On the other hand, the remaining parameters $\tau \in \mathbb{R}^+$ and $\theta_0 \in \mathbb{R}$ can be considered as random variables with log-normal and Gaussian distributions respectively, so that

$$\tau \sim \text{log-}\mathcal{N}(\mu_\tau, \sigma_\tau), \quad \theta_0 \sim \mathcal{N}(\mu_{\theta_0}, \sigma_{\theta_0}),$$

Once defined the random model parameters, the calibration goal is to find the parameter set $\pi = (\mu_\tau, \mu_{\theta_0}, \sigma_\tau, \sigma_{\theta_0})$ that best captures the uncertainty of the data within a specific $(1-\alpha)$-confidence region. A significance level of $\alpha = 0.05$ has been chosen. To calibrate the model, the Multi-Objective Particle Swarm Optimization (MOPSO) bioinspired evolutionary algorithm have been applied [5]. The algorithm follows these steps:

1. Generate L particles. Each particle represents a set of parameters $\pi = (\mu_\tau, \mu_{\theta_0}, \sigma_\tau, \sigma_{\theta_0})$, which characterizes $\tau \sim \text{log-}\mathcal{N}(\mu_\tau, \sigma_\tau)$ and $\theta_0 \sim \mathcal{N}(\mu_{\theta_0}, \sigma_{\theta_0})$ random model parameters.

2. Generate a set $P = \{p^i(t;\pi) : \; t = 0, \ldots, 107; \; i = 1, \ldots, n\}$ of n model simulations using the random model in Equation (1) with n sampled pairs $\{\tau, \theta_0\}_{i=1}^{n}$ from the model parameters distributions (Monte Carlo sampling), and considering $\Delta\kappa = 0$.

3. Compute the objetive functions with set P and real data $\hat{\mathbf{p}}$, and the time instants set $T = \{0, \ldots, 107\}$. Two orthogonal or antagonistic objective functions have been applied:

   - *Inside-outside error function*: if we define $q_t = q(t;P)$ and $Q_t = Q(t;P)$ as the $\alpha/2$ and $1 - \alpha/2$ quantiles, respectively, of the set $\{p^i(t;\pi)\}_{i=1}^{n}$ given a time instant t, the inside-outside error function F_{io} is defined as

     $$F_{io}\left(\hat{\mathbf{p}}, P\right) = \sum_{t \in T} \min\left\{|\hat{p}_t - q_t|, |\hat{p}_t - Q_t|\right\} \mathbb{1}_{\hat{p}_t \notin [q_t, Q_t]} \,, \tag{2}$$

     where $\mathbb{1}_{\hat{p}_t \notin [q_t, Q_t]}$ is the indicator function.

   - *Standard deviation error function*: if we define $\sigma_t = \sigma(t;P)$ as the standard deviation of the set $\{p^i(t;\pi)\}_{i=1}^{n}$ given a time instant t, the standard deviation error function F_σ is defined as

     $$F_\sigma(P) = \sum_{t \in T} \sigma(t;P). \tag{3}$$

   As both functions are orthogonal, we are facing a multi-objective optimization problem.

4. Check if $\left[F_{io}\left(\hat{\mathbf{p}}, P\right), F_\sigma(P)\right]$ is a local best and/or a global best. It is considered as local best if the particle is not Pareto dominated by any of the previous local bests, and global best if the solution is not Pareto dominated by any of the previous global bests.

5. Update particle randomly or via the velocity term, using the classical PSO implementation [6].

6. Repeat from 2 until convergence is achieved.

Generally, the result of optimization a set of solutions (the Pareto front) called Pareto-optimal solutions, which are not dominated by any other solution.

3 Results

After the application of the optimization algorithm on the model and the data, a Pareto front has been obtained, from which a good solution has been chosen, preferring solutions with a low inside-outside error F_{io}, ensuring that the majority of the data are within the 95%-confidence interval. Based on these criteria, the chosen solution has been

$$\begin{aligned} &\pi^* = (\mu_\tau, \sigma_\tau, \mu_{\theta_0}, \sigma_{\theta_0})^* = (-1.1507, 0.0210, -3.9327, 0.4523)\,, \\ &\tau \sim \text{log-}\mathcal{N}(-1.1507, 0.0210), \;\; \theta_0 \sim \mathcal{N}(-3.9327, 0.4523), \\ &[F_{io}^*, F_\sigma^*] = [0.3233, 2.6597]. \end{aligned} \tag{4}$$

The probability distributions of the model parameters and the fit to the data are shown in Figure 1.

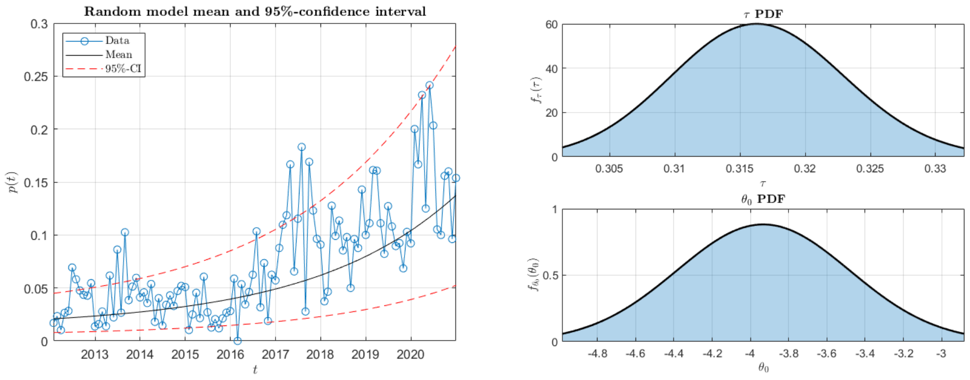


Figure 1: Random model expected value and 95%-confidence interval (left) and model parameters PDFs (right).

4 Conclusions

A general model for antibiotic resistance proportion dynamics considering antibiotic consumption have been analyzed and successfully applied to a real-world case of antibiotic resistance: the colistin-resistant *A. baumannii*. Additionally, a complete random calibration method based on the MOPSO algorithm and Monte Carlo method have been performed to find the random model parameters that best fit to real resistance data series, successfully capturing the randomness within a 95%-confidence interval.

References

[1] European Center of Disease Prevention and Control (ECDC), https://www.ecdc.europa.eu. [Accessed: 14/06/2022].

[2] T. M. Barbosa and S. B. Levy. The impact of antibiotic use on resistance development and persistence. *Drug resistance updates*, 3(5):303–311, 2000.

[3] B. G. Bell, F. Schellevis, E. Stobberingh, H. Goossens and M. Pringle. A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. *BMC Infectious Diseases*, 14(1):1–25, 2014.

[4] Y. Cai, D. Chai, R. Wang, B. Liang and N. Bai. Colistin resistance of acinetobacter baumannii: clinical reports, mechanisms and antimicrobial strategies. *Journal of Antimicrobial Chemotherapy*, 67(7):1607–1615, 2012.

[5] C. A. Coello Coello and M. S. Lechuga. MOPSO: a proposal for multiple objective particle swarm optimization. *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*. IEEE, 2002.

[6] F. Marini and B Walczak. Particle swarm optimization (PSO). a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 149:153–165, December 2015.

[7] J. O'Neill. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. Technical report, Review on Antimicrobial Resistance, UK Governement, 07 2016. [Accessed: 31/10/2021].

[8] A. Pormohammad, K. Mehdinejadiani, P. Gholizadeh, M.J. Nasiri, N. Mohtavinejad, M. Dadashi, S. Karimaei, H. Safari and T. Azimi. Global prevalence of colistin resistance in clinical isolates of acinetobacter baumannii: A systematic review and meta-analysis. *Microbial Pathogenesis*, 139:103887, 2020.

[9] R. A. Smith, N. M. M'ikanatha and F. Read Andrew. Antibiotic resistance: a primer and call to action. *Health Communication*, 30(3):309–314, 2015.

[10] M. Sundqvist. Reversibility of antibiotic resistance. *Upsala Journal of Medical Sciences*, 119(2):142–148, 2014.

[11] M. Sundqvist, P. Geli, D.I. Andersson, M. Sjülund-Karlsson, A. Runehagen, H. Cars, K. Abelson-Storby, O. Cars and G. Kahlmeter. Little evidence for reversibility of trimethoprim resistance after a drastic reduction in trimethoprim use. *Journal of antimicrobial chemotherapy*, 65(2):350–360, 2010.

[12] F. C. Tenover. Mechanisms of antimicrobial resistance in bacteria. *The American Journal of Medicine*, 119(6):S3–S10, 2006.

[13] Valencian Government. Microbiological Surveillance Network of the Valencian community. http://www.sp.san.gva.es/sscc/. [Accessed: 31/10/2021].

[14] World Health Organization (WHO). Antimicrobial resistance. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance, 2020. [Accessed: 21/07/2022].

3

Relative research contributions towards the application and materialization of Wenner four-point method on concrete curing

L. Andrés[♭1], J. H. Alcañiz[♭], T. P. Real[♮], P. Suárez[♢]

(♭) Universidad Católica San Antonio de Murcia,
Av. de los Jerónimos 135, Guadalupe de Maciascoque, Murcia, Spain.
(♮) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(♢)Idvia 2020 Horizonte 2020 SL.
Av. d'Aragó 30, València, Spain.

1 Introduction

Concrete is one of the most widely used materials in construction. This is due to the numerous advantages it offers in terms of durability, versatility, and cost, among others. Cement is the hydraulic binder of concrete, so the compressive strength depends, to a large extent, on the chemical reaction of cement hydration [1]. Consequently, this publication deals with the determination of the compressive strength of concrete during the cement hydration process, and the prediction at early ages of the compressive strength at day 28.

For this purpose, the chemical reaction of the cement hydration process has been studied, in which tricalcium silicate, dicalcium silicate, tricalcium aluminate, tetracalcium aluminoferrite react with water to form ettringite, calcium hydroxide and calcium silicate gel [2]. In addition, an internal structure of capillaries and cavities arises with ions dissolved in the water acting as a reagent. Under these conditions, when an electric field is applied, an accelerated movement of ions occurs.

This explains how increases in electrical resistivity indicate the decrease in porosity, which is the fundamental mechanism that characterizes concrete hydration and compressive strength gain over time. However, it is influenced by several factors related to concrete mix composition (such as cement type, cement content, water-cement ratio, and use of admixtures and supplementary cementations materials). These factors influence the microstructure of concrete as well as the pore solution chemistry, and therefore influence its electrical resistivity.

2 Methodology

With this concept, a system capable of measuring the ions through the electrical resistivity measurement has been developed. The developed system is an own hardware development based on

[1] landres3@alu.ucam.es

Wenner's four-point method. Throughout the development of the hardware, a study of the optimal electrode design, excitation methodology, power supply system and communication system has been carried out.

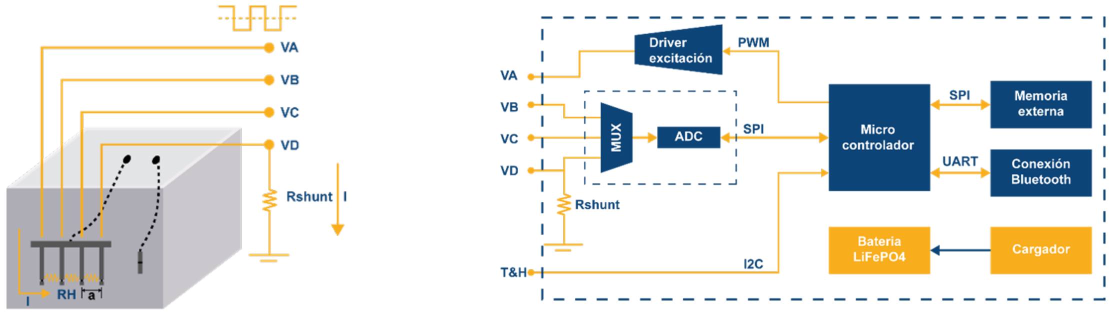


Figure 1: Harware scheme to resisitivity measures.

Finally, the resistivity value is obtained from the following expression

$$\rho = 4\pi a \frac{R_{shunt}(V_B - V_C)}{V_D} \tag{1}$$

Once the way to measure the electrical resistivity of the concrete has been established, an algorithm based on artificial intelligence has been developed to determine the relationship between resistivity and compressive strength. As well as the methodology to predict at early ages the value that the compressive strength of the concrete will acquire on day 28. For the training of the algorithm, compressive strength and electrical resistivity data have been taken during the cement hydration process in different dosages.

The aim of this document is to train the neural network to obtain the curves that define the hardening of the analysed dosages. From these curves it will be possible to determine the real resistance of an instrumented element using the system as well as to identify anomalies in early stages of curing.

3 Test description

For the demonstration of the functioning, the following tests have been considered:

- A total of 2 dosages will be carried out. Both dosages will be cured under controlled (indoor) conditions D1 and D2.

Table 1: Composition of two dosages for testing.

	Mix Design (Kg/m^3)			Admixtures	Admixtures Dosage (% cement)	Cement Quantity (kg/m^3)	SLAG Quantity (kg/m^3)	Total Water (l/m^3)
	Sand 0-4	Crushed 12/19	Gravel 25					
D1	923	395	648	MasterGlenium Sky 698	0.5	330	--	155
D2	923	395	648	MasterGlenium Sky 698	0.5	230	100	160

- One of the dosages (D1) will be repeated and left to cure under not controlled conditions (outdoor) (D1').

- Two mixtures (Test) shall be carried out on each dosage. For each of the mixtures, density, porosity (% air) and consistency (concrete slump test) measurements shall be taken.
- A total of 16 test specimens shall be made from each test. Two specimens will be instrumented with the systems and the other 14 will be used to carry out two compressive strength tests at 1, 2, 7, 14, 21, 28, 60 days after mixing.
- The measured data together with the data collected by the hardware system will be used to extract the theoretical resistivity and resistivity evolution curves and the Resistivity-Resistivity curve for each dosage under ideal conditions, as well as for the third dosage at outdoor conditions.

4 Neural Network training

The system performs periodic measurements of resistance (Ohms) over test piece and the environmental conditions temperature and humidity each two hours. The resistivity values are strongly influenced by environmental conditions, therefore the system performs periodic measurements of temperature and humidity in order to correct the measured values.

The curves of all dosages together are shown below to see the similarity between tests of the same dosage and the differences between different dosages.

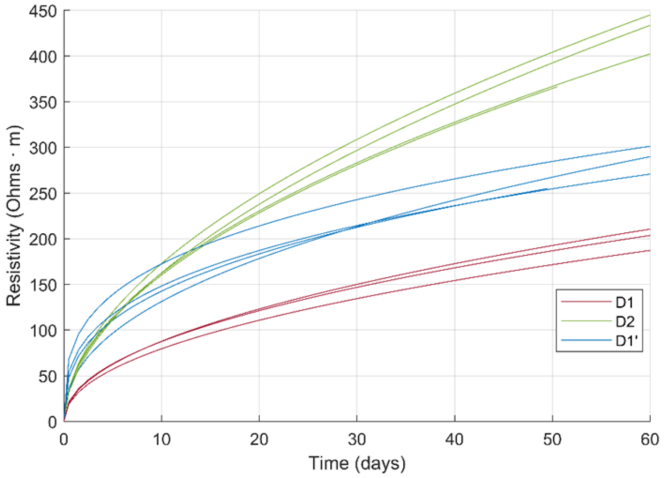


Figure 2: Corrected resisitivity measures of dosages

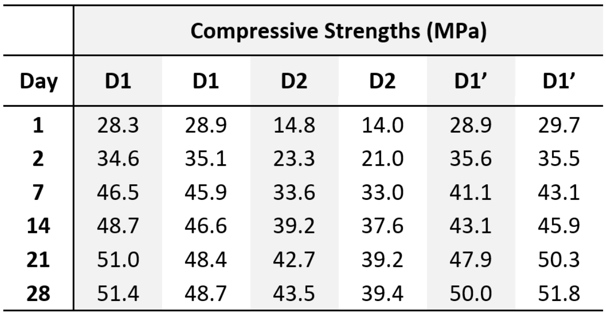
This is due to the fact that outdoor conditions (heat and wind) favour the evaporation of water from the concrete, reducing the amount of water left to react with the cement and harden. This loss of water causes a higher increase in resistivity than the curing process without water loss.

As the amount of water is less than necessary for the chemical reaction to take place, the chemical reaction is not completed, not all of the cement hardens and therefore it can be expected that the dosage 3 will not reach the expected strength.

4.1 Compressive strength

Over 14 specimens was carried out two compressive strength tests at 1, 2, 7, 14, 21, 28 and 60 days after mixing.

Table 2: Measures of compressive strengths.



	Compressive Strengths (MPa)					
Day	**D1**	**D1**	**D2**	**D2**	**D1'**	**D1'**
1	28.3	28.9	14.8	14.0	28.9	29.7
2	34.6	35.1	23.3	21.0	35.6	35.5
7	46.5	45.9	33.6	33.0	41.1	43.1
14	48.7	46.6	39.2	37.6	43.1	45.9
21	51.0	48.4	42.7	39.2	47.9	50.3
28	51.4	48.7	43.5	39.4	50.0	51.8

4.2 Neural Network training

The selected network structure consists of two hidden layers, the first of 15 neurons and the second of 7.

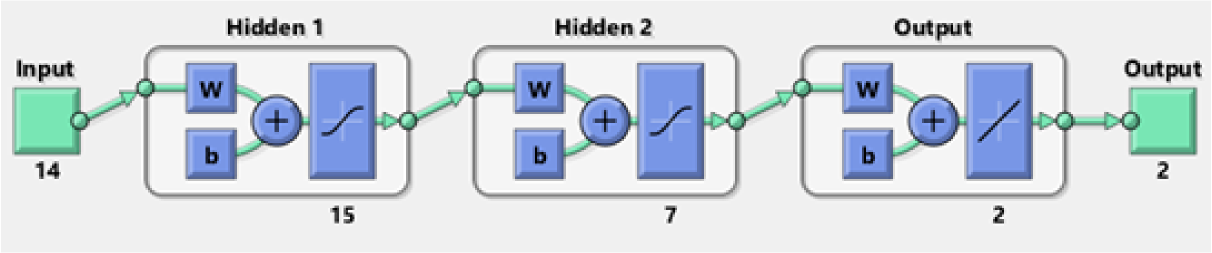
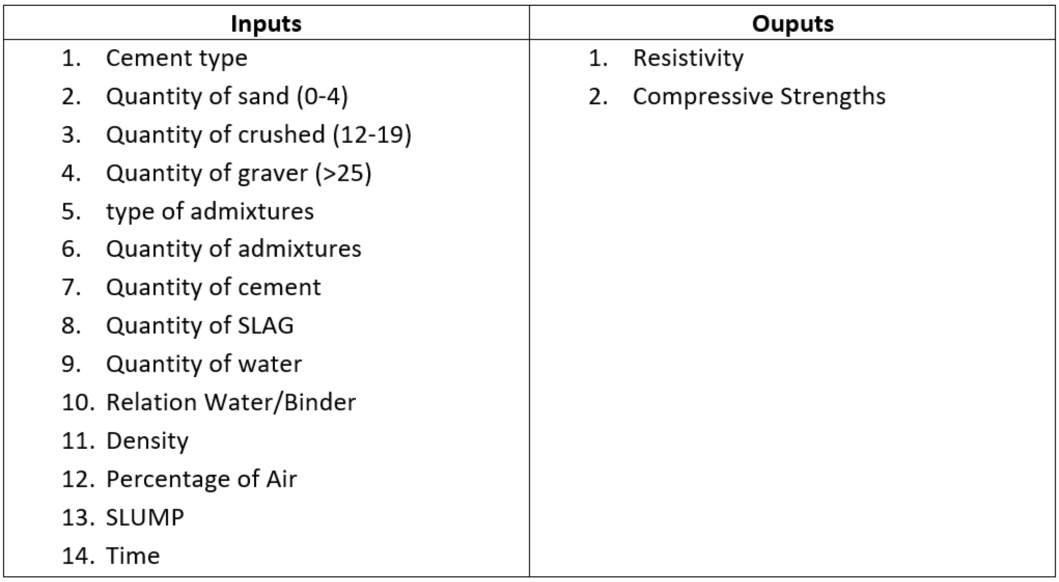


Figure 3: Neural Network structure

Whose inputs and outputs are:

Table 3: Neural Network inputs and outputs.



Inputs	Ouputs
1. Cement type	1. Resistivity
2. Quantity of sand (0-4)	2. Compressive Strengths
3. Quantity of crushed (12-19)	
4. Quantity of graver (>25)	
5. type of admixtures	
6. Quantity of admixtures	
7. Quantity of cement	
8. Quantity of SLAG	
9. Quantity of water	
10. Relation Water/Binder	
11. Density	
12. Percentage of Air	
13. SLUMP	
14. Time	

5 Results and conclusions

The training results show a mean squared error of less than 0.0008 and a regression greater than 0.99 for training dataset, validation dataset and total dataset. These results show satisfactory training, however, as the inputs and outputs are normalised, they do not show the real error that is made.

In order to have an estimation of the error it is most appropriate to compare the real curves with the theoretical curves for each of the dosages.

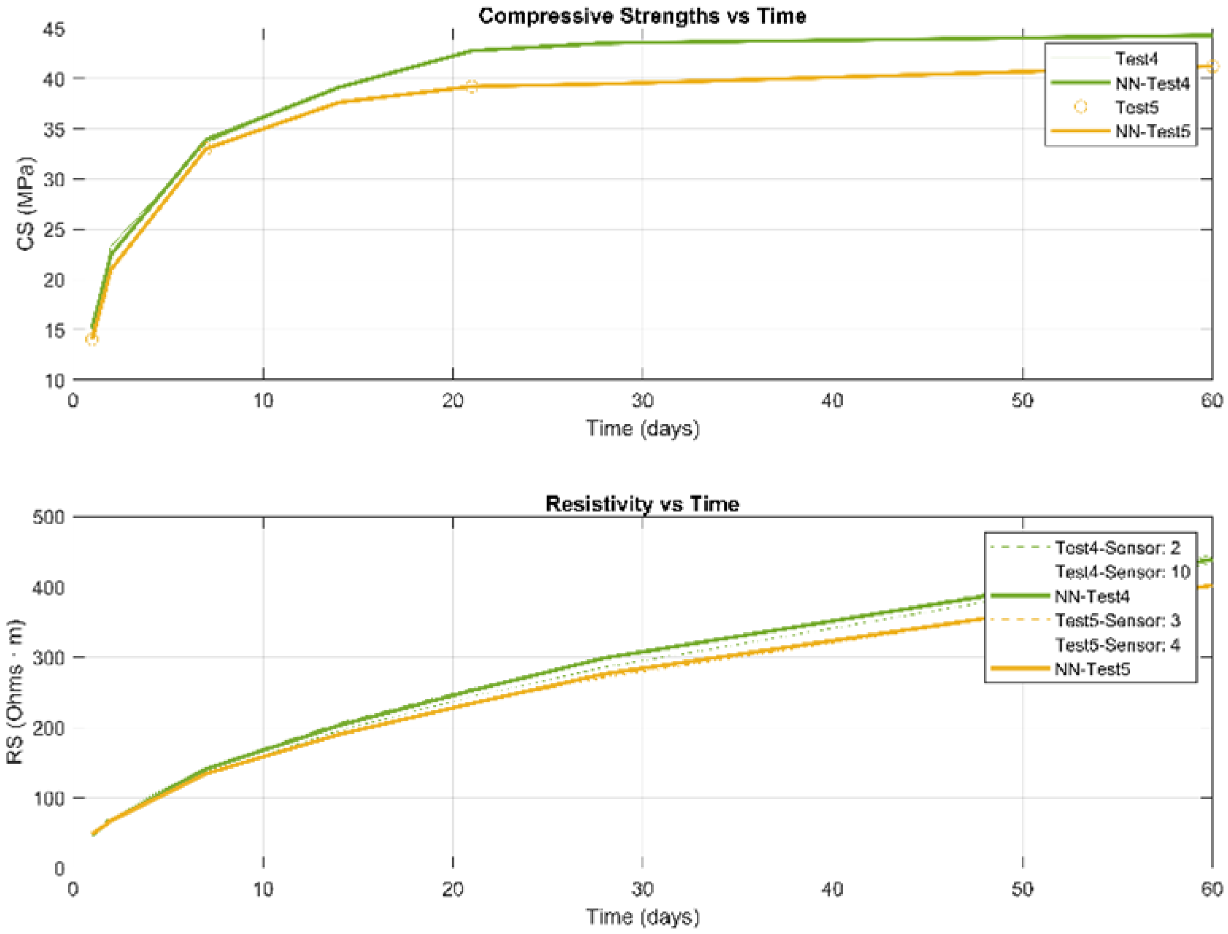


Figure 4: Theoretical curves (RNN) and measured curves fo D2.

As can be seen in the graphs above, the values obtained by the neural network agree with the measured data. For a more objective analysis, the errors have been calculated for each of the measurements of resistivity (RS) and strengths (RC). According to the following expression:

$$RX\ Error(\%) = 100\frac{abs(RX_{real} - RX_{RNN})}{RX_{real}}, \quad \text{where } X = C\ or\ S \tag{2}$$

Table 4: RC Error (%) of each test. Mean (RC Error) = 0.4253%.

RC Error (%)							
Test	Days						
	1	2	7	14	21	28	60
2	0.412	0.413	0.106	0.014	0.059	0.366	0.119
3	-	0.006	-	0.016	-	0.011	-
4	2.164	3.600	0.927	0.224	0.133	0.038	0.035
5	0.019	-	0.007	-	0.016	-	0.037
6	0.964	2.176	0.796	0.073	0.079	0.171	0.151
7	-	0.016	-	0.014	-	0.024	-

Table 5: RS Error (%) of each test. Mean (RS Error) = 4.9467 %.

RS Error (%)								
Test	Sensor	Days						
		1	2	7	14	21	28	60
2	13	1.865	6.841	5.552	4.095	4.497	7.591	4.055
	14	9.511	3.501	4.101	5.243	4.532	0.961	4.319
3	1	6.961	1.100	1.809	0.529	1.392	0.130	0.963
4	2	4.637	7.896	1.641	1.782	1.222	0.436	1.299
	10	5.954	0.984	5.250	3.691	3.462	4.612	1.291
5	3	0.889	5.992	0.770	1.220	0.902	1.125	0.024
	4	1.215	4.249	0.586	-0.135	0.031	1.964	0.096
6	8	14.965	20.688	16.306	9.221	6.907	6.591	4.376
	9	16.166	4.338	3.117	7.888	8.332	7.088	4.865
7	5	34.166	15.440	5.112	4.983	3.273	0.770	2.194
	11	7.131	13.823	10.355	3.652	1.068	0.483	4.698

This result shows that is possible to determine the optimum moment to remove the shoring, to remove the formwork or to demould. In addition, the compressive strength of the concrete after the hydration process of the cement can be known at an early stage.

References

[1] College of Resource and Environment, S. U., High-performance superplasticizer based on chitosan. *Biopolymers and Biotech Admixtures for Eco-Efficient Construction Materials*, 131-150, 2016

[2] Yuan, Y., Ji, Y., Modeling corroded section configuration of steel bar in concrete structure. *Construction and Building Materials*, 23:2461–2466, 2009

4

Application of mathematical models and multicriteria analysis to establish an optimized priorization for the maintenance of bridges in a network

L. Andrés[♭],[1] F. Ribes[♮], E. Fernández[♢] and J. Maldonado[♭]

(♭) Idvia 2020 Horizonte 2020 SL,
Av. d'Aragó 30, València, Spain.
(♮) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(♢) ISA INTERVIAL,
10, Cerro El Plomo 5630, Las Condes, Región Metropolitana, Chile.

1 Introduction

Infrastructure maintenance investment needs are growing yearly due to the increasing ageing of the infrastructure and the higher stresses it is exposed to. Therefore, in addition to ensuring safety, maintenance investment optimisation must be a priority for infrastructure managers, as delaying investment increases the costs and risks of ageing infrastructure.

Nowadays, most maintenance plans are designed based on the findings of traditional visual inspections. The effectiveness of these visual inspections depends entirely on the skills and experience of the bridge inspector, and these defects must have a visible manifestation. The transmission of information to the assessing engineer, who is responsible for deciding on any necessary action, is as accurate and consistent as possible and is also a critical factor.

In recent years, non-destructive techniques (NDT) have been developed to extend the range of identifiable defects (i.e. both visible and non-visible). However, even if an inspector employs such equipment, the inspection still has the same spot and discrete nature that characterises and limits the usefulness of traditional visual inspections.

Faced with the decision that the Manager of these structures must make as to which bridges would be most strategic and beneficial to install such systems, several questions arise that this study has addressed.

[1] laura.andres@idvia.es

2 Methods

2.1 Problem definition

As is well known, all structures require periodic maintenance, which, if not applied in response to the needs of the bridge or at the right time, can compromise the structure's functionality or even its stability. This makes it mandatory to ask the following questions:

- What maintenance actions does the bridge need?
- How often does it need them?

At this point, monitoring based on the implementation of advanced SHM systems based on developing a Digital Twin of the bridge provides a valuable aid. These techniques allow us to know the bridge's structural needs, thus minimising the investment in maintenance by acting at the optimum time through the necessary maintenance or reinforcement operation.

Bearing in mind that all bridges need maintenance and that structural monitoring is a tool that allows optimising and prioritising maintenance actions, it can be stated that having a high percentage of monitored bridges can be interesting. However, this objective cannot be achieved in the short term. Based on this approach, the question is: **Which bridges should be monitored first?**

To this end, all the available data and background information on the bridges of the networks under study is analysed based on a series of criteria with which we can rank the bridges that should be monitored first in the most objectively and exhaustively way possible.

2.2 Choosing evaluation criteria

The different criteria considered for the evaluation of the bridge networks are set out below. Parameters that have a high impact on bridge maintenance should be chosen. They can be divided in:

- **Strategic parameters:** traffic intensity and alternative routes.
- **Structural parameters:** age, structural health and bridge length.
- **Environment parameters:** hydrology, scour, seismicity and liquefaction.

2.3 Scoring system

For each criterion, a parameter has been defined ranging from 0 to 50 depending on its importance on bridge maintenance. Each of the parameters considered in the previous section are shown below.

- **Traffic intensity**. SHM should prioritize high-demand bridges in the network. TMDA (Tránsito Medio Diario Anual) can be used to score this criterion and obtain the Traffic Intensity Index (TII). Evolution of this index can be seen in Figure 1 left.
- **Alternative routes**. The more time required to take an alternative route, the higher the Redundancy Index (RI) should give importance to the poorly connected bridges (Figure 1 right).

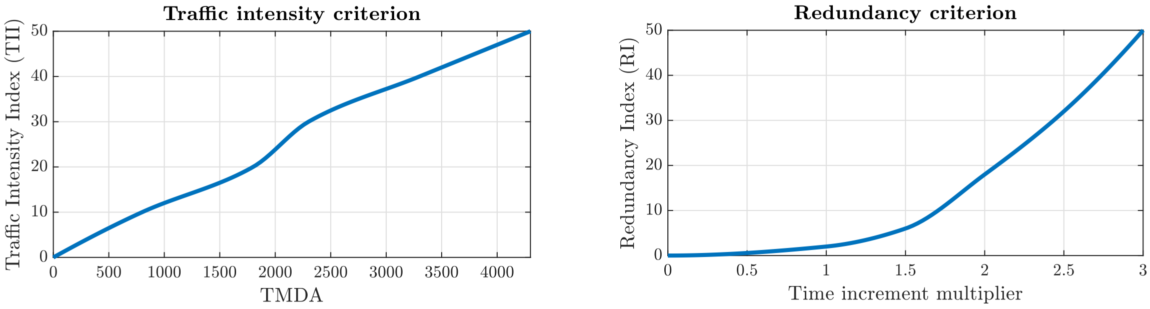


Figure 1: Traffic Intnesity Index (left), Redundancy Index (right).

- **Age**. Older structures are more attractive than new ones in Advanced SHM Systems. For this reason, Age Index (AI), that can be seen in Figure 2 left, is defined as giving less weight to newer bridges.
- **Structural Health**. Structural health is also an interesting parameter for determining which structures SMH Systems should be installed. The visual inspection score defines Structural Health Index (SHI). This score is based on the number of defects found during the inspection (Figure 2 right).

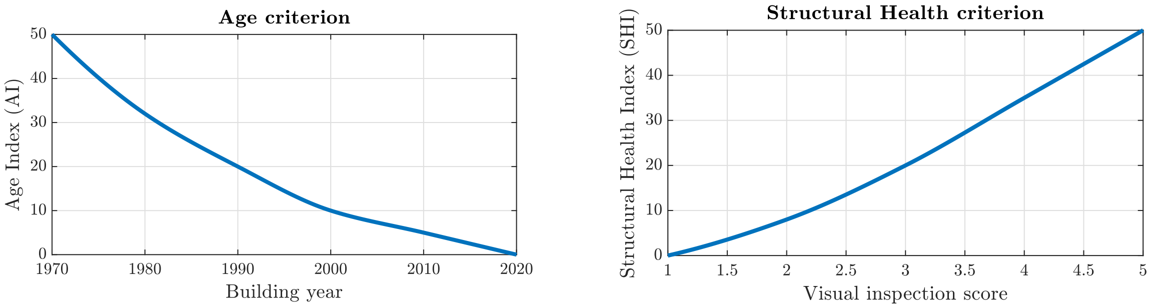


Figure 2: Age Index (left), Structural Health Index (right).

- **Bridge length**. The bigger the bridge size is, the more maintenance investment is required to ensure the safety of its users. Bridge Length Index (BLI) is defined to keep this in mind. Evolution of this index can be seen in Figure 3 left.
- **Hydrology**. The riverbed behind the bridge should be considered if it has any of its piles in the river. The riverbed type, near vegetation and Strahler order are combined to define the Hydraulic Index (HI) shown in Figure 3 right.

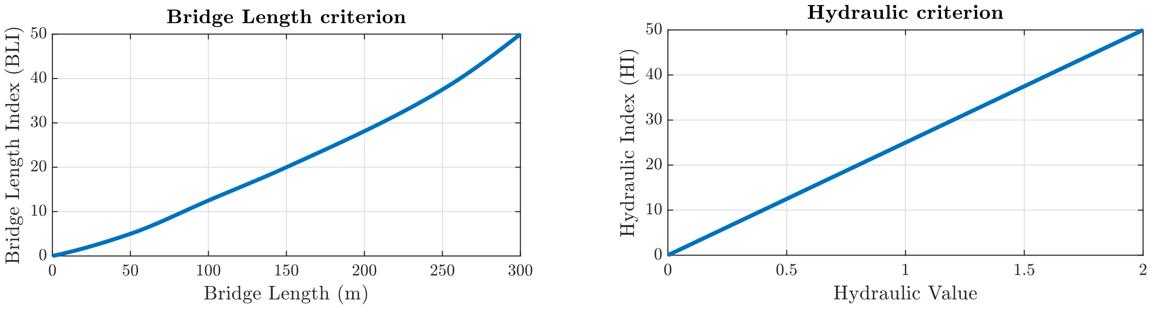


Figure 3: Bridge Length Index (left), Hydraulic Index (right).

- **Scour**. Scour is a phenomenon that can be very dangerous to the structure in some situations. Therefore, this parameter must be considered a criterion to determine which bridges to monitor first (Figure 4 left).

- **Seismicity**. Seismic activity is essential in some situations since these phenomena can affect structure stability and integrity. Therefore a seismic criterion is made based on the seismic zone of the structure to analyze as shown in Figure 4 right. Liquefaction is also a phenomenon to be considered. It is linked to the terrain type and seismic activity and occurs more frequently in cohesionless and saturated soils. The liquefaction Index (LI) is defined in terms of the ISL (Índice de Susceptibilidad a la Licuefacción) and is a modifier of the Seismicity Index.

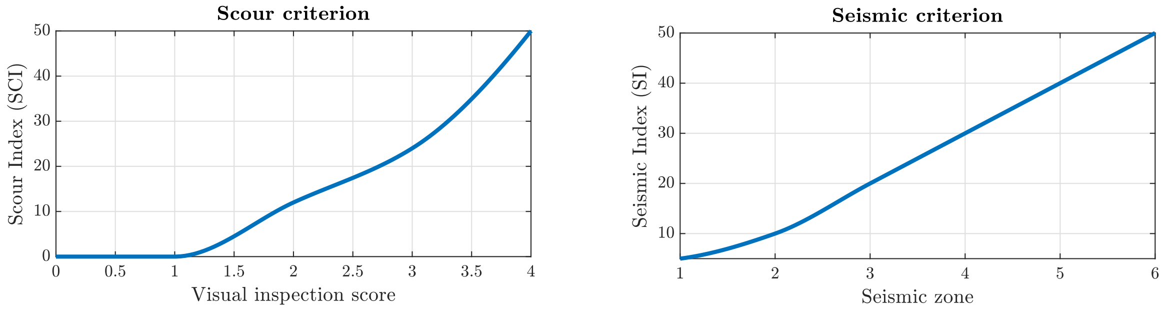


Figure 4: Scour Index (left), Seismic Index (right).

In order to obtain the final score, a weighted average of each index is made in Equation 1, where the weight of each index can be varied according to the analyst's needs. The weights assigned for each criterion are shown in Table 1

$$Score = \sum_{i=1}^{8} \frac{PI_i \cdot W_i}{100} \tag{1}$$

PI_i	**TII**	**RI**	**AI**	**SHI**	**BLI**	**HI**	**SCI**	**SI**
W_i (%)	10	5	10	20	10	10	20	15

Table 1: Assigned weights for each index.

3 Results

With the application of the scoring system to every bridge of the network, structures with higher scores can be selected. In these bridges, the installation of SMH Systems should be prioritized in order to minimize costs.

As an example of application, some of the bridges belonging to the "Ruta de los Ríos" (Chile) have been analysed (Figure 5). A higher score will indicate a higher monitoring priority, allowing, at a glance, to decide which bridges in the network to apply advanced SHM systems.

Rucaco Bridge, San Pedro Bridge and Cruces Bridge are the three highest-scoring bridges in this case (Figure 6). Therefore, these bridges should be prioritised for SHM installation.

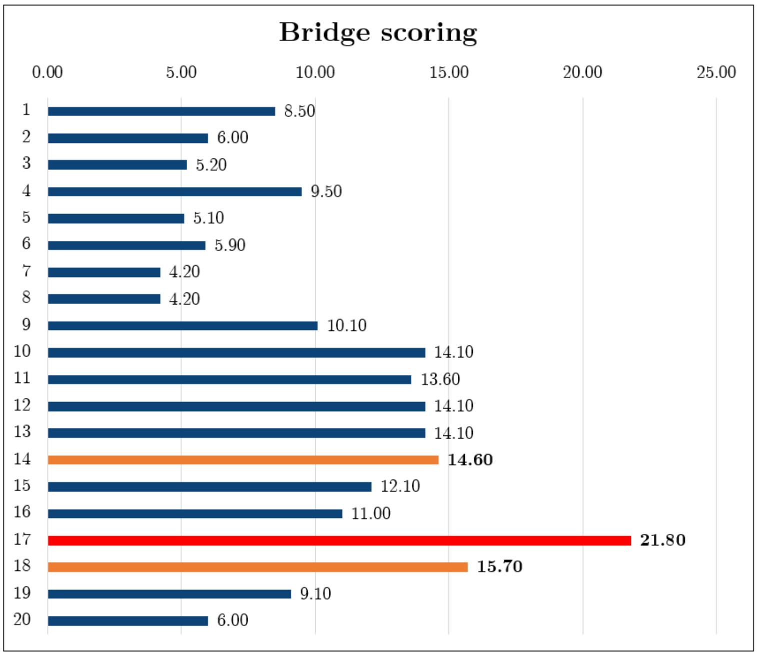
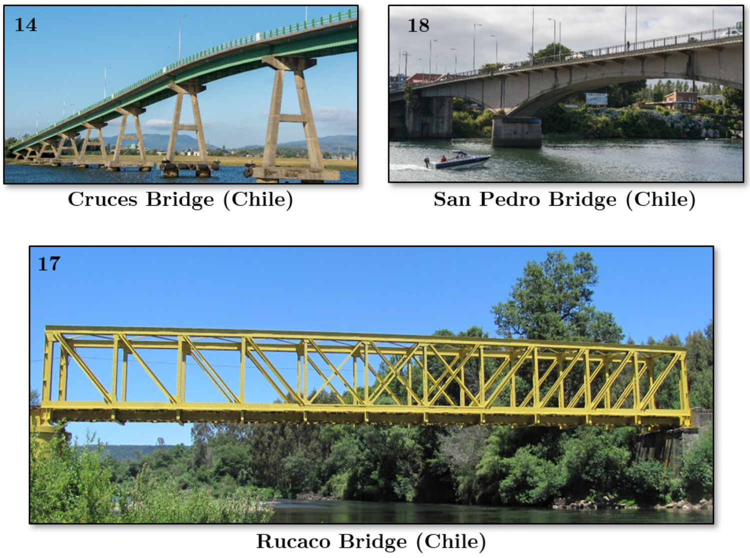


Figure 5: Bridge scoring in “Ruta de los Ríos”.



Cruces Bridge (Chile)

San Pedro Bridge (Chile)

Rucaco Bridge (Chile)

Figure 6: Highest-scoring bridges in “Ruta de los Ríos”.

4 Conclusions

Advanced SMH Systems can provide more frequent or continuous information about the structure, but they cannot be installed in every network bridge. A tool to select in which bridges to install SMH has been developed based on different parameters affecting costs (strategic, structural and environmental). As output, a bridge rank can be obtained, and the structures with a higher score are more suitable to install SHM Systems.

References

[1] Leyton, F., mapa de peligro sísmico actualizaco. Centro Sismológico Nacional, Universidad de Chile. Santiago, 2014.

[2] SERNAGEOMIN, Servicio Nacional de Geología y Minería, Mapa Geológico de Chile. Escala 1:1.000.000, 2002.

[3] Leyton F., Ruiz S. and Sepúlveda S., Reevaluación del peligro sísmico probabilístico en Chile central *Andean Geology*, págs. 455-472, 2012.

[4] Youd T. and Perkins D., Mapping liquefaction-induced ground failure potential *Journal of the Geotechnical Engineering Division*, 1978.

5

Probabilistic analysis of scalar random differential equations with state-dependent impulsive terms via probability density functions

V. Bevia ♭,[1] J. C. Cortés ♭ M. Jornet # and R.J. Villanueva♭

(♭) Instituto de Matemática Multidisciplinar (Imm),
Universitat Politècnica de València (UPV)
Camí de Vera s/n, València, Spain.

(#) Departament Matemàtiques,
Universitat de València (UV),
Burjassot (València), Spain.

1 Introduction

Most phenomena observed in nature can be described as a function changing smoothly over time. However, sometimes the system suddenly changes its state, requiring special mathematical tools to model its dynamics correctly. This is common in biology, medicine, or engineering when determining the effectiveness of specific impulsive-type control strategies. In ecology, these are known as harvesting models [5, 9, 10].

In this contribution, we will study, from a probabilistic standpoint, the following random Initial Value Problem (IVP):

$$\begin{cases} \dfrac{\mathrm{d}X(t,\omega)}{\mathrm{d}t} &= g(X(t,\omega),t,\mathbf{A}(\omega)) - \displaystyle\sum_{k=1}^{N}\Gamma_k(\omega)\delta(t-t_k)X(t,\omega), \quad t>t_0, \\ X(t_0,\omega) &= X_0(\omega). \end{cases} \tag{1}$$

Here t_0 denotes a real number; $X_0(\omega)$, $\mathbf{A}(\omega) := (A_1(\omega),\dots,A_m(\omega))$ and $\{\Gamma_k(\omega)\}_{k=1}^{N}$ are assumed to be independent absolutely continuous random variables defined on the Hilbert space $\mathrm{L}^2(\Omega,\mathbb{R})$, whose elements are real-valued random variables with finite variance and $(\Omega,\mathcal{F},\mathbb{P})$ denotes a complete probability space [6]; $\delta(t-t_k)$ stands for the Dirac delta function [4] acting at the prefixed time instants $t=t_k$, $k=1,\dots,N$ and g is known as the *(scalar) field function* satisfying certain conditions that will be specified later. Finally, $X(t,\omega)$ denotes the solution of the random IVP (1)

[1] vibees@doctor.upv.es

2 Methods and Results

2.1 Pathwise solution

The deterministic Laplace transform [2] and the usual conditions required for the existence and uniqueness of ODEs have allowed us to construct a right-continuous pathwise solution of the random IVP (1), given by

$$X(t,\omega) = X_0(\omega) + \int_{t_0}^{t} g(X(s,\omega), s, \mathbf{A}(\omega))\mathrm{d}s - \sum_{k=1}^{N} \Gamma_k(\omega) X(t_k,\omega)\mathrm{H}(t-t_k), \quad t \geq t_0, \tag{2}$$

$$X(t_k,\omega) = \frac{X(t_k^-,\omega)}{1+\Gamma_k(\omega)} = X(t_k^+,\omega), \quad \omega \in \tilde{\Omega}. \tag{3}$$

2.2 Probability Density Function evolution

RDEs verifies a *probability conservation property*; that is, the total probability in the phase space is conserved through time. This fact gives an evolution PDE which is verified by its 1-PDF. The theorem can be stated as follows

Theorem 1. *[1] Let $b(\cdot,t) : \mathbb{R} \longrightarrow \mathbb{R}$ be a Lipschitz-continuous function for all $t \in (t_0,\infty)$, and continuous in t. Let $X(t,\omega)$, $t \geq t_0$, $\omega \in \Omega$ be the stochastic process verifying the following RDE in the almost-surely or mean square sense:*

$$\begin{cases} \dfrac{\mathrm{d}X(t)}{\mathrm{d}t} &= b(X(t),t), \quad t > t_0, \\ \\ X(t_0) &= X_0 \in \mathrm{L}^2(\Omega,\mathbb{R}). \end{cases} \tag{4}$$

Let $\mathcal{D}$ be a set such that $\{X([t_0,\infty),\omega)\}_{\omega\in\Omega} \subset \mathcal{D}$. Then, the 1-PDF of the stochastic process $X(t)$, denoted by $f = f_{X(t)}$, verifies the Liouville PDE:

$$\begin{cases} \partial_t f(x,t) + \partial_x[b\, f](x,t) = 0, \quad x \in \mathcal{D}, \quad t > t_0, \\ \\ f(x,t_0) = f_0(x), \quad x \in \mathcal{D}, \\ \\ \partial_x f(x,t) = 0, \quad x \in \partial\mathcal{D}, \quad t \geq t_0, \end{cases} \tag{5}$$

where f_0 is the PDF of $X_0 = X_0(\omega)$.

When the RDE has random parameters, IVP (5) becomes a family of deterministic PDE problems indexed in the realizations, $\mathbf{a}$, of the random parameter vector, $\mathbf{A} = \mathbf{A}(\omega)$,

$$\begin{cases} \partial_t f(x,t\,|\,\mathbf{a}) + \partial_x[b(x,t,\mathbf{a}) f(x,t\,|\,\mathbf{a})] = 0, \quad x \in \mathcal{D} \subseteq \mathbb{R}, \quad t > t_0, \\ \\ f(x,t_0\,|\,\mathbf{a}) = f_0(x), \quad x \in \mathcal{D}. \end{cases} \tag{6}$$

The PDF of the RDE solution (independent of parameter realizations) is obtained by marginalizing the joint PDF of both the solution and the parameter vector $\mathbf{A}$, which, using the conditional PDF can be written as:

$$f(x,t) = \int_{\mathbb{R}^m} f(x,t\,|\,\mathbf{a}) f_{\mathbf{A}}(\mathbf{a})\mathrm{d}\mathbf{a} = \mathbb{E}_{\mathbf{A}}[f(x,t\,|\,\mathbf{A})], \tag{7}$$

where $f_{\mathbf{A}}$ is the parameters' joint PDF and $\mathbb{E}$ denotes the expectation operator. This shows that the PDF can be obtained by solving (6) for all realizations $\mathbf{a}$ of $\mathbf{A}$ and then computing its mean.

However, a *priori*, global-in-time existence of a solution to the Liouville equation can only be assured when the field function $b(\cdot, t)$ is Lipschitz continuous, uniformly in t. The field of the RDE class under study, $b(x,t) = g(x,t) - \sum_k \gamma_k \delta(t-t_k)x$, does not verify this hypothesis at the impulse times $\{t_k\}_{k=1}^N$. We want to obtain a condition such as (3), but for the PDF. This will allow the computation of the evolution of f_0, accurately capturing the discontinuities at the impulse times.

Let us turn back to the set of conditions in (3), which are identities between random variables. We are going to make use of the RVT theorem, which can be written as follows:

Theorem 2. *[2,8] Let $\mathbf{X}$, $\mathbf{Y} : \Omega \to \mathbb{R}^M$ be two random vectors with PDFs $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$, respectively. Assume that there is a one-to-one, C^1 function $\mathbf{h}$ such that $\mathbf{X} = \mathbf{h}(\mathbf{Y})$. Then, denoting $\mathbf{h}^{-1}$ as the inverse mapping of $\mathbf{h}$,*

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{h}^{-1}(\mathbf{x})) \left| \frac{\partial \mathbf{h}^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right|, \tag{8}$$

where $\left| \frac{\partial \mathbf{h}^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right|$ denotes the absolute value of the determinant of the Jacobian matrix.

Now, the Liouville equation describes the evolution of the PDF between impulse times, whereas at any impulse time, applying the RVT theorem, we obtain:

$$f(x, t_k) = \mathbb{E}_{\Gamma_k}[f\left(x\,(1+\Gamma_k), t_k^-\right)|1+\Gamma_k|], \quad \forall x > 0, \quad k = 1, \ldots, N, \tag{9}$$

because of the relations at (3).

3 Results

In this contribution, we are going to deal with the impulse-harvest generalized logistic model with a finite number of captures, say N,

$$\begin{aligned} X'(t,\omega) &= \alpha(t)\, r(\omega) X(t,\omega) \left(1 - \left(\frac{X(t,\omega)}{K(\omega)}\right)^{\nu(\omega)}\right) - \sum_{n=1}^N \Gamma_n(\omega)\delta(t-t_n)X(t,\omega), \\ X(t_0,\omega) &= X_0(\omega), \end{aligned} \tag{10}$$

where $t \geq t_0$ and $\omega \in \tilde{\Omega}$. As usual, t is interpreted as the time, the parameter r is the growth ($r > 0$) or decay ($r < 0$) rate, and K is the carrying capacity. The differential equation is generalized by adding two terms: a positive, monotonically growing function $\alpha(\cdot)$ and a constant positive term ν. The first term, $\alpha(\cdot)$, allows controlling the so-called *lag phase*, which is the growth phase in which the population under study has not yet achieved a fully exponential growth. In particular, we have chosen [7]:

$$\alpha(t) := \frac{q(\omega)}{q(\omega) + \mathrm{e}^{-m(\omega)\, t}}, \quad q, m > 0 \text{ a.s.}$$

The latter, ν, is a power that controls how fast the carrying capacity K is approached and is known as *deceleration* term. When $\nu = 1$, the classical logistic differential equation is obtained. And when ν tends to 0, the Gompertz equation is given. The incorporation of both the function $\alpha(\cdot)$ and the power ν allows for more flexible S-shaped curves to model growth phenomena over time.

3.1 Tumor removal

Radiotherapy, chemotherapy, and direct retrieval of a fraction of a tumor mass are some of the main techniques used to treat cancer. The first two treatments have a prolonged effect of tumor destruction, whereas the latter can be modeled via a delta-type impulse function because of the sudden extraction of the tumor mass with respect to the total treatment. Let us model this problem as (10), where the parameter vectors are chosen as follows:

- The initial tumor size $X_0 \sim \mathrm{N}|_{(0,1)}(0.15, 0.01)$, where $\mathrm{N}|_{(0,1)}$ is a normal distribution truncated on the interval $(0, 1)$.
- Variables q and m will be given the same deterministic values as in the previous example: $q = 1$ and $m = 4$.
- We consider $r \sim \mathrm{N}|_{(0,1)}(0.15, 0.0075)$, $\nu \sim \mathrm{Unif}(1, 1.25)$ and $K \sim \mathrm{Unif}(0.9, 1)$.
- We are going to consider 5 removals with equally distributed intensity given by $\Gamma \sim \mathrm{N}|_{\mathbb{R}^+}(2, 0.01)$, at times $T_{\mathrm{Tumor}} = \{t_1 = 15, t_2 = 25, t_3 = 35, t_4 = 45, t_5 = 55\}$.

Figure 1, shows the mean and 95%-confidence intervals according to the prefixed parameters and removal times. It is seen how, after each removal, the tumor size grows according to the un-removed size of the tumor. Interestingly, the confidence interval amplitude before each removal is higher than the uncertainty after the removal. Indeed, since all removals are distributed as $\Gamma \sim \mathrm{N}|_{\mathbb{R}^+}(2, 0.01)$, it is easily seen that each removal takes away half of the tumor (in average), thus reducing the uncertainty after each removal time. This case allows having a long-time prediction with a reduced level of uncertainty while still considering random impulses. This is further seen in Figure 2, where the PDF given as the solution of the Liouville equation in this particular problem setting is shown in every simulated time in T_{Tumor}.

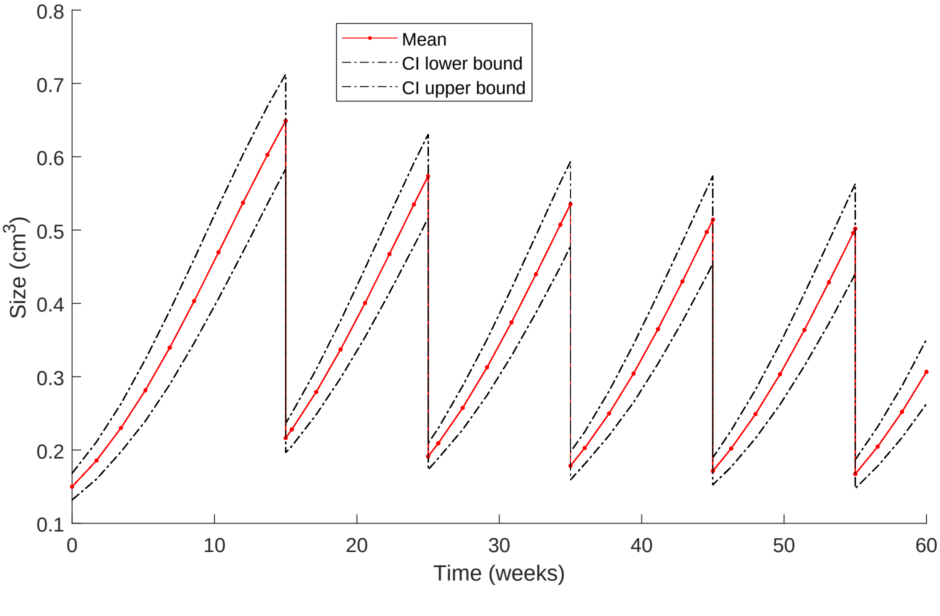


Figure 1: Time evolution of the mean tumor size and a 95% confidence interval with several extractions.

4 Conclusions

In this contribution, we have obtained a pathwise solution to a general random differential equation with a finite number of random-intensity, state-dependent, impulsive terms, with the usual assumptions on the regularity of the field function. Furthermore, we have determined the evolution

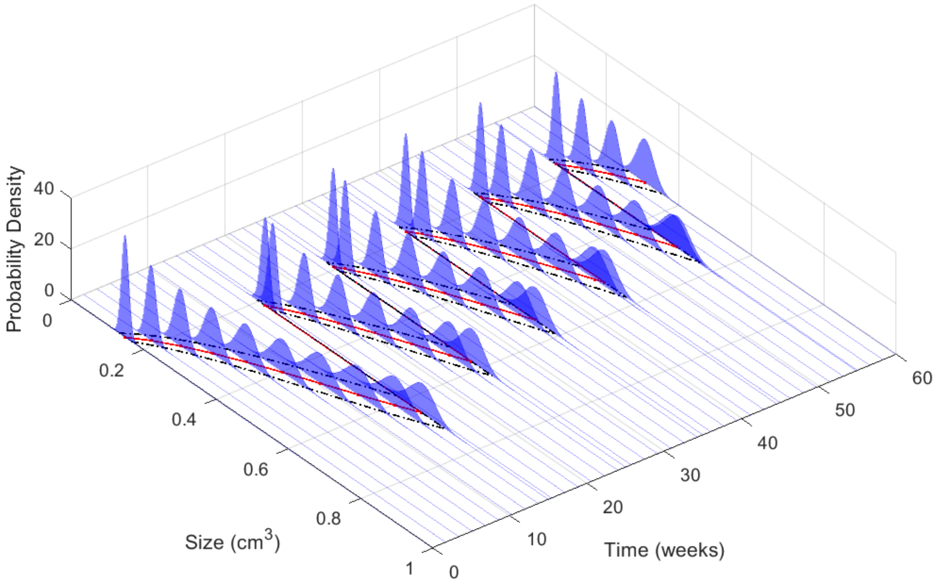


Figure 2: Full view of the PDF evolution simulations at the corresponding time values in T_{Tumor}, together with the mean (red) and 95% confidence intervals (dashed, black). Compare with Figure 1.

of the first probability density function of the solution stochastic process by combining the Liouville equation and the Random Variable Transformation theorem. We have applied our general theoretical findings to a mathematical model emerging from the generalized logistic model with natural growth altered by impulsive terms acting contrarily to its natural dynamics.

As a final note, the work on which this presentation was based has been submitted to a scientific journal as a research article.

Acknowledgments

This work has been partially supported by the Spanish Ministerio de Economía, Industria y Competitividad (MINECO), the Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional (FEDER UE) grant PID2020–115270GB–I00, the Generalitat Valenciana (grant AICO/2021/302) and by el Fondo Social Europeo y la Iniciativa de Empleo Juvenil EDGJID/2021/185. Vicente Bevia acknowledges the doctorate scholarship granted by Programa de Ayudas de Investigación y Desarrollo (PAID), Universitat Politècnica de València.

References

[1] V. Bevia, C. Burgos, J.C. Cortés, A. Navarro, and R.J. Villanueva. Analysing Differential Equations with Uncertainties via the Liouville-Gibbs Theorem: Theory and Applications, pages 1–23. Springer Singapore, Singapore, 2020. ISBN 978-981-15-8498-5. doi: 10.1007/978-981-15-8498-5 1. URL https://doi.org/10.1007/978-981-15-8498-5_1.

[2] V. Bevia, C. Burgos, J.C. Cortés, A. Navarro-Quiles, and R.-J. Villanueva. Uncertainty quantification analysis of the biological Gompertz model subject to random fluctuations in all its parameters. Chaos, Solitons & Fractals, 138:109908, 2020. doi: https://doi.org/10.1016/j.chaos.2020.109908.

[3] P. Dyke. An Introduction to Laplace Transforms and Fourier Series. Springer London, 2014. ISBN 978-1-4471-6394-7.

[4] C. Gasquet and P. Witomski. Fourier Analysis and Applications. Filtering, Numerical Computation, Wavelets. Springer, 1998.

[5] N. Hritonenko and Yu. Yatsenko. Bang-bang, impulse, and sustainable harvesting in age-structured populations. Journal of Biological Systems, 20(02):133–153, 2012. doi: 10.1142/S0218339012500088.

[6] M. Loève. Probability Theory I. Springer New York, NY, New York, 1977. ISBN 978-0-387-90210-4.

[7] Y. Ram, E. Dellus-Gur, M. Bibi, K. Karkare, U. Obolski, M.W. Feldman, T.F. Cooper, J. Berman, and L. Hadany. Predicting microbial growth in a mixed culture from growth curve data. Proceedings of the National Academy of Sciences, 116(29):14698–14707, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1902217116. URL https://www.pnas.org/content/116/29/14698.

[8] T. Soong. Random Differential Equations in Science and Engineering. Academic Press, New York, 1973. ISBN 9780126548501.

[9] X.Huang and B. Yang. Improving energy harvesting from impulsive excitations by a nonlinear tunable bistable energy harvester. Mechanical Systems and Signal Processing, 158:107797, 2021. ISSN 0888-3270. doi: 10.1016/j.ymssp.2021.107797.

[10] X. Zhang, Z. Shuai, and K. Wang. Optimal impulsive harvesting policy for single population. Nonlinear Analysis: Real World Applications, 4(4):639–651, 2003. ISSN 1468-1218. doi: https://doi.org/10.1016/S1468-1218(02)00084-6.

6 Combined and reduced combined matrices

R. Bru [♭, 1] M.T. Gassó[♭, 2] and M. Santana[♮, 3]

(♭) Institut Universitari de Matemàtica Multidisciplinària,
Universitat Politècnica de València,
Camí de Vera s/n, València, Spain.
(♮) Instituto de Matemática,
Universidad Autonóma de Santo Domingo,
Ave. Alma Matér, Santo Domingo, Dominican Republic.

1 Introduction

The combined matrix of a nonsingular matrix M is defined as $\mathcal{C}(M) = M \circ M^{-T}$ where the symbol $\circ$ denotes the Hadamard product of two matrices. Those matrices are known as the relative gain array in Mathematical Control Theory (see [1]) as it is pointed out by Johnson and Shapiro in [10], where some mathematical problems of those matrices are studied. Properties of combined matrices can be seen in [4] and [8] and [2].

The basic properties of combined matrices are

- The sum of the entries of any column and any row of a combined matrix is 1.
- Note that if the combined matrix is nonnegative then it is double stochastic.
- $\mathcal{C}(M)$ preserves the zero entries of M.
- $\mathcal{C}(MD) = \mathcal{C}(DM) = \mathcal{C}(M)$ for any nonsingular diagonal matrix D.
- $\mathcal{C}(PMQ) = P\mathcal{C}(M)Q$, where P and Q are two permutation matrices.
- $\mathcal{C}(M^T) = \mathcal{C}(M^{-1})$.

The name *combined matrix* seems to be used by the first time by Fiedler in [4]. Before this paper, the name Relative Gain Array (RGA) was used either in Control Theory and in the correponding field of Mathematics.

In this work, we explain some applications of combined matrices and introduce the reduced combined matrix, which is a special combined matrix in such a way that the sums of its columns are lower bounds of the condition number of the matrix. In particular, we give properties of the reduced combined matrix and given the relationship with both combined matrices. Matrices for which the lower bound is sharp are also given.

[1] rbru@imm.upv.es

[2] mgasso@mat.upv.es

[3] msantana22@uasd.edu.do

2 Applications

Below are the main applications of combined matrices: (i) Computing double stochastic matrices, (ii) Mapping eigenvalues with diagonal entries, (iii) Control systems, (iv) Bounding the condition number $\kappa(A)$. We are going to discuss these applications mainly those related to control theory and bounding the condition number. The authors of [6] give some algorithms to compute double stochastic matrices, mainly based on some upper Hessenberg matrices and from orthogonal (unitary) matrices. In both cases, double stochastic matrices are obtained.

It is known (see [10]) the following relation.

- Given a diagonalizable matrix $B = [b_{ij}]$ of size $n \times n$
- with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$.
- Then, there exists an invertible matrix A such that

$$B = A\,\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)A^{-1},$$

  and therefore,

$$\begin{bmatrix} b_{11} \\ b_{22} \\ \vdots \\ b_{nn} \end{bmatrix} = \mathcal{C}(A) \begin{bmatrix} \lambda_{11} \\ \lambda_{22} \\ \vdots \\ \lambda_{nn} \end{bmatrix}.$$

  That is, the image of the eigenvalues of the diagonalizable matrix B by the combined matrix A is the diagonal entries of B.

- Two more applications are explained in the sections below.

Example 1. *Given the matrix*

$$M = \begin{bmatrix} 2.4500 & -1.5200 & 0.4500 \\ -1.3700 & 5.8300 & -3.5400 \\ -2.3500 & 0.6700 & 4.4500 \end{bmatrix}$$

its combined matrix is

$$C(M) = \begin{bmatrix} 1.3037 & -0.4118 & 0.1081 \\ -0.1819 & 1.3103 & -0.1284 \\ -0.1218 & 0.1014 & 1.0203 \end{bmatrix}.$$

Note that the sums of its columns and rows are 1.

3 Relative gain array

Suppose we have a MultiInput MultiOutput control system (MIMO) with

- inputs $u_1, u_2, \ldots, u_n$.
- outputs $y_1, y_2, \ldots, y_n$.
- controllers $c_1, c_2, \ldots, c_n$.

Using the controllers c_i, the question is: How to assign (loop pairing) outputs (y_i) with inputs (u_j)?

A MIMO system is given by its transfer matrix $G(s)$, which is a rational matrix. The Relative Gain Array (RGA) is the combined matrix of the transfer matrix evaluated at 0, that is, $G(0) = [g_{ij}]$. It turns out that of $G(0)$ gives information about pairing the output j with input i (see [11]):

- If we pair output j with input i and $g_{ij} < 0$ then the plant will not be stable.
- It is known that $g_{ij} > 0$ is a necessary condition for stability (see [9]) and $g_{ij} > 0$ should be close to 1.

Actually, we have the following result.

Theorem 3. *Let G the transfer matrix of a given multiinput output system (MIMO). Let $\mathcal{C}(G)$ its combined matrix. A necessary condition so that the MIMO system be stable is that the $\mathcal{C}(G)$ has a positive monomial submatrix.*

Then, we have to find a monomial submatrix of order 4 in the RGA matrix to assign outputs j with inputs i. Let us see the following example.

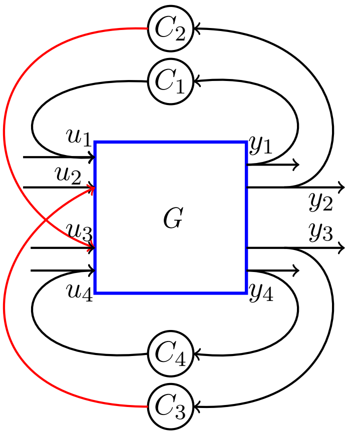
Example 2. *Suppose we have the following transfer matrix*

$$G = \begin{bmatrix} 1.4000 & 0.9000 & 0.2860 & 0.1000 \\ 1.4000 & 1.0000 & 0.4400 & 0.4000 \\ 0.2200 & 0.2640 & 0.0242 & -0.0660 \\ 0.2000 & 0.3000 & 0.0880 & 0.5000 \end{bmatrix}.$$

Then its relative gain array, that is, its combined matrix is

$$C(G) = (RGA) = \begin{bmatrix} 4.6455 & -1.6395 & -2.1126 & 0.1066 \\ -2.8095 & 0.6803 & 3.5374 & -0.4082 \\ -0.9146 & 1.7506 & 0.0076 & 0.1565 \\ 0.0786 & 0.2086 & -0.4324 & 1.1451 \end{bmatrix}.$$

To loop pairing outputs with inputs we have to extract a monomial submatrix of order 4 from the matrix RGA. One possible monomial matrix is one whose nonzero entries are the (1,1), (2,3), (3,2) and (4,4) entries of RGA, Then, the loop pairing can be seen in the following figure.



The loop pairing is output 1 with input 1, Output 2 with input 3, output 3 with input 2, and output 4 with input 4.

4 Combined and reduced matrices and conditioning

Let $M = [m_{ij}]$ be an $n \times n$ nonsingular matrix and denote the ith column by m_i. Consider the factorization

$$M = AF, \tag{1}$$

where the matrix $A = [a_1, a_2, \ldots, a_n]$ with the ith column of A given by $a_i = \dfrac{m_i}{||m_i||_2}$ and the diagonal matrix $F = \text{diag}(||m_1||_2, ||m_2||_2, \ldots, ||m_n||_2)$.

Let $P = [p_{ij}]$ be the inverse of M and denote by p^i the ith row vector of P. Consider the factorization

$$P = M^{-1} = GB. \tag{2}$$

Denoting the rows of a matrix with super indices, we have that $b^i = \dfrac{p^i}{||p^i||_2}$ and the diagonal matrix $G = \text{diag}\left(||p^1||_2, ||p^2||_2, \ldots, ||p^n||_2\right)$.

Now, let us see how we can compute the minimal projection of an invertible matrix by its combined matrix (see [5] and [8]). First, we recall the definitions of combined and reduced combined matrices given in [3].

Definition 1. *Let M be an invertible matrix.*
(i) The matrix

$$\mathcal{C}(M) = M \circ \left(M^{-1}\right)^T$$

is called the ***combined matrix*** *of M, where $\circ$ is the Hadamard product.*
(ii) Let A be the matrix obtained by normalizing the columns of matrix M. The matrix

$$\mathcal{R}(M) = A \circ B^T$$

is said to be the ***reduced combined matrix*** *of M, where B is M^{-1} with normalized rows.*

The results listed below are given in the work [3].

The reduced combined matrix of M^T and M^{-1} has the same relationship as the combined matrix of both matrices, as we see in the following result.

Theorem 4. *Let M be an invertible matrix. Then*

$$\mathcal{R}(M^T) = \mathcal{R}(M^{-1}),$$

where $\mathcal{R}(M^T)$ and $\mathcal{R}(M^{-1})$ are the reduced combined matrices of M^T and M^{-1}, respectively.

The relationship between the combined matrix and the reduced combined matrix is the following.

Theorem 5. *Let M be an invertible matrix. Let $\mathcal{C}(M)$ and $\mathcal{R}(M)$ be the combined and the reduced combined matrix of M, respectively. Then*

$$\mathcal{C}(M) = \mathcal{R}(M)FG, \tag{3}$$

where F and G are the matrices of the factorizations (1) *and* (6)*, respectively.*

Theorem 6. *Suppose the combined and the reduced combined matrix of a matrix M are equal. Then, M has its columns orthogonal.*

Theorem 7. *Let M be a unitary matrix of order n. Then its combined matrix is equal to its reduced combined matrix, that is, $\mathcal{C}(M) = \mathcal{R}(M)$.*

We are going to see that with the reduced combined matrix we can obtain a lower bound of the condition number of M.

Let M be an invertible matrix.The condition number of M is

$$\kappa(M) = ||M||||M^{-1}||,$$

where $||\cdot||$ denotes a matrix norm. Usually, the spectral norm of M is used, that is, $\kappa(M) = ||M||_2||M^{-1}||_2$, where $||M||_2 = \sigma_1$. Here, σ_1 denotes the maximum singular value of M.

Theorem 8. *Let M be an invertible matrix of order n. Let $r_j = \sum_{i=1}^{n} r_{ij}$, $\quad 1 \leq j \leq n$, the sum of the column entries of the reduced combined matrix $\mathcal{R}(M) = [r_{ij}]$. Let $r(M) = \min\limits_{1\leq i\leq n} |r_i|$, Then*

$$1 \leq \frac{1}{r(M)} \leq \kappa(M).$$

The sum of the column entries of the reduced combined matrix $\mathcal{R}(M)$ is an upper bound of the condition number of M.

Example 3. *Consider the invertible matrix M of the first example. The reduced combined matrix is*

$$\mathcal{R}(M) = A \circ B^T = \begin{bmatrix} 0.6465 & -0.1773 & 0.0567 \\ -0.0902 & 0.5641 & -0.0674 \\ -0.0604 & 0.0437 & 0.5355 \end{bmatrix}.$$

- *The column sums of the matrix $\mathcal{R}(M)$ are $r_1 = 0.4959, r_2 = 0.4305$, $r_3 = 0.5248$.*
- *The minimal column sum is $r(M) = 0.4305$.*
- *The lower bound of $\kappa(M) = 5.2436$ is $\dfrac{1}{r(M)} = 2.3228$.*

5 Acknowledgement

This work was partially supported by Dominican Republic FONDOCYT grant number 2021–2022–1D1–121 and by GVA grant CIAICO/2021/162 of Generalitat Valenciana, Spain.

References

[1] Bristol E., On a new measure of interaction for multivariable process control. *IEEE Trans. Autom. Control*, 1:13–134, 1966.

[2] Bru R., Gassó M.T., Giménez I., Santana M., Nonnegative combined matrices. *J. Appl. Math.* vol. 2014, ID 182354, 5 pages, 2014.

[3] Bru R., Gassó M.T., Santana M., Combined matrices and conditioning. *Applied Mathematics and Computation*, 412, ID 126549 (8 pages), 2022.

[4] Fiedler M., Notes on Hilbert and Cauchy matrices. *Linear Algebra Appl.*, 432:351–356, 2010.

[5] Fiedler M., Markham T. L., Combined matrices in special classes of matrices. *Linear Algebra Appl.*, 435:1945–1955, 2011.

[6] Fuster R., Gassó M.T., Giménez I., CMMSE algorithms for constructing doubly stochastic matrices with the relative gain array (combined matrix) $A \circ A^{-T}$. *Journal of Mathematical Chemistry*, 57:1700–1709, 2019.

[7] 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 (Antofagasta)*, 40(1):255–271, 2021.

[8] Horn R. A., Johnson C. R., Topics in Matrix Analysis. Cambridge Univ. Press, 1991.

[9] Hovd M., Skogestad S., Sequencial design of decentralized controllers. *Automatica*, 30(10):1601–1607, 1994.

[10] Johnson C., Shapiro H., Mathematical aspects of the relative gain array. SIAM J. Algebraic Discrete Methods, 7(4):627–644, 1986.

[11] McAvoy T.J., Interaction Analysis: Principles and Applications. Pittsburgh, Instrument Society of America, 1983.

[12] Wilkinson J. H., The Algebraic Eigenvalue Problem. Oxford, Clarendon Press, 1965.

7

Random Fractional Hermite Differential Equation: A full study un mean square sense

C. Burgos $^{\flat}$,[1] T. Caraballo $^{\natural}$ J. C. Cortés$^{\flat}$ and R. J. Villanueva$^{\flat}$

($\flat$) Instituto Universitario de Matemática Multidisciplinar. Universitat Politècnica de València,
Camino de vera, s/n. 46022, Valencia.

($\natural$) Departamento de Ecuaciones Diferenciales y Análisis Numérico. Universidad de Sevilla,
Avenida Reina Mercedes, s/n. 41012-Sevilla.

Abstract

In this contribution a full probabilistic study for the Random Fractional Hermite differential equation is performed. Firstly, applying the random fractional Fröbenius method we will construct a solution convergent in mean square sense. Then, we will obtain reliable approximations for the mean and for the standard deviation taking into account that the solution described by a power series converges in mean square sense. After that, we will go a step further computing first probability density function of the solution. Finally, we show one numerical example to illustrate the theoretical findings.

1 Introduction

Hermite differential equation is applied in quantum physics to obtain the solution of problems which describes the dynamics of particles, as atoms protons and neutrons. An example of it is the quantum harmonic oscillator which is defined by the Schrödinger differential equation. Nevertheless, when we describe the dynamics of quantum particles, there exist non-local and memory effects which are not able to describe by classical derivatives. In the recent years, it has been observed that fractional operators can properly address this complex dynamics, [1]. In this work the classical derivative in the Hermite differential equation is reemplaced by the Caputo fractional operator.

It is also worthly mentioned that the parameters of the differential equations, forcing term, initial conditions, etc. include uncertainties which have to take into account to model accurately a real phenomena. In this contribution, besides including a fractional operator in the Hermite differential equation, we will consider that the parameters of the equation are random variables instead deterministic values.

In this work we will deal with the following Initial Value Problem (IVP)

$$({}^{C}D_0^{2\alpha}Y)(t) - 2t^{\alpha}({}^{C}D_0^{\alpha}Y) + \lambda Y(t) = 0, \quad Y(0) = Y_0, \quad Y'(0) = Y_1, \qquad (1)$$

where λ, Y_0 and Y_1 are random variables and $({}^{C}D_0^{\alpha}Y)$ is the Caputo derivative of order $\alpha \in]0,1]$ of the stochastic process $Y(t)$ defined in mean square (m.s). The operator $({}^{C}D_0^{2\alpha}Y)$ is given by [2]

$$({}^{C}D_0^{2\alpha}Y) :={}^{C} D_0^{\alpha}({}^{C}D_0^{\alpha}Y(t))$$

[1] clabursi@posgrado.upv.es

This contribution is organized as follows: Section 2 is devoted to construct a m.s. convergent solution of the IVP (1). In Section 3, we will obtain approximations for the mean and for the standard deviation of the solution. Then, in Section 4, we go an step further and applying the Random Variable Transformation Technique [3], we will obtain approximations for the first probability density function (1-PDF) of the solution Finally, the last Section, Section 5, is devoted to illustrate the theoretical findings with a numerical example.

2 Constructing mean square convergent solutions

This section is devoted to apply the random fractional Fröbenius method to compute the solution stochastic process of the RFIVP (1).

The fractional Fröbenius method allows us to obtain a solution in terms of a generalized power series,

$$Y(t)=\sum_{m=0}^{\infty} Y_m t^{\alpha m}. \tag{2}$$

Computing the coresponding Caputo fractional derivatives, $({}^{C}D_0^{\alpha})$ and $({}^{C}D_0^{2\alpha})$, of $Y(t)$, and substituting these expressions in the IVP (1), we can obtain the coefficients Y_m. Thus, the solution is given by

$$\begin{aligned} Y(t) =& Y_0\left(1+\sum_{m=1}^{\infty}\left[\frac{t^{2m\alpha}}{\Gamma(2\alpha m+1)}\left(\sum_{i=0}^{m-1}\lambda^{i+1}(-1)^{i+1}G_{m-1,i}\right)\right]\right) \\ &+Y_1\left(t^{\alpha}+\sum_{m=1}^{\infty}\left[\frac{\Gamma(\alpha+1)t^{(2m+1)\alpha}}{\Gamma((2m+1)\alpha+1)}\left(\sum_{i=0}^{m}\lambda^{i}(-1)^{i}\hat{G}_{m,i}\right)\right]\right), \end{aligned} \tag{3}$$

where

$$G_{m,i}=\begin{cases} \displaystyle\sum_{j_1<j_2<\ldots<j_{m-i}} 2\frac{\Gamma(2j_1\alpha+1)}{\Gamma((2j_1-1)\alpha+1)}\cdot 2\frac{\Gamma(2j_2\alpha+1)}{\Gamma((2j_2-1)\alpha+1)}\cdot\ldots\cdot 2\frac{\Gamma(2j_{m-i}\alpha+1)}{\Gamma((2j_{m-i}-1)\alpha+1)}, & \text{if } i<m,\\ 1, & \text{if } m=i,\\ 0, & \text{otherwise,}\end{cases} \tag{4}$$

and

$$\hat{G}_{m,i}=\begin{cases} \displaystyle\sum_{j_1<j_2<\ldots<j_{m-i}} 2\frac{\Gamma((2j_1-1)\alpha+1)}{\Gamma((2j_1-2)\alpha+1)}\cdot 2\frac{\Gamma((2j_2-1)\alpha+1)}{\Gamma((2j_2-2)\alpha+1)}\ldots 2\frac{\Gamma((2j_{m-i}-1)\alpha+1)}{\Gamma((2j_{m-i}-2)\alpha+1)}, & \text{if } i<m,\\ 1, & \text{if } m=i,\\ 0, & \text{otherwise.}\end{cases} \tag{5}$$

As it can be observed, the solution is described by a generalized random power series. Assuming the following hypotheses, we can guarantee that (3) is convergent in mean square sense.

- **H1:** The inputs parameters Y_0, Y_1 and λ are independent random variables.
- **H2:** The random variable λ is bounded random variable, i.e. exist b_1 and b_2 finite such that $b_1<\lambda(\omega)<b_2$ for each $\omega\in\Omega$.

3 Approximating the mean and the standard deviation

Once a mean square convergent solution for the IVP (1) is obtained we compute reliable approximations for the mean and for the standard deviation of the solution.

To do it, it is important to take into account that mean square convergence of the random series (3)-(5) guarantees the convergence of the approximations of the mean and the standard deviation by truncating at a specific order, say $M > 0$ integer, the series solution, [3, Th. 4.3.1]. This leads to the following approximations for the mean and for the second order moment

$$\begin{aligned}\mathbb{E}[Y_M(t)] =& \mathbb{E}[Y_0]\left(1+\sum_{m=1}^{M}\left[\frac{t^{2m\alpha}}{\Gamma(2\alpha m+1)}\left(\sum_{i=0}^{m-1}\mathbb{E}[\lambda^{i+1}](-1)^{i+1}G_{m-1,i}\right)\right]\right)\\ &+\mathbb{E}[Y_1]\left(t^{\alpha}+\sum_{m=1}^{M}\left[\frac{\Gamma(\alpha+1)t^{(2m+1)\alpha}}{\Gamma((2m+1)\alpha+1)}\left(\sum_{i=0}^{m}\mathbb{E}[\lambda^{i}](-1)^{i}\hat{G}_{m,i}\right)\right]\right)\end{aligned} \tag{6}$$

and

$$\begin{aligned}\mathbb{E}[Y_M^2(t)] =& \mathbb{E}[Y_0^2]\Bigg(1+2\sum_{m=1}^{M}\left[\frac{t^{2\alpha m}}{\Gamma(2\alpha m+1)}\left(\sum_{i=0}^{m-1}\mathbb{E}[\lambda^{i+1}](-1)^{i+1}G_{m-1,i}\right)\right]\\ &+\left(\sum_{m=1}^{M}\sum_{n=1}^{M}\left[\frac{t^{2\alpha(m+n)}}{\Gamma(2\alpha m+1)\Gamma(2\alpha n+1)}\left(\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\mathbb{E}[\lambda^{i+j+2}](-1)^{i+j+2}G_{m-1,i}G_{n-1,j}\right)\right]\right)\Bigg)\\ &+\mathbb{E}[Y_1^2]\Bigg(t^{\alpha}+2t^{\alpha}\sum_{m=1}^{M}\left[\frac{\Gamma(\alpha+1)t^{(2m+1)\alpha}}{\Gamma((2m+1)\alpha+1)}\left(\sum_{i=0}^{m}\mathbb{E}[\lambda^{i}](-1)^{i}\hat{G}_{m,i}\right)\right]\\ &+\sum_{m=1}^{M}\sum_{n=1}^{M}\left[\frac{\Gamma(\alpha+1)^2t^{(2n+2m+2)\alpha}}{\Gamma((2m+1)\alpha+1)\Gamma((2n+1)\alpha+1)}\left(\sum_{i=0}^{m}\sum_{j=0}^{n}\mathbb{E}[\lambda^{i+j}](-1)^{i+j}\hat{G}_{m,i}\hat{G}_{n,i}\right)\right]\Bigg)\\ &+2\mathbb{E}[Y_0]\mathbb{E}[Y_1]\Bigg(t^{\alpha}+t^{\alpha}\sum_{m=1}^{M}\left[\frac{t^{2\alpha m}}{\Gamma(2\alpha m+1)}\left(\sum_{i=0}^{m-1}\mathbb{E}[\lambda^{i+1}](-1)^{i+1}G_{m-1,i}\right)\right]\\ &+\sum_{m=1}^{M}\left[\frac{\Gamma(\alpha+1)t^{(2m+1)\alpha}}{\Gamma((2m+1)\alpha+1)}\left(\sum_{i=0}^{m}\mathbb{E}[\lambda^{i}](-1)^{i}\hat{G}_{m,i}\right)\right]\\ &+\sum_{m=1}^{M}\sum_{n=1}^{M}\left[\frac{\Gamma(\alpha+1)t^{2\alpha m}t^{(2n+1)\alpha}}{\Gamma((2n+1)\alpha+1)\Gamma(2\alpha m+1)}\left(\sum_{i=0}^{m-1}\sum_{j=0}^{n}\mathbb{E}[\lambda^{i+j+1}](-1)^{i+j+1}G_{m-1,i}\hat{G}_{n,j}\right)\right]\Bigg),\end{aligned} \tag{7}$$

respectively.

Taking into account (6) and, (7) and $\sigma[Y_M(t)] = \sqrt{E[Y_M(t)^2] - E[Y_M(t)]^2}$, we can compute the standard deviation.

4 Computing approximations for the 1-PDF of the solution SP

So far, approximations for the two first statistical moments have been obtained. Nevertheless, sometimes higher one-dimensional moments are also needed. The 1-PDF of the solution allows us to compute these higher moments. Applying the Random Variable Transformation technique [3, Ch. 2] to the truncated solution of order M we can construct approximations for the 1-PDF. After some involved computations, one gets

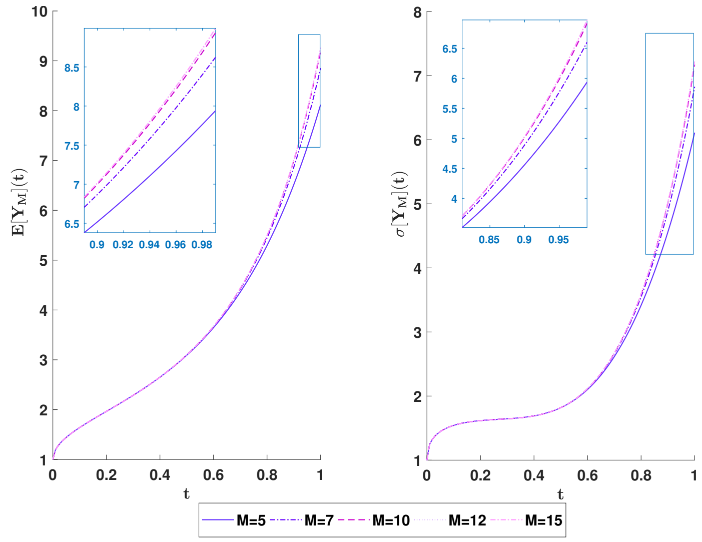


Figure 1: Mean and standard deviation of the solution for different orders of truncation M.

$$f_{Y_M(t)}(y) = \int_{\mathcal{D}(Y_1,\lambda)} f_{Y_0}\left(\frac{y - y_1\left(t^{\alpha} + \sum\limits_{m=1}^{M}\left[\frac{\Gamma(\alpha+1)t^{(2m+1)\alpha}}{\Gamma((2m+1)\alpha+1)}\left(\sum\limits_{i=0}^{m}\lambda^{i}(-1)^{i}\hat{G}_{m,i}\right)\right]\right)}{1+\sum\limits_{m=1}^{M}\left[\frac{t^{2\alpha m}}{\Gamma(2\alpha m+1)}\left(\sum\limits_{i=0}^{m}\lambda^{i+1}(-1)^{i+1}G_{m-1,i}\right)\right]}\right) f_{Y_1}(y_1) f_{\lambda}(\lambda)$$
$$\cdot \frac{1}{\left|1+\sum\limits_{m=1}^{M}\left[\frac{t^{2\alpha m}}{\Gamma(1\alpha m+1)}\left(\sum\limits_{i=0}^{m-1}\lambda^{i+1}(-1)^{i+1}G_{m-1,i}\right)\right]\right|}\mathrm{d}y_1\mathrm{d}\lambda. \tag{8}$$

Assuming that the PDF of random variable Y_0, f_{Y_0}, is Lipschitz, it can be rigorously proved that (8) converges to the 1-PDF of the solution stochastic process, $Y(t)$, as $M \to \infty$.

5 Numerical Examples

This section is addressed to numerically illustrate the previous theoretical results. We will take $\alpha = 0.5$. We will assume that, $\lambda \sim Be(2,3)$, where $Be(2,3)$ denotes the beta distribution of parameters $(2,3)$; $Y_0 \sim Ga(1,1)$ and $Y_1 \sim N(2,\sqrt{2}^2)$, where $Ga(1,1)$ stands for the Gamma distribution of parameters $(1,1)$ and $N(2,\sqrt{2}^2)$ Gaussian distribution of parameters $(2,\sqrt{2}^2)$.

Figure 1 shows the mean and the standard deviation of the solution for different orders of truncation $M \in \{5,7,10,12,15\}$. We can observe in the zoomed area the convergence as M increases.

The 1-PDF has been plotted in Figure 2 considering different orders of truncation M. Each subplot corresponds to different $t \in \{0.25, 0.5, 0.75\}$. To verify the convergence as the order of truncation, M, increases, a zoom has been performed in each subplot.

References

[1] K. Gökhan-Atman, H. Şirin, *Nonlocal Phenomena in Quantum Mechanics with Fractional Calculus*, Reports on Mathematical Physics, 86, 2, 2020, 263-270, https://doi.org/10.1016/S0034-4877(20)30075-6

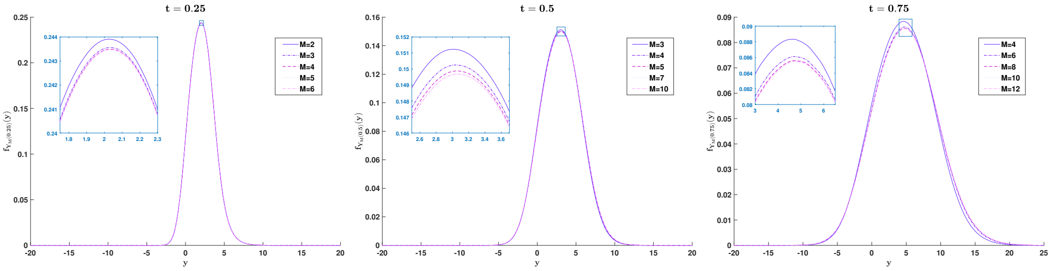


Figure 2: 1-PDF of the solution, (8), for different $t \in \{0.25, 0.5, 0.75\}$.

[2] C. Burgos, J.C. Cortés, L. Villafuerte, R. J. Villanueva. *Extending the deterministic Riemann Liouville and Caputo operators to the random framework: A mean square approach with applications to solve random fractional differential equations.* Chaos, Solitons Fractals. Elsevier. 2017, 102, 305-318.

[3] T. T. Soong. *Random differential equations in science and engineering.* Academic Press. 1973. ISBN: 9780080956121

[4] M. Al-Refai, M. Syam, Q. Al-Mdallal. (2017). *On the fractional Legendre equation and fractional Legendre functions.* Progr Fract Differ Appl. 2017 3(2), 93-102.

8

Mathematical modelling of frailty and dependency in basic activities of daily living in general population aged 70 years

S. Camacho Torregrosa[♭,1] C. Santamaría Navarro[♭] and X. Albert Ros[♮]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(♮) Centro de Salud de Moncada
Conselleria de Sanitat
Avenida Mediterráneo 2, Moncada, Valencia

1 Introduction

Frailty is a clinical state in a person that increases the vulnerability to develop adverse outcomes such as dependency and mortality when it is exposed to a stressor [Caballero-Mora(2017)]. Frailty is a predictor of mortality, disability, mobility loss, institutionalization, falls and cardiovascular disease [Rizos and Soler(2013)].

There are two conceptual definitions of frailty: physical frailty [Fried et al.(2001)]based on Fried's criteria, and frailty based on the deficit accumulation [Rockwood(2005)] (includes social assessment, comorbidity, among other elements). The frailty indices are derived from the deficit accumulation definition. Frail-VIG index [Amblàs-Novellas et al.(2017)] is a frailty index based on the comprehensive geriatric assessment (CGA, VIG in Spanish). The CGA is considered the gold standard for management of frailty in elderly people. This index is composed with variables recorded during the usual clinical evaluation process. The Frail-VIG Index establishes a situational diagnosis of the severity of the patient's frailty.

There is a relationship between frailty, disability and comorbidity, but there are not the same concepts. Between 23% and 26% of the frail elderly people do not have disability or comorbidity [Alcalá et al.(2010)]. Furthermore, the physical frailty and the two stages of disabilities (disability in instrumental and basic activities of daily living) follow a hierarchy along a continuum [Zamudio-Rodríguez et al.(2020)].

There are multiple factors associated with frailty and dependency. However, there are few studies that predict who will be frail or dependent in clinical settings. In a systematic review published in 2022 [Grootven and van Achterberg(2022)], it was concluded that no revised predictive model of functional status can be recommended for implementation in practice. Most of the studies aimed to predict frailty are based in the Fried criteria, such as the model made by Vieira de Sousa [de Sousa et al.(2018)], or the nomogram developed by Bing-Ru [Dong et al.(2021)]. Other models predicted frailty conditions (no frailty), like mortality, urgent hospitalization, disability, fracture or emergency admission [Tarekegn et al.(2020)].

[1] sucatoval@gmail.com

Our aim is to develop predictive models to know the probability of being frail according to the Frail-VIG Index or dependent for basic activities of daily living (using the Barthel Index) at a given age, according to clinical and sociodemographic variables.

2 Methods

This is an observational population-based study, in people aged 70 year or over from two lists of general practitioners of the Moncada's Primary Health Care. Sociodemographic data were collected: gender, marital status, educational level and number of cohabitants. Dependence was measured with the Barthel Index, and frailty with the Frail-VIG Index [Amblàs-Novellas et al.(2017)] and the Short Physical Performance Battery (SPPB) [Guralnik et al.(1994)]. We requested the chronicity level to the Conselleria de Sanitat of Valencia, which classifies the chronicity in four categories: level 0, level 1, level 2 and level 3, meaning healthy people, chronic patients with low, moderate and high complexity, respectively.

The Barthel Index is a questionnaire that measures performance in activities of daily living. It was categorized in the next groups, based in the Shah criteria [Shah et al.(1989)]: no dependent (IB$\geq$95), mild dependency (IB 90-65), moderate-severe dependency (IB 25-60), and absolute dependency (IB$\leq$20). Frail-VIG Index is a questionnaire with 22 items that assess 25 deficits classified in eight domains: functional, nutritional, cognitive, emotional, social, geriatric syndromes, severe symptoms and diseases. The range goes from 0 to 1, and categorizes frailty in: no frail (<0.20), mild frailty (0.20-0.35), moderate frailty ($0.36 - 0.50$) and severe frailty (>0.50). The SPPB [Guralnik et al.(1994)] assess the lower extremity physical performance status. It consist in three test: test of standing balance (tandem, semi-tandem and side-by-side stands), walking speed test, and the ability to rise from a chair test. The range goes from 0 to 12. The cut of point we established was "seven", because it was the best cut-off point identificated in the Fradea study [Abizanda et al.(2012)] to define frailty. Frail-VIG index includes the Barthel Index, for that reason we calculated a new cut-off point of the Frail-VIG Index without the Barthel Index item.

This study was approved by the "Ethic Investigation Corporative Committee of Primary Health Care from the Comunitat Valenciana", in 01/31/2019, registration number 18/10. Participants or representatives were informed and signed the consentiment.

Data were anonymized, and the study complied with the data protection regulation "Organic Law 3/2018, of December 5, concerning protection of personal data and guarantee of digital rights".

2.1 Survival Analysis

The time from the age of 70 years old until the occurrence of one particular event is consider as the *survival time.* Survival technics are used to evaluate two different events or end-points:

1. Time until dependence measured with the Barthel Index (*time free of dependence*) .

2. Time until frailty measured with the Frail-VIG index (*time free of frailty*).

Two studies were performed taking into account these two end-points. The main objective in both studies was to establish prognostic factors for dependence and frailty in population over 70 years old. Kaplan-Meier estimator for the survival functions and the Log-Rank test for comparing survival functions of different data groups were used to establish the prognostic factors in a univariate analysis. A p-value obtained < 0.05 was consider significant. The Cox model was used as a multivariate analysis to explore the relationship between the survival time of a patient and the explanatory variables. The selected models were obtained using a variable selection strategy suggested in [Collett(2003)]. The resulting models made it possible to define risk groups according to the characteristics of the patients and to predict the probability of dependence or frailty at different ages (70, 85 and 95 years old).

Harrell's concordance index (C-index) [Harrell et al.(1982)] was used to asses models accuracy (discrimination). The C-index is the non-parametric version of the Area Under the Curve (AUC) that can take into account censored data. It ranges from $C = 0.5$ (random discrimination) to $C = 1$ (perfect discrimination). This area represents the probability that for two randomly chosen patients, the patient with characteristics of a worse prognosis has the event under study before.

Statistical analysis were done using R [Team(2010)]

3 Results

A total of 416 patients were considered into the dataset. Median age was 77 years old, 180 patients (43.3%) were male and 236 (56.7%) female. Other patients characteristics were collected for this study defining the following variables: *marital status* (with partner or others situations); *educational level* (without basic studies or with studies); *number of cohabitants* (with three levels: live alone, 2 cohabitants and 3 or more cohabitants) and *chronicity level* (with two levels: healthy or low complexity chronic disease and medium/high complexity chronic disease). Among the 416 patients, 111 individuals (26.7%) were classified as dependent by the Barthel index and 122 cases (29.3%) as frail person by the Frail-VIG index.

Univariate analysis established marital status, frail-VIG index and chronicity level as predictive variables of dependence and the variables marital status and chronicity level as predictive factors of frailty. The resulting Cox models for both endpoint (dependence and frailty) are shown in Table 1. Both models identified as prognostic factors to predict dependence and frailty: gender, marital status and chronicity level. The concordance index c obtained for both models were 0.69 for dependence model and 0.70 for frailty model.

Both fitted models establish that female patients have more risk to be dependent or frail along time respect male patients; patients without partner reduce the risk of dependency and frailty respect those patients with partner and finally, patients with a medium or severe chronic disease increase risk of dependency and frailty (more than 4-fold) respect patients healthy or with a low complexity chronic disease. These 3 variables code each one with two levels define 8 different characteristic groups (Table 2). The predicted probabilities of dependence and frailty for a patient belonging each group of characteristic at 70, 85 and 95 years old are shown in Table 3. Corresponding survival functions are depicted in Figure 1.

	Dependence		Frailty	
Covariates	HR	*p* value	HR	*p* value
Gender				
Male, Female	1.99 (1.27-3.11)	0.002	1.60 (1.06-2.41)	0.026
Marital status				
With partner, No partner	0.31 (0.19-0.49)	<0.001	0.33 (0.22-0.52)	<0.001
Chronicity level				
Healthy/Low level, Medium/Severe level	2.80 (1.74-4.50)	<0.001	4.53 (2.66-7.71)	<0.001

Table 1: Hazard Ratios (HR), 95% CI (in parenthesis) and *p* value for the probability of dependence and frailty. Estimates using Cox model.

Male	With partner	Healthy/Low Chronic level	Group 1
		Moderate/Severe Chronic level	Group 2
	Without partner	Healthy/Low Chronic level	Group 3
		Healthy/Low Chronic level	Group 4
Female	Without partner	Healthy/Low Chronic level	Group 5
		Moderate/Severe Chronic level	Group 6
	Without partner	Healthy/Low Chronic level	Group 7
		Healthy/Low Chronic level	Group 8

Table 2: Characteristic Groups.

		Dependence			Frailty	
Group	70 yr.	85 yr.	95 yr.	70 yr.	85 yr.	95 yr.
G1	0.99 (0.99-1.00)	0.76 (0.65-0.90)	0.04 (<0.001-0.47)	0.99 (0.99-1.00)	0.79 (0.67-0.91)	0.17 (0.04-0.65)
G2	0.98 (0.98-0.99)	0.47 (0.34-0.65)	<0.01 (<0.01-0.07)	0.98 (0.98-1.00)	0.33 (0.22, 0.50)	<0.01 (<0.01-0.05)
G3	0.99 (0.99-1.00)	0.92 (0.87-0.97)	0.37 (0.18-0.76)	0.99 (0.99-1.00)	0.99 (0.87-0.97)	0.55 (0.36-0.85)
G4	0.99 (0.99-1.00)	0.80 (0.70-0.90)	0.64 (0.01-0.38)	0.99 (0.99-1.00)	0.69 (0.58-0.83)	0.07 (0.01-0.33)
G5	0.99 (0.99-1.00)	0.59 (0.43-0.79)	<0.01 (<0.01-0.03)	0.99 (0.99-1.00)	0.68 (0.54-0.86)	0.06 (0.01-0.56)
G6	0.97 (0.94-0.99)	0.22 (0.11-0.43)	<0.01 (<0.01-0.02)	0.98 (0.96-1.00)	0.17 (0.09-0.35)	<0.01 (<0.01-0.02)
G7	0.99 (0.99-1.00)	0.85 (0.78-0.93)	0.14 (0.04, 0.51)	0.99 (0.99-1.00)	0.88 (0.82-0.95)	0.38 (0.20-0.73)
G8	0.99 (0.98-0.99)	0.63 (0.53-0.75)	<0.01 (<0.01-0.11)	0.99 (0.99-1.00)	0.53 (0.45-0.68)	0.01 (<0.01-0.14)

Table 3: Free of dependency probability and free of frailty probability at 70, 85 and 95 years old, and 95% CI (in parenthesis).

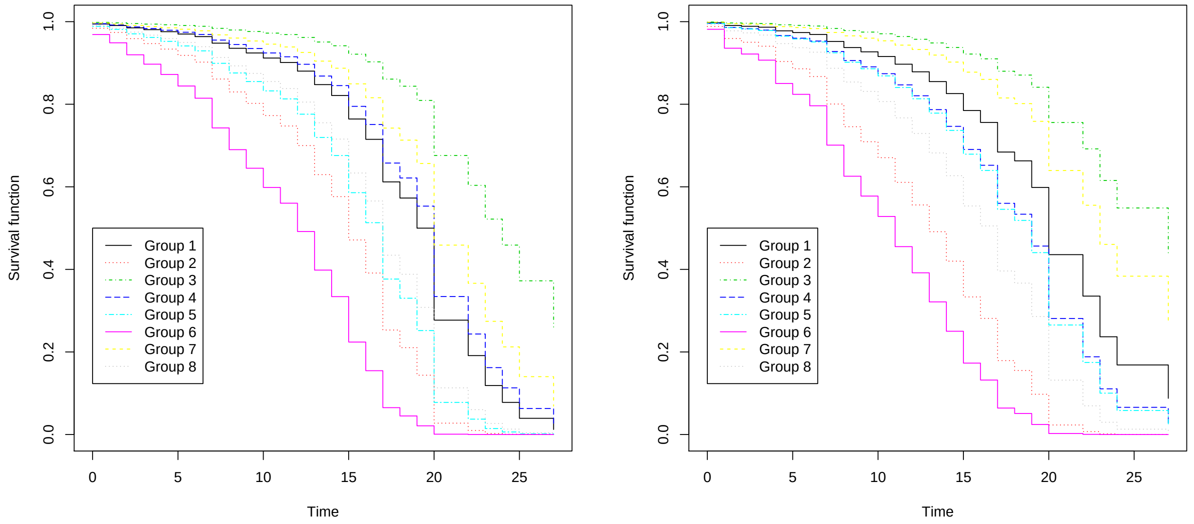


Figure 1: Survival functions (time free of dependency on left and time free of frailty on right) for the eight characteristic gropus.

4 Conclusions

We developed a model to predict dependency and another one to predict frailty in general population. Included variables in both models were gender, marital status and chronicity level. These variables are easy to collect in clinical practice except chronicity level, but usually the computerized medical history (ABUCASIS system in Comunitat Valenciana) shows it.

We found differences compared to previous studies on marital status: in our study be married is a risk factor for the dependency and frailty. In the Kogima study [Kojima et al.(2020)], be single implies twice the risk of frailty than married person. On the other hand, in the Trevisan study [Trevisan et al.(2016)], single man had more risk of frailty, and widow female had less risk than married female. Both studies valued frailty with Fried criteria, not with a frail index. This topic could be the subject of future studies.

Several studies and models [Grootven and van Achterberg(2022),Arnau et al.(2016)] concluded that the SPPB and gait of speed (included in the SPPB) could be variables to identify risk of disability. However, this variables were not significant in our study.

There are important differences between groups in the free time of frailty or dependency. Therefore, there are two points to consider in clinical practice: the modificable variables are the patient's diseases, assessed with the chronicity level or with Frail-VIG index; and risk groups must be identified in order to approach them.

Next step will be to study the effect of frailty (estimated with Frail-VIG Index and SPPB) and dependency in mortality in general population.

References

[Abizanda et al.(2012)] P. Abizanda, L. Romero, P.M. Sánchez-Jurado, P. Atienzar-Núñez, J.L. Esquinas-Requena, and I. Garcia-Nogueras. Association between functional assessment instruments and frailty in older adults: The fradea study. *Journal of Frailty & Aging*, pages 1–7, 2012. doi: 10.14283/jfa.2012.25.

[Alcalá et al.(2010)] María Victoria Castell Alcalá, Ángel Otero Puime, María Teresa Sánchez Santos, Araceli Garrido Barral, Juan Ignacio González Montalvo, and María Victoria Zunzunegui. Prevalencia de fragilidad en una población urbana de mayores de 65 años y su relación con comorbilidad y discapacidad. *Atención Primaria*, 42(10):520–527, 2010. doi: 10.1016/j.aprim.2009.09.024.

[Amblàs-Novellas et al.(2017)] Jordi Amblàs-Novellas, Joan Carles Martori, Núria Molist Brunet, Ramon Oller, Xavier Gómez-Batiste, and Joan Espaulella Panicot. Índice frágil-VIG: diseño y evaluación de un índice de fragilidad basado en la valoración integral geriátrica. *Revista Española de Geriatría y Gerontología*, 52(3):119–127, 2017. doi: 10.1016/j.regg.2016.09.003.

[Arnau et al.(2016)] Anna Arnau, Joan Espaulella, Marta Serrarols, Judit Canudas, Francesc Formiga, and Montserrat Ferrer. Risk factors for functional decline in a population aged 75 years and older without total dependence: A one-year follow-up. *Archives of Gerontology and Geriatrics*, 65:239–247, 2016. doi: 10.1016/j.archger.2016.04.002.

[Caballero-Mora(2017)] Ángeles Caballero-Mora. Knowing frailty at individual level: A systematic review wp leader: Ucsc work package: Wp4. 2017.

[Collett(2003)] D. Collett. *Modelling Survival Data in Medical Research.* Chapman-Hall/CRC, segunda edition, 2003.

[de Sousa et al.(2018)] Jacy Aurelia Vieira de Sousa, Maria Helena Lenardt, Clóris Regina Blanski Grden, Luciana Kusomota, Mara Solange Gomes Dellaroza, and Susanne Elero Betiolli. Physical frailty prediction model for the oldest old. *Revista Latino-Americana de Enfermagem*, 26(0), 2018. doi: 10.1590/1518-8345.2346.3023.

[Dong et al.(2021)] Bing-Ru Dong, Xiao-Qing Gu, Hai-Ying Chen, Jie Gu, and Zhi-Gang Pan. Development and validation of a nomogram to predict frailty progression in nonfrail chinese community-living older adults. *Journal of the American Medical Directors Association*, 22(12):2571–2578.e4, 2021. doi: 10.1016/j.jamda.2021.05.020.

[Fried et al.(2001)] L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W. J. Kop, G. Burke, and M. A. McBurnie. Frailty in older adults: Evidence for a phenotype. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(3): M146–M157, 2001. doi: 10.1093/gerona/56.3.m146.

[Grootven and van Achterberg(2022)] Bastiaan Van Grootven and Theo van Achterberg. Prediction models for functional status in community dwelling older adults: a systematic review. *BMC Geriatrics*, 22(1), 2022. doi: 10.1186/s12877-022-03156-7.

[Guralnik et al.(1994)] J. M. Guralnik, E. M. Simonsick, L. Ferrucci, R. J. Glynn, L. F. Berkman, D. G. Blazer, P. A. Scherr, and R. B. Wallace. A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission. *Journal of Gerontology*, 49(2):M85–M94, 1994. doi: 10.1093/geronj/49.2.m85.

[Harrell et al.(1982)] Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. doi: 10.1001/jama.1982.03320430047030.

[Kojima et al.(2020)] Gotaro Kojima, Kate Walters, Steve Iliffe, Yu Taniguchi, and Nanako Tamiya. Marital status and risk of physical frailty: A systematic review and meta-analysis. *Journal of the American Medical Directors Association*, 21(3):322–330, 2020. doi: 10.1016/j.jamda.2019.09.017. URL https://doi.org/10.1016/j.jamda.2019.09.017.

[Rizos and Soler(2013)] Luis Romero Rizos and Pedro Abizanda Soler. Fragilidad como predictor de episodios adversos en estudios epidemiológicos: revisión de la literatura. *Revista Española de Geriatría y Gerontología*, 48(6):285–289, 2013. doi: 10.1016/j.regg.2013.05.005.

[Rockwood(2005)] K. Rockwood. A global clinical measure of fitness and frailty in elderly people. *Canadian Medical Association Journal*, 173(5):489–495, 2005. doi: 10.1503/cmaj.050051.

[Shah et al.(1989)] Surya Shah, Frank Vanclay, and Betty Cooper. Improving the sensitivity of the barthel index for stroke rehabilitation. *Journal of Clinical Epidemiology*, 42(8):703–709, 1989. doi: 10.1016/0895-4356(89)90065-6.

[Tarekegn et al.(2020)] Adane Tarekegn, Fulvio Ricceri, Giuseppe Costa, Elisa Ferracin, and Mario Giacobini. Predictive modeling for frailty conditions in elderly people: Machine learning approaches. *JMIR Medical Informatics*, 8(6):e16678, 2020. doi: 10.2196/16678.

[Team(2010)] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. http://www.R-project.org.

[Trevisan et al.(2016)] Caterina Trevisan, Nicola Veronese, Stefania Maggi, Giovannella Baggio, Marina De Rui, Francesco Bolzetta, Sabina Zambon, Leonardo Sartori, Egle Perissinotto, Gaetano Crepaldi, Enzo Manzato, and Giuseppe Sergi. Marital status and frailty in older people: Gender differences in the Progetto veneto anziani longitudinal study. *Journal of Women's Health*, 25(6):630–637, 2016. doi: 10.1089/jwh.2015.5592.

[Zamudio-Rodríguez et al.(2020)] Alfonso Zamudio-Rodríguez, Luc Letenneur, Catherine Féart, José Alberto Avila-Funes, Hélène Amieva, and Karine Pérès. The disability process: is there a place for frailty? *Age and Ageing*, 49(5):764–770, 2020. doi: 10.1093/ageing/afaa031.

Jordan structures of an upper block echelon matrix

B. Cantó♭,[1] R. Cantó♭ and A.M. Urbano♭

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Some methods adapted to structured classes of matrices have been studied recently. An important class of structured matrices due to its applications is the sign regular (SR) matrices, that is, matrices whose all their minors of the same order have the same sign (see e.g. [9] and the references therein). Relevant subclasses of SR matrices are formed by irreducible totally nonnegative (ITN) matrices and by totally nonpositive (t.n.p.) matrices.

Recall that a matrix A is irreducible if there is not a permutation matrix P such that

$$PAP^T = \begin{bmatrix} B & C \\ O & D \end{bmatrix},$$

where O is an $(n-r) \times r$ zero matrix $(1 \leq r \leq n-1)$. A matrix $A \in \mathbb{R}^{n\times n}$ is called ITN matrix if it is irreducible and all its minors are nonnegative. Otherwise, a matrix $A = (-a_{ij}) \in \mathbb{R}^{n\times n}$ is called t.n.p. matrix if all its minors are nonpositive. We have two subclasess of t.n.p. matrices, we say that $A = (-a_{ij}) \in \mathbb{R}^{n\times n}$ is a type-I t.n.p. matrix if it is t.n.p. and $-a_{11} < 0$. Otherwise, $A = (-a_{ij}) \in \mathbb{R}^{n\times n}$ is called type-II t.n.p. matrix if it is t.n.p., $-a_{11} = 0$, $-a_{12} < 0$ and $-a_{21} < 0$. All these matrices have been studied by several authors (see, for instance, [2, 6–8]).

Let $A \in \mathbb{R}^{n\times n}$ be an ITN matrix or a t.n.p. matrix with $\text{rank}(A) = r$ and principal rank $p\text{-rank}(A) = p$, that is, r is the size of the largest invertible square submatrix of A, and p is the size of the largest invertible principal submatrix of A. We say [1] that the sequence of integers $\alpha = \{h_1, h_2, \ldots, h_p\} \in \mathcal{Q}_{p,n}$ is the sequence of the first p-indices of A if for $j = 2, \ldots, p$ we have $\det(A[h_1, h_2, \ldots, h_{j-1}, h_j]) \neq 0$ and $\det(A[h_1, h_2, \ldots, h_{j-1}, t]) = 0$, $\ h_{j-1} < t < h_j$.

Note that if A is totally nonnegative matrix without null rows or columns, then $h_1 = 1$. In [1] the authors use this sequence to study the linear dependence relations between rows or columns of totally nonnegative matrices.

A triple (n, r, p) is called $(1, h_2, \ldots, h_p)$-realizable if there exists an ITN matrix or a t.n.p. matrix $A \in \mathbb{R}^{n\times n}$ with rank r, principal rank p, and $H = \{1, h_2, \ldots, h_p\}$ is the sequence of its first p-indices. It is known that the Jordan structure corresponding to the zero eigenvalue in the Jordan canonical form of A can be complex but not arbitrary. In [2] the authors proved that the number of Jordan canonical forms of ITN matrices associated with a realizable triple (n, r, p) is $P^p_{n-r}(n-p)$, that is, the number of partitions of $n-p$ into exactly $n-r$ parts with the largest part at most p, and presented an algorithm to obtain these Jordan structures. So, the matrix A

[1] bcanto@mat.upv.es

has $n-r$ zero Jordan blocks whose sizes are given by the Segre characteristic of A relative to its zero eigenvalue [10]. This sequence, denoted by $S=(s_1,s_2,\ldots,s_{n-r})$, satisfies

$$\begin{cases} s_1 \leq \min\{r-p+1,p\} \\ s_i \leq s_{i-1}, \quad i=2,\ldots,n-r \\ \sum_{i=1}^{n-r} s_i = n-p. \end{cases}$$

It is known that some properties that ITN matrices satisfy without prescribed p-indices, are not satisfied when they are prescribed. For instance, by [7], the upper bound for the maximum rank of an ITN matrix A associated with a realizable triple (n,r,p) is $n-\left\lceil\frac{n-p}{p}\right\rceil$, but this bound can be lower when the sequence of the first p-indices is prescribed, see [1]. On the other hand, if the sequence of the first p-indices is prescribed, then the number of the zero-Jordan structures admissible for a realizable triple (n,r,p) is less than or equal to this number when the sequence is not prescribed.

In [5] the authors construct an ITN or a t.n.p. matrix A using the product $A=LU$ where L is a lower triangular matrix with all its nonzero elements equal to 1, and U is an upper block echelon TN matrix of size $n\times n$, with rank equal to r, principal rank equal to p and with the same zero-Jordan structure and sequence of the first p-indices as matrix A. We recall that a matrix is an upper echelon matrix if the first nonzero entry in each row (leading entry) is to the right of the leading entry in the row above it and all zero rows are at the bottom. A matrix is upper block echelon if each nonzero block, starting from the left, is to the right of the nonzero blocks below and the zero blocks are at the bottom. A matrix is a lower (block) echelon matrix if its transpose is an upper (block) echelon matrix.

Moreover, in [5] the authors show that this matrix U, can be transformed by similarity and permutation similarity, into a matrix $\widehat{T}=XUX^{-1}$. For that, in this work we obtain all possible rows and columns of $\widehat{T}$ that allow us to obtain the matrices U used to construct some ITN matrices or t.n.p. matrices with a zero-Jordan structure admissible when the sequence of the first p-indices is prescribed.

2 Obtaining rows and columns of matrix $\widehat{T}$

Let U be an upper block echelon matrix of size $n\times n$, such that $\operatorname{rank}(U)=r$, p-$\operatorname{rank}(U)=p$ and $H=[h_1,h_2,\ldots,,h_p]$ as the sequence of its first p-indices. Let $\widehat{T}$ be the matrix obtained from U by similarity and permutation similarity following the steps given in [5]. This matrix has the following structure

$$\widehat{T}=\left[\begin{array}{c|c} A_1 & O \\ \hline O & T \end{array}\right],$$

where $A_1\in\mathbb{R}^{p\times p}$ is a non-derogatory nonsingular matrix, with one Jordan block of size p corresponding to its unique eigenvalue $\lambda=1$ and T is an upper block echelon nilpotent matrix of size $q\times q$, where $q=n-p$. The block partition of T by rows and columns is $[t_1,t_2,\ldots,t_{p-1},t_p]$, that is,

$$T=\begin{array}{c} \left[\begin{array}{ccccccc} O & T_{12} & T_{13} & T_{14} & \cdots & T_{1,p-1} & T_{1p} \\ O & O & T_{23} & T_{24} & \cdots & T_{2,p-1} & T_{2p} \\ O & O & O & T_{34} & \cdots & T_{3,p-1} & T_{3p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ O & O & O & O & \cdots & T_{p-2,p-1} & T_{p-2,p} \\ O & O & O & O & \cdots & O & T_{p-1,p} \\ O & O & O & O & \cdots & O & O \end{array}\right] \end{array} \begin{array}{c} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_{p-2} \\ t_{p-1} \\ t_p \end{array},$$

where $t_i = h_{i+1} - h_i - 1, \quad i = 1, \dots, h_p$, being $h_{p+1} = n$ and its rank is $r_2 = r - p$.

Since the zero-Jordan blocks of T and U are the same, we are going to calculate all possible rows and columns of T to obtain matrices U with the admissible zero-Jordan structures. To construct T, we consider the following conditions: each row and column has at most one 1 and the rest of their entries are 0; and each 1 of a nonzero row is always strictly to the right of the 1 of the row above it. The Jordan structure of T is obtained from the ranks of the successive powers of the matrix.

We calculate all admissible rows of T that allow us to obtain the rank r of the matrix U. The matrix T must have rank r_2, for that, the ones are distributed along the $q - t_p$ rows of T, considering the upper block echelon structure. The different possibilities are the combinations without repetition of $q - t_p$ elements taken r_2 in r_2, that is $\begin{pmatrix} q - t_p \\ r_2 \end{pmatrix}$.

Algorithm 0 (see Annex) calculates the admissible rows. The inputs of this Algorithm are the size n, the rank r, the principal rank p and the sequence of the first p-indices H; the output are all possibilities for the rows of T given by the matrix CR.

On the other hand, it is known that the block structure limits the available columns to calculate the rank r_2 for the matrix T. Then, we calculate, for each admissible row, the different combinations of columns that we can choose to obtain this rank taking into account the structure of T. For that, we use Algorithm 0 (see Annex). The inputs are the triple (n, r, p), the sequence H and one row R from CR; the output is the matrix CCR for each row R. CCR has the different combinations of columns that we can choose to obtain the rank r_2.

We give the following example to clarify the algorithms.

Example 1. We consider a triple $(12, 9, 5)$ and the sequence of the first p-indices $H = \{1, 3, 6, 8, 11\}$. The block partition of T is [1, 2, 1, 2, 1], that is

$$T = \left[\begin{array}{c|cc|c|cc|c}
0 & \star & \star & \star & \star & \star & \star \\ \hline
0 & 0 & 0 & \star & \star & \star & \star \\
0 & 0 & 0 & \star & \star & \star & \star \\ \hline
0 & 0 & 0 & 0 & \star & \star & \star \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & \star \\
0 & 0 & 0 & 0 & 0 & 0 & \star \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right].$$

The different possibilities for the rows of T that we obtain using Algorithm 0 are:

$$CR^T = \begin{bmatrix}
3 & 2 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
4 & 4 & 3 & 3 & 3 & 4 & 3 & 3 & 3 & 2 & 2 & 2 & 2 & 2 & 2 \\
5 & 5 & 5 & 4 & 4 & 5 & 5 & 4 & 4 & 5 & 4 & 4 & 3 & 3 & 3 \\
6 & 6 & 6 & 6 & 5 & 6 & 6 & 6 & 5 & 6 & 6 & 5 & 6 & 5 & 4
\end{bmatrix}.$$

If we choose, for instance, the row $R_2 = [2,\ 4,\ 5,\ 6]$, the Algorithm 0 shows *"The rank cannot be obtained"*, because there are not enought columns to obtain rank r_2 with this rows. Otherwise, if we consider $R_8 = [1,\ 3,\ 4,\ 6]$ then we obtain its corresponding columns:

$$(CCR_8)^T = \begin{bmatrix}
4 & 3 & 3 & 3 & 2 & 2 & 2 \\
5 & 5 & 4 & 4 & 5 & 4 & 4 \\
6 & 6 & 6 & 5 & 6 & 6 & 5 \\
7 & 7 & 7 & 7 & 7 & 7 & 7
\end{bmatrix}.$$

Once we have obtained the corresponding rows and columns that allow us to obtain the rank r_2, we continue as follows. For each row and one of its corresponding columns, for example the row $R = [1,\ 3,\ 4,\ 6]$ of CR and the column $C = [3,\ 5,\ 6,\ 7]$ of CCR_8, we construct a matrix $RC_{2\times(r-p)} = [R; C]$ and we follow the steps described below to obtain the paths of pairs $P_z = (i, j)$, $z = 1, 2, \ldots$. To clarify the procedure we follow it using Example 1.

Step 1. Construct a permutation table with two rows, RU and $\widehat{RT}$. The first entries of the RU are the sequence of the first p-indices H of the matrix U, and the following are the missing number from 2 to n in ascending order. The entries of the $\widehat{RT}$ are consecutive numbers from 1 to n. In Example 1 we have

RU	1	3	6	8	11	2	4	5	7	9	10	12
$\widehat{RT}$	1	2	3	4	5	6	7	8	9	10	11	12

.

Step 2. Obtain a new matrix $\widehat{RC}$, adding p to each entry of RC. That is

$$\widehat{RC} = RC + 5 * \text{ones}(2,5) = \begin{bmatrix} 6 & 8 & 9 & 11 \\ 8 & 10 & 11 & 12 \end{bmatrix}.$$

Step 3. Obtain the positions in the matrix U that correspond to the new entries obtained in Step 2, using the permutation table. For that, we look for the numbers of $\widehat{RC}$ in $\widehat{RT}$ and change them with the corresponding number in RU. In Example 1 is

$$\widehat{RC} = \begin{bmatrix} 6 & 8 & 9 & 11 \\ 8 & 10 & 11 & 12 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 5 & 7 & 10 \\ 5 & 9 & 10 & 12 \end{bmatrix}.$$

Step 4. Create a matrix P_z where the rows contain the numbers obtained in Step 3 plus the p-indices in increasing order. Match the inputs i_t from the first row with its corresponding j_t, $t = 1, \ldots, r-p$ and obtain the path P_z. In Example 1 is

$$P_z = \{(2,3), (3,5), (5,6), (6,8), (7,9), (8,10), (10,11), (11,12)\}.$$

These steps are implemented in Algorithm 0 (see Annex). Finally, with P_z and using Algorithms 5 and 6 from [5], we obtain the matrix U and then, the matrix A that can be an ITN matrix, a type-I t.n.p. matrix or a type-II t.n.p. matrix. All these matrices have the same sequence of its first p-indices and the same zero-Jordan structure (see [3,4]).

Example 2. With Example 1, the row $R = [1\ \ 3\ \ 4\ \ 6]$, the column $C = [3\ \ 5\ \ 6\ \ 7]$ and the path $P_z = \{(2,3), (3,5), (5,6), (6,8), (7,9), (8,10), (10,11), (11,12)\}$, using Algorithm 5 from [5], we obtain the matrix U and $S = (4, 2, 1)$

$$U = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 2 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 3 & 3 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}.$$

Applying Algorithm 6 from [5], we can obtain the matrices ITN, type-I t.n.p. or type-II t.n.p. all of them associated with the triple $(12, 9, 5)$, with the same sequence of its first p-indices $H = \{1, 3, 6, 8, 11\}$ and the same zero-Jordan structure given by the Segre characteristic $S = (4, 2, 1)$.

3 Conclusions

This work improves the result given in [5] because in that paper we only obtain one ITN matrix, type-I t.n.p. matrix or type-II t.n.p. matrix, both of them associated with a triple (n, r, p) with the same sequence of its first p-indices H and the same zero-Jordan structure. Now, in this contribution, we extend the results and obtain all admissible rows and columns that allow us to calculate matrices, with an specific structure, associated with the triple (n, r, p) and with a sequence of its first p-indices H but with all possible Jordan structures.

Acknowledgements.This research was partially supported by the Generalitat Valenciana under the grant CIAICO/2021/162.

References

[1] Cantó, R., Urbano, A.M., On the maximum rank of totally nonnegative matrices *Linear Algebra Appl.*, 551:125–146, 2018.

[2] Cantó, B., Cantó, R., Urbano, A.M., All Jordan canonical forms of irreducible totally nonnegative matrices *Linear Multilinear Algebra*, 69(13):2389-–2409, 2021.

[3] Cantó, B., Cantó, R., Urbano, A.M., Irreducible totally nonnegative matrices with a precribed Jordan structure *Linear Algebra App.*, 609:129–151, 2021.

[4] Cantó, B., Cantó, R., Urbano, A.M., On totally nonpositive matrices associated with a triple negatively realizable *RACSAM REV R ACAD A*, 134:1–25, 2021.

[5] Cantó, B., Cantó, R., Urbano, A.M., Jordan structures of irreducible totally nonnegative matrices with a prescribed sequence of the first p-indices *RACSAM REV R ACAD A*, 116(84):1–27, 2022.

[6] Cantó, R., Peláez, M.J., Urbano, A.M., On the characterization of totally nonpositive matrices *SeMA Journal*, 73:347–368, 2016.

[7] Fallat, S.M., Johnson, C.R., Totally Nonnegative Matrices. Oxford, Princeton Ser. Appl. Math., 2011.

[8] Fallat, S.M., van den Driessche, P., On Matrices with all minors negative *Electronic Journal of Linear Algebra*, 7:92–99, 2000.

[9] Peña, J.M., On nonsingular sign regular matrices *Linear Algebra and its Applications*, 359:1–3, 2003.

[10] Shapiro, H., The Weyr characteristic. London, The American Mathematical Monthly, 1999.

4 Annexes

Algorithm 1 $CCR = \text{pairsPz}(n, r, p, H, R, C)$

```
$PT = p * \text{ones}(r-p, 2) + [R'\ C']$;   $RT = [1:n]$;   $RU = [0]$;
for $j = 1:n$ do
    $m = \text{ismember}(RT(1,j)H)$;
    if $m\ = 1$ then
        $RU = [RU, RT(1,j)]$;
    end if
end for
$RU = [H, RU(1, 2:n+1-p)]$;
for $j = 1:r-p$ do
    $Pz(j,:) = [RU(1, PT(j,1))RU(1, PT(j,2))]$;
end for
$Pz = [\text{sort}([S(:,1);\ H'])\text{sort}([S(:,2);\ H'])]$;   $Pz = Pz(2:r,:)$;
```

Algorithm 2 $CR = \text{rows}(n, r, p, H)$

```
$q = n - p;$   $r2 = r - p;$   $HA = [H, n + 1];$   $a2 = length(HA);$
$R = \text{combnk}([1 : q - (HA(a2) - HA(a2 - 1) - 1)], r2);$
```

Algorithm 3 $CCR = \text{columnsR}(n, r, p, H, R)$

```
$q = n - p;$   $r2 = r - p;$   $HA = [H, n + 1];$   $C = \text{combnk}([HA(2) - HA(1) : q], r2);$
$a2 = \text{size}(C, 1);$   $B = [0];$
for $i = 1 : p$ do
    $B(i + 1) = B(i) + HA(i + 1) - HA(i) - 1;$
end for
$E = [0]; i = 1;$
while $i <= r2$ do
    $j = 1;$
    while $j <= p$ do
        $m = \text{ismember}(R(i), [B(j) + 1 : B(j + 1)]);$
        if $m == 1$ then
            $E = EB(j + 1)];$   $j = p + 1;$
        else
            $j = j + 1;$
        end if
    end while
    $i = i + 1;$
end while
$E = E(2 : r2 + 1);$   $CCR = \text{zeros}(1, r2);$   $i = 1;$
while $i <= a2;$ do
    $d = C(i, :) - E;$   $h = d > 0;$   $l = h * \text{ones}(r2, 1);$
    if $l == r2$ then
        $CCR = [CCR; C(i, :);$
    end if
    $i = i + 1;$
end while
$a3 = \text{size}(CCR, 1);$
if $a3 == 1$ then
    The rank cannot be obtained
else
    $CCR = CCR(2 : a3, :);$
end if
```

10

Modeling inflammation in wound healing

A. Patrick[♭,1] and B. Chen-Charpentier[♭,2]

(♭) Department of Mathematics,
University of Texas at Arlington
Arlington, Texas 76019-0408 , USA.

1 Introduction

A wound is an injury to living tissue in the body caused by a cut or an impact. The body has mechanisms to repair the wound. These mechanisms of wound healing are organized into the overlapping stages, namely, homeostasis, inflammation, proliferation, and remodeling. Within minutes of the happening of the wound, blood clots are formed and the bleeding stops. This is the first provisional matrix formed for the wound. Next, in the inflammation stage the focus of the mechanisms implemented will be to remove pathogens and debris . The turnover of the provisional matrix proceeds in the proliferation and remodeling stages. Understanding the mechanisms in wound healing is of interest in improving wound care and determining causes of diseases.

According to MedlinePlus [6] by the United States National Library of Medicine, wounds are ”injuries that break the skin or other body tissues” and can include ”cuts, scrapes, scratches, and punctured skin.” The formation of a wound involves a disruption in skin and other tissue integrity. Where and how this integrity of the body composition and function is disrupted is what characterizes the type of wound and what can elucidate the overall pathology to be diagnosed. Some types of wounds include those caused by mechanical stress. Examples of these types of wounds are penetrating wounds, blunt force trauma wounds (i.e., abrasions, lacerations, skin tears), closed wounds (i.e., contusions, hematomas), etcetera. Mechanisms that cause chronic wounds are another important type of wound to study since these types of wounds can lead to amputation and death. Nussbaum et al. [9] points out that individuals who are at risk are elderly, disabled, or in general, individuals who cannot care from themselves and individuals with pre-existing skin or immunological conditions.

In addition to being involved in the before mentioned wounds, the wound healing process is involved in cardiovascular issues. For example, when blood does not circulate properly in the myocardium, oxygen is not able to be distributed to cells. Heart cells undergo necrosis and tissues die, hence the sequence of phenomena that characterizes what is known as mycardium infarction (i.e., heart attack). This context of wound healing has been of interest due to this repair process leading to adverse after effects. The dynamics of this process have been studied by methods incorporating experimental data and differential equation modeling such as in [4].

The dynamical processes of wound healing is not just important for treating wounds but in addressing economical related concerns. Research by Nussbaum et al. [9] conducted a retrospective

[1] apatrick@mavs.uta.edu
[2] bmchen@uta.edu

analysis of the cost of chronic wound care. Some prevalent wound types among Medicare beneficiaries were surgical infections, diabetic infections, traumatic wounds, skin disorders, and venous infections. Data from 2014 revealed that there was a high prevalence amount individuals who were 75 years or older, and total Medicare expenditure estimates reached between 3 billion and 96.8 billion dollars where surgical wounds were one of the most expensive.

Mathematical modeling provides a means to help provide framework and understanding It is a tool to develop theories and try scenarios that may be hard or impossible to test under experimentation. Ordinary differential equation models are useful when studying dynamics over time in a nonspecific unit of space. Previous studies have incorporated this methodology to study wound healing phenomena such as the recovery after a myocardial infarction [4], keloid and hypertorphic scarring [1], relationship between transforming growth factor β and tissue tension [8], and acute wound healing [7], just to name a few.

There are several types of models of wound healing. Ordinary differential equations models do not model dependence in space, but concentrate on change over time for a determined unit of space. These are the most common models. Among the papers with wound models that have been presented, Reynolds et al. [11] focused on inflammation and anti-inflammation with their state variables being activated phagocytes, tissue damage, and anti-inflammatory mediators. Activated phagocytes are representative of inflammation and include neutrophils and macrophages. The anti-inflammatory variable is representative of mediators including cortisol and interleukin-10 (IL-10). Here subsystems based on mass action kinetics are presented and the quasi-steady state assumption is implemented on the local response and resting phagocytes. Subsequent to model construction, different scenarios of pathogen growth rate and initial conditions are presented in scenarios where the wound is under circumstances that result in a healthy outcome, aseptic death, and septic death.

Cooper et al. [2] built a model expanding on Reynolds. The inflammation state variable replaced by more specialized inflammatory variables, neutrophils and macrophages. Estrogen and cortisol inhibition and enhancement factors are also implemented. A set of parameters that result in dynamical assumptions are simulated, the assumptions being that neutrophils peak between 0.75 and 2.75 days; the peak for macrophages is between 2.75 and 6.25 days; and that macrophage levels dropped below 0.1 by day 20. For some of the state variables, the rates per cell units were adapted from Reynolds. More specifically, pathogen carrying capacity was chosen as 10^6 cells per unit space per unit time which, in order for the model to be well defined, sets P-units for the state variable P to be the same. The units for these cells are important when comparing results to experimental conditions, when combining results from different experiments, or when comparing parameters from different model experiments. For the Cooper study the purpose was to simulate general behavior, so units for some state variables such as debris were set to be arbitrary.

Torres et al. [13] utilizes these cell dynamics with experimental results. For pathogens, neutrophils, M1 and M2 macrophages, units are chosen to be 10^7 units. Note that the interaction mechanisms such as phagocytation and activation which has units pathogen units per inflammation cell units per time and cells per pathogen units per time, respectively, will be affected by these chosen units. For example, the background immune defense parameter, k_{bp}, activation of local background immune response by pathogen, value would change when switching from 10^6 P-units to 10^7 P-units.

Some studies that modeled the proliferation and remodeling stage are Jin et al. [4] and Segal et al. [12]. Jin et al. incorporates macrophages, MMP-9, TGF-β, fibroblasts, and collagen to model the healing process after a myocardial infarction. They assume inhibition of MMP-9 by TGF-β due to the induced presence of TIMP-1. Subsequent to model formation they validate their model and parameter values by comparing model output to experimental data.

Segal et al. [12] constructs a model incorporating inflammation, pathogens, fibroblasts, and collagen. They include three types of fibroblasts (proliferating fibroblasts, migrating fibroblasts, and active fibroblasts). The collagen state variable is chosen to be a percentage where 0 represents the wound not being filled and 1 being the wound being filled. Values are allowed to go above 1 to account for scarring. They also include both inhibition of collagen deposition and degradation influenced by the current amount of collagen fibers formed. It is assumed the closer the wound is from being filled, the less need for fibroblasts which reduces amount of collagen being deposited. Moreover, they assume that inflammation cells can release enzymes that degrade collagen. To determine values for their parameters experimental data for collagen where scaled so that the highest values was a little above one. Subsequently, the resulting model was tested using low and high values of pathogen that resulted in high collagen deposition and low collagen disposition, respectively.

As mentioned above, the main four stages of wound healing are: 1. Hemostasis, blood coagulates. 2. Inflammation, sterilization of the wound. 3. Proliferation, tissue regrowth. 4. Remodeling, formation of new epithelium and scar tissue. In this paper, we will construct a model of the inflammation phase. It will extend the results of [2] and [13] by adding more species and adding and modifying certain hypotheses based on newer research.

2 Methods

2.1 State variables

The model considers the following variables: 1. Pathogens, P, that are assumed to proliferate logistically. 2. Debris, P_t. Upon initiation of the wound and the physical obstruction of tissue, these dead cells and possibly outside debris must be removed. 3. Neutrophils, N, which are the first phagocytic cells introduced to the wound during the inflammation stage. Neutrophils phagocytize pathogens and debris and afterwards commit apoptosis. Apoptotic Neutrophils, A_N, that can influence some critical parts of the wound healing process, namely, the continuation of the inflammation stage, the resolution of debris removaland the polarization between macrophages. M1 and M2 Macrophages, M_1 and M_2, have diverse functions in wound healing. There is spectrum of macrophage types. For our application, we will choose to categorize macrophages with more pro-inflammatory functions as M1 macrophages and macrophages with more anti-inflammatory functions as M2 macrophages. There are many chemical factors present, but in order not to over complicate the model, their quantities are assumed proportional to the cells producing them.

2.2 Parameters

Parameter values were estimated based on previous findings in the literature. One estimate utilizing a list of assumptions and another estimate based upon experimental data. For the first estimate, the following assumptions were implemented:

1. Neutrophils peak around day 1 day post injury (dpi)
2. M1 macrophages peak between 2 - 3 dpi
3. M2 macrophages peak at least 1 dpi after M1 macrophages peak and once M2 macrophages peak they are the dominant macrophages present.

Initial conditions where assumed to be $P_0 = 1$ and $Pt_0 = 2$ which has a higher starting pathogen density then in Cooper et al. [2] to simulate a wound with more pathogenic result.

For the method utilizing data, parameters were found that minimized the output from experimental data (Torres et al., [13]). In Torres, the experiment captures wound dynamics under the conditions that there is pathogen, but minimal debris. Pathogen starts small, but due to increase in carrying capacity the pathogen density increases causing the onset of the inflammation process. For the initial conditions we assume $P_t = 0.001$ and let P_0 be an optimization argument. Due to the nature of the experiment, another state variable B for broth was included. To account how the broth affects the carrying capacity, another parameter $k_{kbp\infty}$ was included. The inflammation dynamic from this experiment is modeled by the system with a modification on the logistic growth term for pathogens and with the inclusion of a new equation for broth which is the following:

$$\frac{dB}{dt} = -k_{bp}BP.$$

The pathogen equation was modified to the following:

$$\frac{dP}{dt} = k_{pg}P(1 - \frac{P}{P_\infty + k_{kbp\infty}B}) - \frac{k_{pb}s_bP}{\mu_b + k_{bp}P} - k_{pn}PN(1 + k_{en}E) - k_{pm}P(M_1 + M_2)(1 + k_{em}E).$$

The fmincon MATLAB function [5] was used to fit the solution to the data using the following equation weighted least squares function:

$$\min_p \sum_{i=1}^{n} (\frac{y_i - y(t_i, p_i)}{\sigma_i})^2. \tag{1}$$

2.3 The model

$$\begin{aligned}
\frac{dP}{dt} &= k_{pg}P(1 - \frac{P}{P_\infty}) - \frac{k_{pb}s_bP}{\mu_b + k_{bp}P} - k_{pn}PN(1 + k_{en}E) - k_{pm}P(M_1 + M_2)(1 + k_{em}E) \\
\frac{dP_t}{dt} &= \mu_{an}A_N - k_{ptn}P_tN(1 + k_{en}E) - k_{ptm1}P_tM_1(1 + k_{em}E) \\
&\quad - k_{ptm2}P_tM_2(1 + k_{em}E) - \mu_{pt}P_t \\
\frac{dA_N}{dt} &= k_{an}N - k_{anm1}A_NM_1(1 + k_{em}E) - k_{anm2}A_NM_2(1 + k_{em}E) - k_{ann}N(1 + k_{en}E) \\
&\quad - d_{an}A_N - \mu_{an}A_N \\
\frac{dN}{dt} &= R_N\frac{S_{nr}}{\mu_{nr} + R_N}\frac{1}{(1 + \frac{E}{E_{ninf}})^2} - k_{an}N, \;\; R_N = k_{npt}P_t + k_{np}P + k_{nan}A_N \\
\frac{dM_1}{dt} &= R_{M1}\frac{s_{mr}}{\mu_{mr} + R_{M1} + R_{M2}} - k_{m1m2}A_NM_1 + k_{m2m1}M_2 - \mu_{m1}M_1, \\
&\quad R_{M1} = k_{mpt}P_t + k_{m1p}P + k_{m1n}N + \frac{k_{m1m1}M_1}{1 + (\frac{E}{E_{M\infty}})^2} + k_{m1an}A_N \\
\frac{dM_2}{dt} &= R_{M2}\frac{s_{mr}}{\mu_{mr} + R_{M1} + R_{M2}} + k_{m1m2}A_NM_1 - k_{m2m1}M_2 - \mu_{m2}M_2, \\
&\quad R_{M2} = k_{m2m2}M_2 + k_c.
\end{aligned} \tag{2}$$

3 Results

Numerical simulations were performed using different initial conditions to take into account the severity of the wound. The peaks of Neutrophils, and of macrophages M1 and M2 were at the correct times and heights. The wound healed in all simulations with the parameters used. The numerical simulations were done using the routine ode45 of MATLAB [5] To analyze the model behavior in regard to the parameters, a bifurcation analysis was conducted in XPPAUTO [3]. The following are analyses conducted for pathogen growth and carrying capacity.

Bifurcation analysis was conducted with respect to kpg, the growth rate of the pathogen. XPPAUTO indicated bifurcation points in three main locations: between 15 and 16, at 54.956, and a Hopf bifurcation at 15.446. The XPPAUTO program was subsequently run from the periodic solutions from the Hopf points, which resulted in unstable solutions. Different kpg values where tested around the bifurcation points. For high enough pathogen and debris initial conditions, kpg = 20 and kpg = 80 resulted in an unhealthy outcome indicated by the high end behavior for M1 and M2 macrophages. For kpb, the source of background local response, there where two bifurcation points found: a Hopf bifurcation at kpb = 16.215 and at kpb = 17.57 . For low values of kpb the steady state is relatively high for all state variables and also the transient end behaviors are high. For ub, the intrinsic decay of local response, there where two bifurcation points found: at ub = 0.2085 and at a Hopf bifurcation ub = 0.1748 . As ub increases the steady state of each variable steadily increases but after a high enough value this steady state does not change much. Around the bifurcation point there is oscillatory behavior in apoptotic neutrophils. For snr, the source of resting neutrophils, there was one bifurcation point found at snr = 4.222 . For small snr, the steady state for debris, apoptotic neutrophils, M1 macrophages, and M2 macrophages is relatively high and it goes down as snr increases. After snr is big enough the steady state starts increasing at a steady rate and for the other state variables except pathogens the steady state remains relatively higher.

A global sensitivity analysis of each of the state variables with respect to all the parameters was done. The chosen sensitivity analysis method used was the Fourier Amplitude Sensitivity Test (FAST). This was implemented within the SAFE package [10] written in MATLAB. The FAST method is a variance based method that implements ANOVA decomposition and uses the Fourier series to estimate the total model variance. The Fourier transform is used to decompose the variance of the model output described by each parameter. The sensitivity indices are the proportion of the variance attributable to the factor of interest over the total variance and has a range between 0 and 1. This method was chosen since it is computationally efficient and can be used for non-linear, non-monotonic models [14]. This analysis is conducted utilizing each state variable as the output, and afterwords the average of all the state variables as the output.

4 Conclusions

Models can capture multifaceted elements that control outcomes, but these models are a tool that apply results based on current knowledge. As the experimental literature on these topics expand, the more accurate these scientific models will be as the assumptions are updated. The use of the model is dependent on the specifications the model was built upon. Models are supplemental in providing a means to elucidate outcomes and effects and can help build integrity in the experimental literature where unavoidable limitations such as those in *in vivo* vs *in vitro*, animal vs human research, and also the consideration that immune response from a wound in a controlled environment may not be the same as in a specimen that is exposed to different environments.

Established models lead to further analysis that important in analyzing the behavior under a set of specified assumptions and circumstances. Methods such as sensitivity analysis is useful for identifying which factors are important. These findings can be used to help factor in consideration for experimental design. For example, the influx of monocytes when compared to other factors may have a high sensitivity index for the overall inflammation stage implying experimental investigations of changes in the expected inflammation output may consider monocyte recruitment. The bifurcation results show that there are certain condition sunder which wounds do not heal, and how changing certain values of the parameters the wound will heal. This can be very useful in practice when wounds do not heal.

References

[1] C. Cobbold and J. Sherratt. Mathematical Modelling of Nitric Oxide Activity in Wound Healing can explain Keloid and Hypertrophic Scarring. *Journal of theoretical biology*, 204:257–88, June 2000.

[2] R. L. Cooper, R. A. Segal, R. F. Diegelmann, and A. M. Reynolds. 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.

[3] B. Ermentrout and A. Mahajan. Simulating, analyzing, and animating dynamical systems: a guide to xppaut for researchers and students. *Appl. Mech. Rev.*, 56(4):B53–B53, 2003.

[4] Y. Jin, H.-C. Han, J. Berger, Q. Dai, and M. Lindsey. Combining experimental and mathematical modeling to reveal mechanisms of macrophage-dependent left ventricular remodeling. *BMC Systems Biology*, 5:60 – 60, 2010.

[5] MATLAB. *MATLAB R2021a*. The MathWorks Inc., Natick, Massachusetts, 2021.

[6] Medlineplus. Wounds and Injuries. Publisher: National Library of Medicine.

[7] N. B. Menke, J. W. Cain, A. Reynolds, D. M. Chan, R. A. Segal, T. M. Witten, D. G. Bonchev, R. F. Diegelmann, and K. R. Ward. An in silico approach to the analysis of acute wound healing. *Wound Repair and Regeneration*, 18(1):105–113, Jan. 2010. Publisher: John Wiley & Sons, Ltd.

[8] K. E. Murphy, C. L. Hall, S. W. McCue, and D. Sean McElwain. A two-compartment mechanochemical model of the roles of transforming growth factor β and tissue tension in dermal wound healing. *Journal of Theoretical Biology*, 272(1):145–159, Mar. 2011.

[9] S. R. Nussbaum, M. J. Carter, C. E. Fife, J. DaVanzo, R. Haught, M. Nusgart, and D. Cartwright. An Economic Evaluation of the Impact, Cost, and Medicare Policy Implications of Chronic Nonhealing Wounds. *Value in Health*, 21(1):27–32, Jan. 2018. Publisher: Elsevier.

[10] F. Pianosi, F. Sarrazin, and T. Wagener. A matlab toolbox for global sensitivity analysis. *Environmental Modelling & Software*, 70:80–85, 2015.

[11] A. Reynolds, J. Rubin, G. Clermont, J. Day, Y. Vodovotz, and G. Bard Ermentrout. A reduced mathematical model of the acute inflammatory response: I. Derivation of model and analysis of anti-inflammation. *Journal of theoretical biology*, 242(1):220–236, Sept. 2006.

[12] R. Segal, R. Diegelmann, K. Ward, and A. Reynolds. A Differential Equation Model of Collagen Accumulation in a Healing Wound. *Bulletin of mathematical biology*, 74:2165–82, July 2012.

[13] M. Torres, J. Wang, P. J. Yannie, S. Ghosh, R. A. Segal, and A. M. Reynolds. 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.

[14] C. Xu and G. Gertner. Understanding and comparisons of different sampling approaches for the Fourier Amplitudes Sensitivity Test (FAST). *Computational Statistics & Data Analysis*, 55:184–198, Jan. 2011.

11

The geometry behind PageRank rankings

Gonzalo Contreras-Aso[♭,♮,1], Regino Criado[♭,♮], and Miguel Romance[♭,♮]

(♭) Department of Applied Mathematics, Material Science and Electronic Technology,
Universidad Rey Juan Carlos,
28933 Móstoles, Madrid, Spain.

(♮) Laboratory of Mathematical Computation on Complex Networks and their Applications,
Universidad Rey Juan Carlos,
28933 Móstoles, Madrid, Spain.

1 Introduction

It has been almost 25 years since the invention of the PageRank algorithm [1], which established Google as the world's most used search engine ever since. This algorithm relies on the network of hyperlinks between different websites, which can be described as a directed graph $G = (\mathcal{N}, \mathcal{E})$ where $n, m \in \mathcal{N}$ represent webpages and $(n, m) \in \mathcal{E}$ iff there is a hyperlink between in webpage n pointing to webpage m. Furthermore, this graph as a representation in terms of its adjacency matrix $A = (a_{nm}) \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}$, where $a_{nm} = 1$ if $(n, m) \in \mathcal{E}$, otherwise $a_{nm} = 0$.

Let us establish some notation which will come in handy. We will denote the number of nodes as $|\mathcal{N}| = N$. We say that a vector $v \in \mathbb{R}^N$ is *stochastic* if all of its components are positive ($v_i > 0$, $\forall\, i = 1, ..., N$) and if it has unit 1-norm ($|v|_1 = 1$). We will also denote the vector of ones as $e = (1, ..., 1) \in \mathbb{R}^N$.

The PageRank algorithm computes the only stochastic vector $\pi(G, \alpha, v) \in \mathbb{R}^N$ solution to the eigenvalue problem

$$\pi^T = \pi^T(\alpha P + (1-\alpha)e\, v^T) \tag{1}$$

where $\alpha \in (0, 1)$ is the *damping factor*, $v \in \mathbb{R}^N$ stochastic is the *personalization vector*, and P is the row-normalizedadjacency matrix of G.

The interpretation behind this equation is the following: consider a random walker on a node in the graph, following a simple set of rules. With probability α, it will follow the out-edges of the node, with equal probability for each (this corresponds to the first term, due to the row-normalization), with probability $1 - \alpha$ it will instead teleport at random to another node in the network, with the chance to teleport to node i given by v_i. Then, the PageRank vector π is the stationary distribution of the random walker on the nodes of the network. This interpretation represents an average user, surfing the web following links until searching something else completely unrelated.

Although at first [1] they were not explicitly mentioned, these network theoretical foundations were soon recognized by the scientific community and ignited a torrent of research in understanding the main features of this algorithm. In particular, the dependency of the PageRank vector on α was thoroughly studied (see for instance [2]), but something that remained understudied is the relation between the personalization vector v and the PageRank vector.

In this work we attempt to shed some light in this matter, in relation to centrality controllability.

[1] gonzalo.contreras@urjc.es

2 PageRank controllability

Our goal is to understand how far PageRank scores can be altered after making changes in some of its ingredients. At first glance, there are two changes that turn out to be dead ends in order to attain full control of the centrality scores:

- Modifying the weights of the underlying graph G: this won't allow us to set an arbitrary centrality vector as the PageRank vector. The reason for this is the row-normalization of the adjacency matrix in order to obtain P, this renders any single out-edge (those from nodes with out-degree one) weightless to the eyes of PageRank.

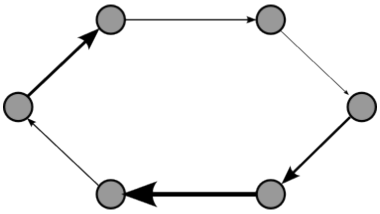


Figure 1: A simple counterexample, the directed cycle C_6. Weighting it is pointless, due to the row-normalization.

- Modifying the damping factor: it would be unreasonable to think that changing a single parameter could reach the whole space of centrality vectors. In fact, it's easy to check that for $\alpha = 1$ the PageRank solution is the eigenvector associated to the spectral radius $\rho(P)$, but for $\alpha = 0$ the result will depend on the choice of personalization vector v. This already hints at the interplay between the damping factor and the personalization vector when trying to control the PageRank centrality.

We want to make this interplay between the damping factor and the personalization vector, in relation to centrality scores, explicit. In other words, can any PageRank vector be set for a given graph and damping factor if we have control over the personalization vector used in the algorithm?

The answer is no: in general there is no positive ($\langle v, e_i \rangle > 0$, $\forall 1 \leq i \leq N$) solution. Nevertheless, we can study the conditions under which π_0 actually has an associated personalization vector v. In fact, in [4] the following theorem regarding the above conditions is proven:

Theorem I (Existence of the personalization vector). Given $G = (V, E)$ and $\pi_0 \in \mathbb{R}_+^N$ with $|\pi_0|_1 = 1$, then there exists $v \in \mathbb{R}_+^N$ with $|v|_1 = 1$ personalization vector such that

$$\pi(G, \alpha, v) = \pi_0 \iff \langle \pi_0, e_j \rangle > \alpha \pi_0^T P e_j \quad \forall 1 \leq j \leq N \tag{2}$$

This is a highly restrictive condition for the existence of a personalization vector, as it consists of N inequalities which must be satisfied together, for a fixed choice of α. We can, however, relax these constraints while retaining control over the centrality, if we consider controlling centrality "rankings" rather than centrality "scores". The idea is that, for most practical applications, one is not interested in the specific centrality score of each node. Instead, what's interesting is which nodes are more important than others: the ranking (first, second, third, ...) of the nodes by their importance.

Dealing with centrality rankings seems harder than with specific scores, however in the PageRank case there is a nice geometrical interpretation which will allow us to crack the problem of their control. Consider the N-simplex defined as

$$\triangle_N = \{x \in \mathbb{R}_+^n; \quad |x|_1 = 1\}. \tag{3}$$

This represents both the space of all possible personalization vectors and the space of all possible PageRank vectors of graphs with n nodes. Therefore, we can see PageRank as the following map:

$$\begin{aligned} \pi(G, \alpha, \cdot): \quad & \triangle_N \longrightarrow \triangle_N \\ & v \hookrightarrow \pi(G, \alpha, v) \end{aligned} \tag{4}$$

The key point in this geometric viewpoint is that we can associate each possible ranking to a portion of the simplex. Consider the midpoint of the simplex, given by normalization of e, i.e. $\tilde{e} = \frac{1}{N} e = \frac{1}{N} \sum_{i=1}^{N} e_i$. We can then define the hyperplanes bisecting the simplex through the midpoint $\tilde{e}$ and any combination of $n-2$ vertices as

$$\mathcal{H}_N^{i,j} = \left\{ \sum_{k=0,\, k \neq i,j}^{N} \lambda_k e_k \;\middle|\; e_0 \equiv \tilde{e},\ \lambda_k \in \mathbb{R} \right\}. \tag{5}$$

The relevance of this construction is that it provides us with a way to classify the points $\pi \in \triangle_n$ according to their ranking, as shown in 2.

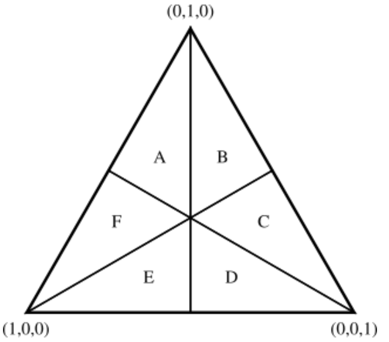


Figure 2: Different ranking regions in the $n = 3$ case. For instance, $\pi = (\pi_1, \pi_2, \pi_3) \in A \setminus \partial A$ corresponds to $\pi_2 > \pi_1 > \pi_3$, while the intersection between triangles would lead to equal scores, e.g. $\pi \in B \cap C$ would correspond to $c_2 = c_3 > c_1$.

In this light, we can see that there is ranking control if and only if

$$\tilde{e} = \frac{1}{n} e \in \mathrm{Im}(\pi), \quad \tilde{e} = \frac{1}{n} e \notin \partial \mathrm{Im}(\pi). \tag{6}$$

The argument here is identical to that of the hyperplanes: $\pi = \tilde{e}$ is the point in ranking space where $c_1 = c_2 = ... = c_N$. And given that all hyperplanes $\mathcal{H}_n^{i,j}$ pass through $\tilde{e}$ by construction, all ranking regions are ϵ away from it. Thus, moving ϵ away in any direction will lead to different rankings. The following theorem, proven in [4], formalizes this idea:

Theorem II (Necessary condition for ranking control) Given a network $G = (V, E)$ and $\alpha = (0, 1)$, then it is possible to obtain obtain any ranking of the nodes under the PageRank $\pi(G, \alpha, v)$ for a $v \in \triangle_n$ if and only if

$$\frac{1}{\alpha} > \max_j \left(\sum_{i=1}^{n} P_{ij} \right) \tag{7}$$

This tells us that if we have $\alpha < 1/\max_j \sum_i P_{ij}$, we can always find the chosen ranking with an appropriate personalization vector.

2.1 Real network implications

It is left for us to discuss the implications of the above inequality. While it is far less restrictive than 2, it is nevertheless a hard bound; in networks with thousands of nodes, this condition will likely force α to be very small.

In fact, we have computed the maximum value of α satisfying that bound for several real directed networks (all fetch from the KONECT network repository [5] and the CASOS network repository), and indeed these values are incredibly small. This is shown in Figure 3.

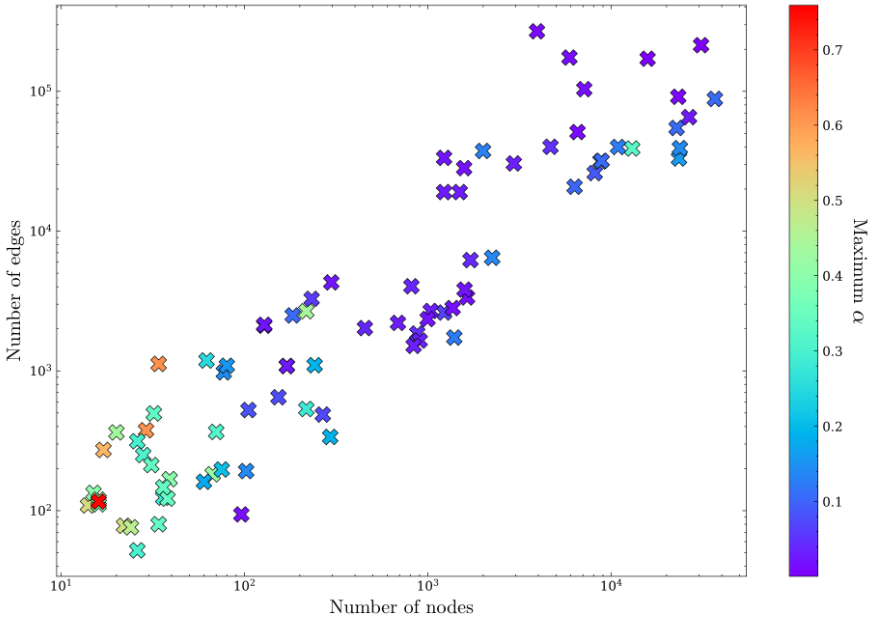


Figure 3: Scatter plot showing the number of edges against the number of nodes for 84 different real networks, with datapoints colored based on the maximum value of α providing ranking control.

As we can see in the Figure, networks with their maximum α (that saturating the bound) are mostly below 0.4, with a handful of exceptions above 0.6.

3 Conclusions and outlook

Knowing that most applications of PageRank use relatively high values of α, around 0.85, the above discussion implies that there is absolutely no ranking control in those situations. But this is good news, as it implies that the orderings yielded with PageRank can't be judiciously modified.

The described methods and geometric ideas are not restricted to the standard PageRank algorithm: there are other ones which can benefit from theorems 1 and 2, so long as they feature a map like 4. In view of this fact, an analogous treatment is carried out in [4] for the biplex PageRank algorithm, an alternative to the standard PageRank one studied in [6].

References

[1] L. Page, S. Brin, R. Motwani and T. Winograd. "The PageRank Citation Ranking: Bringing Order to the Web". *Stanford InfoLab*, 1999-66, 1999

[2] P. Boldi, M. Santini and S. Vigna. "PageRank as a Function of the Damping Factor". *Proceedings of the 14th International Conference on World Wide Web*, 557–566 (2005)

[3] E. García and F. Pedroche and M. Romance. “On the localization of the personalized PageRank of complex networks”. *Linear Algebra and its Applications*, 439, 3, 640-652, (2013)

[4] G. Contreras-Aso, R. Criado and M. Romance. “Parametric control of PageRank centrality rankings: a geometrical approach”, *Preprint* (2022)

[5] J. Kunegis. “KONECT - The Koblenz Network Collection”, *Proc. Int. Conf. on World Wide Web Companion*, 1343–1350 (2013)

[6] F. Pedroche, M. Romance and R. Criado. “A biplex approach to PageRank centrality: From classic to multiplex networks”, *Chaos* 26, 065301 (2016)

12

Surveillance model of the evolution of the plant mass affected by Xylella fastidiosa in Alicante (Spain)

José Juan Cortés Plana,[1] María Teresa Signes Pont,
Joan Boters Pitarch and Higinio Mora Mora

Department of Computer Science and Technology,
Polytechnic School, University of Alicante
Carretera San Vicente del Raspeig s/n,
03690 San Vicente del Raspeig - Alicante, Spain.

1 Introduction

Xylella fastidiosa (Xf) is an insect-transmitted bacterial plant pathogen associated with serious diseases in a wide range of plants. The vectors or transmitting agents of the disease are insects that feed on the xylem of plants. Once the vector bites and sucks on an infected plant, it acquires the bacteria that remain in the insect's feeding structures and will transmit it to the next plant it bites for nourishment. It has no cure.

Since 2013, several outbreaks appeared in the Mediterranean area: in 2017, the threat reached the almond trees of the province of Alicante (Valencian Community, Spain)

In general, the rapid spread of invasive organisms is the result of the drastic alteration of natural environments by human activities as well as global trade. Our mobile society is redistributing species on earth at a rate that challenges ecosystems, threatens human health and produces economic stress [1]. Climate change is also allowing certain bacteria to occupy new niches that were not suitable before [2].

Farmers do not agree with the policies of the European Union to eradicate: 81600 trees are going to be cut down, so their incomes have dropped dramatically.

In this paper we present, a surveillance model that aims to prevent and minimize as far as possible the impact of the expansion of Xylella fastidosa. For that, we control the evolution of plant mass since it is the main factor that allows the nymphs of the insect to develop

2 Methods

2.1 Temperature and humidity impact on the propagation of Xf

This, which is of transcendental importance since it affects the bacterium due to its thermophilic nature, has made it possible to establish models that determine the potential risk of contamination based on the minimum temperature. Although these are still preliminary results, the maps developed by the Valencian Institute of Agricultural Research (IVIA) provide a first approximation of

[1] jj.cortes@ua.es

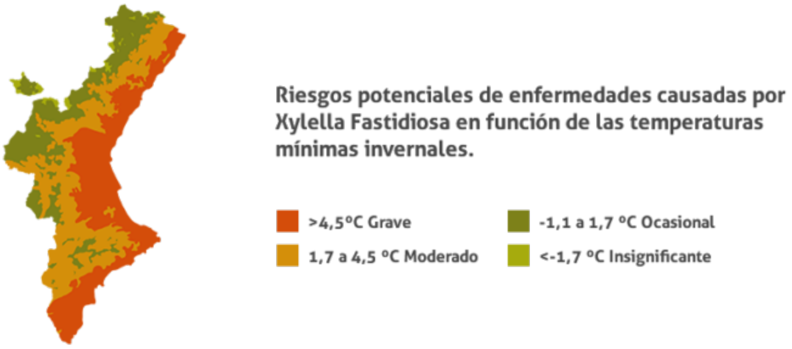

Riesgos potenciales de enfermedades causadas por Xylella Fastidiosa en función de las temperaturas mínimas invernales.
>4,5°C Grave
-1,1 a 1,7 °C Ocasional
1,7 a 4,5 °C Moderado
<-1,7 °C Insignificante

Figure 1: Impact of temperature on the propagation of Xylella fastidiosa Comunidad Valenciana (Spain) [3]

the areas with a potentially more favorable climate for the development of this disease, as shown in Fig.1, [3]

In addition to the temperature maps of Purcell that divide the total territory into four temperature areas approximately parallel to the Mediterranean coast (see Figure 1), other works such as [4]. We have observed humidity as it affects the spread of the pest Xylella fastidiosa (Xf)

As we have temporary photos at a certain moment, depending on the level of rain, wind, temperature, we could obtain a model based on the plant mass based on some variables that would give more weight to the model. These variables are basically humidity. We could establish an improved model to that of the Conselleria Agricultura of Generalitat Valenciana of temperature Figure 1, with depending on the amount of rain and the amount of heat the plant mass that would act as a reservoir for the vectors that spread the plague

2.2 Collecting data to control the evolution of the plant mass

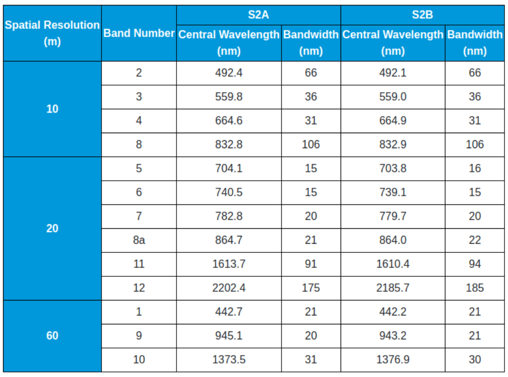
We will use the available data from the ESA (European Space Agency) Copernicus project. Copernicus is a project that consists in observing the earth through satellites. Specifically, we will use images taken by the Sentinel-2 project. Sentinel-2 is a polar-orbiting, high-resolution multispectral imaging mission for land monitoring to provide, images of vegetation, land and water cover, inland waterways, and coastal areas. It has two satellites in orbit, Sentinel-2A which was launched on June 23, 2015 and Sentinel-2B followed on March 7, 2017.

Sentinel-2 is designed to provide images that can be used to distinguish between different types of crops, as well as data on numerous plant indices, such as leaf area index, leaf chlorophyll content, and plant water content, leaves, all of which are essential to accurately monitor plant mass which is the main factor affecting the vector of expansion

Sentinel 2 has various forms of image processing, we will use level 2A.

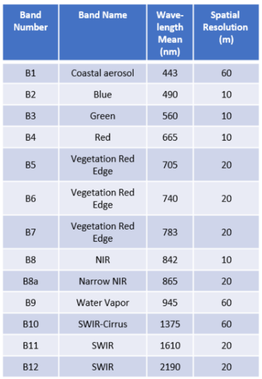
Level 2A processing includes scene classification and atmospheric correction applied to Level 1C orthoimage products of the upper atmosphere (TOA). The main level 2A output is an orthoimage background-of-atmosphere (BOA) corrected reflectance product.

Additional outputs are an Aerosol Optical Thickness (AOT) map, a Water Vapor (WV) map, and a Scene Classification Map (SCM) along with Quality Indicators (QI) for cloud and snow probabilities at a given distance (Table 1). 60m resolution. Level 2A output image products will be resampled and generated with equal spatial resolution for all bands (10 m, 20 m or 60 m). Standard distributed products contain the envelope for all resolutions in three different folders:



Spatial Resolution (m)	Band Number	S2A		S2B	
		Central Wavelength (nm)	Bandwidth (nm)	Central Wavelength (nm)	Bandwidth (nm)
10	2	492.4	66	492.1	66
	3	559.8	36	559.0	36
	4	664.6	31	664.9	31
	8	832.8	106	832.9	106
20	5	704.1	15	703.8	16
	6	740.5	15	739.1	15
	7	782.8	20	779.7	20
	8a	864.7	21	864.0	22
	11	1613.7	91	1610.4	94
	12	2202.4	175	2185.7	185
60	1	442.7	21	442.2	21
	9	945.1	20	943.2	21
	10	1373.5	31	1376.9	30

(a) Table 1 Xpectral bands of Sentinel 2



Band Number	Band Name	Wave-length Mean (nm)	Spatial Resolution (m)
B1	Coastal aerosol	443	60
B2	Blue	490	10
B3	Green	560	10
B4	Red	665	10
B5	Vegetation Red Edge	705	20
B6	Vegetation Red Edge	740	20
B7	Vegetation Red Edge	783	20
B8	NIR	842	10
B8a	Narrow NIR	865	20
B9	Water Vapor	945	60
B10	SWIR-Cirrus	1375	60
B11	SWIR	1610	20
B12	SWIR	2190	20

(b) Table 2 Band's name of Sentinel 2

10 m: contains spectral bands 2, 3, 4, 8, a true color image (TCI), and AOT and WV maps scaled from 20 m.

20 m: contains spectral bands 2 - 7, bands 8A, 11 and 12, a true color image (TCI), a scene classification map (SCL), and an AOT and WV map. The B8 band is omitted because B8A provides more accurate spectral information.

60m: Contains all components of the 20m product resized to 60m, plus bands 1 and 9. Cirrus band 10 is omitted as it does not contain surface information.

The 13 spectral bands of Sentinel-2 (Table 2) range from the Visible (VNIR) and Near Infra-Red (NIR) to the Short Wave Infra-Red (SWIR):

4 x 10 metre Bands: the three classical RGB bands ((Blue (4̃93nm), Green (560nm), and Red (6̃65nm)) and a Near Infra-Red (8̃33nm) band.

6 x 20 metre Bands: 4 narrow Bands in the VNIR vegetation red edge spectral domain (7̃04nm,7̃40nm, 7̃83nm and 8̃65nm) and 2 wider SWIR bands (1̃610nm and 2̃190nm) for applications such as snow/ice/cloud detection, or vegetation moisture stress assessment.

3 x 60 metre Bands mainly focused towards cloud screening and atmospheric correction (4̃43nm for aerosols and 9̃45nm for water vapour) and cirrus detection (1̃374nm). Radiometric resolution is the capacity of the instrument to distinguish differences in light intensity or reflectance. The greater the radiometric resolution, the more accurate the sensed image will be. [9]

Radiometric resolution is routinely expressed as a bit number, typically in the range of 8 to 16 bits. The radiometric resolution of the MSI instrument is 12 bit, enabling the image to be acquired over a range of 0 to 4095 potential light intensity values. citeefsa

As we have images from 2015, we can monitor the province of Alicante since Xf entered in 2017, in the province of Alicante.

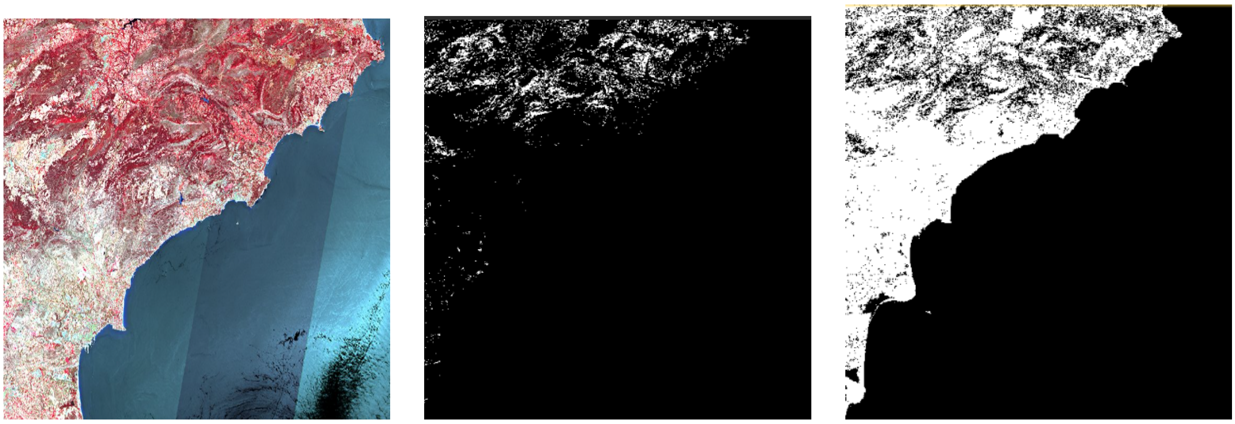
Figure 2 shows the most common shades in a false color composition are:

Magenta red indicates healthy and well-developed vegetation. Intensity and saturation of red allows estimates of the growth cycle of plant areas Pink, less dense plant areas or with less developed vegetation or areas of early growth. White, areas with little or no vegetation. Dark blue or black indicates the presence of water. Metallic gray or blue represents cities or populated areas. Beige or golden transition zones such as dry meadows or scrub.

Here the primary colors red, green and blue have been used to other layers that do not correspond to the wave amplitude that corresponds naturally. But this is very useful in the visual analysis of vegetation, for example, since it sometimes allows us to distinguish by discrimination covers that have an apparent visible behavior but have a different appearance in electromagnetic radiation. For example in this image composition the magenta red represents vigorous vegetation such as irrigated crops, meadows or forests. The study of the intensity and saturation of red allows estimates of the growth cycle of plant areas, even comparing them with images taken previously. The pink color shows areas such as less dense vegetation, or areas of early growth. White represents areas of no vegetation, such as bare ground, sand, clouds, snow. The dark blue-black areas fully or partially covered with water. Metallic gray or blue represents cities or populated areas. The beige or golden transition zones such as dry meadows or scrub.

From which we can extract the areas where we have vegetation, areas where there is vegetation in white and a black background where there is no vegetation, Figure 3 and Figure 4.

Or areas where there is no vegetation upside down, black area where there is vegetation.



(a) Infrared

(b) Vegetation white

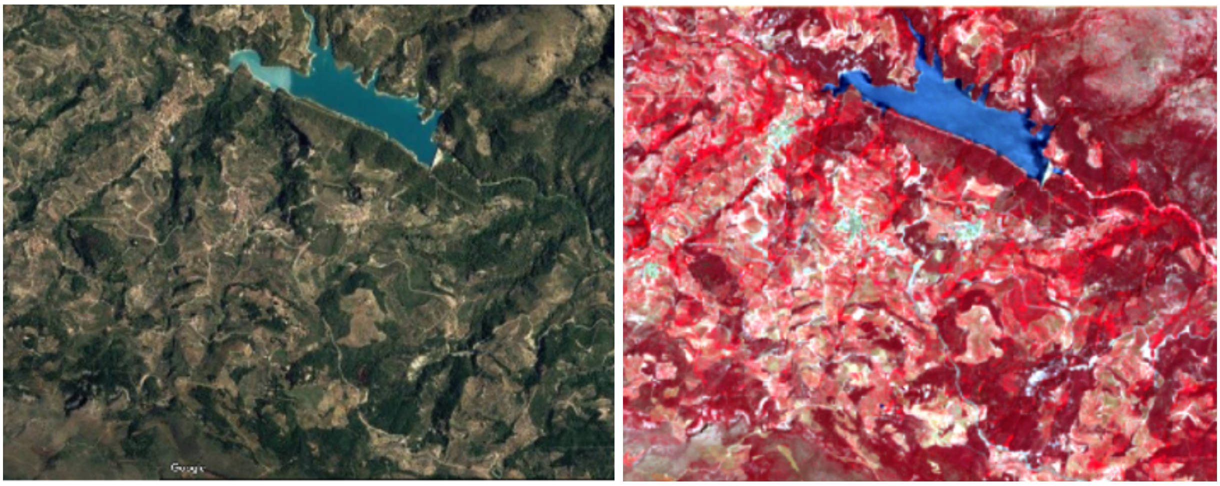
(c) Vegetation black

Here we have another photo taken of the Guadalest reservoir (Figure 5) where the first case of Xf was detected, on the peninsula in 2017. Figure 5 we have an image using the bands that interest us, which are B2, B4 and B8

This does not give information to be able to see how the plant mass evolves, with Google Maps we can also see how the plant mass evolves, but by discriminating frequencies we can separate plant mass from shrubs and plant mass from trees. We have solved this problem with Sentinel-2. This allows us to model through climate models and establish a risk map based on rainfall, temperature, wind, humidity.

With this experiment we can label on the one hand the plant mass for a certain area. Also based on environmental parameters such as we can download from Sentinel-2 photos of the same area from a few years ago, we can monitor the plant mass of a certain place. With what we could see where plant mass is being destroyed and where that plant mass is growing, in a study area. With these data we can extrapolate how the plant mass has evolved over time depending on variables of temperature, rain, wind, etc. As the studies of [6] affirm, we will carry out two studies in the control of plant mass, one in summer and the other in autumn. In addition, within the plant mass [5] we studied two subtypes: the vegetation of the soil of the orchards and the woody areas. We have in Sentinel-2 since 2016 of the entire Valencian Community with which we have a corpus of images that we can predict based on a series of climatic data and a series of photos predict if the plant mass is going to grow or decrease in a certain area . The idea would be to define areas vulnerable to the epidemic and establish a risk map for pest surveillance as provided by the model [7], of treating all plant material in these areas in spring, which would slow down the rate Of reproduction.

These data allows us establishing areas where there are a risk of transmission and monitoring these areas is recommended. This monitoring could be done with Drones as has with this experi-



(a) Natural view (Google) (b) Infrared view

Figure 4: First Xf. in Spain.

ment we can label the plant mass for a certain area. Also based on environmental parameters such as we can download photos of the same area from a few years ago from Google Maps. With what we could see where plant mass is being destroyed and where that plant mass is growing, in a study area.

Based on the mathematical models described, we are making projections of the entire Comunidad Valenciana with which we have a corpus of images that we can predict based on a series of climatic data and a series of photos whether the plant mass will increase or decrease in a certain area. The idea would be to define areas vulnerable to the pest and establish a risk map for pest surveillance as provided by the model [7], of treating all plant material in these areas in spring, which would slow down the rate of reproduction. Where we would establish several colors based on high risk (red), moderate risk (yellow), no changes (white), no risk (green)

3 Results

Working with these bands we have some samples that show how the plant mass evolves depending on climatic variables, temperature, humidity. This allows us to train a multilayer perceptron to provide zones where the plant mass grows or decreases. Figure 4 shows the data flow diagram with citedatos, citeefsa After generating a risk map we mark with colors: red for high risk (plant mass grows), green for low risk (plant mass increases slightly), yellow for medium risk. The white has the same vegetable mass.

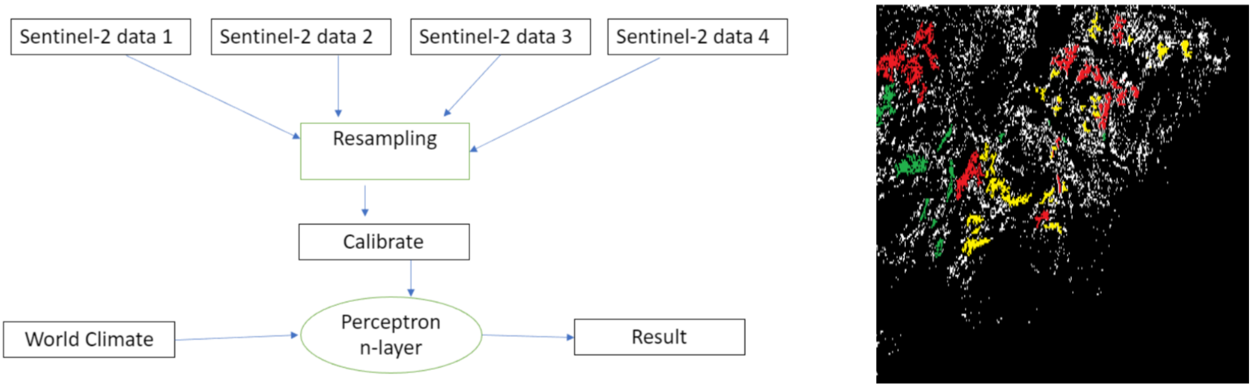
4 Conclusions

By monitoring the plant mass with the help of the photos from Sentinel 2, and the variables Temperature and Humidity, we can establish risk maps.

These areas should be monitored when an increase in plant mass is detected

This work may be improved by increasing the resolution. It would be advisable to monitor the areas to be studied with drones with higher resolution hyperspectral cameras

As a future work we could also foresee reforestation of a fire. The photos taken by Sentinel-2 are free and we can have one every 5 days depending on the cloudiness.



(a) Data flow diagram

(b) Result

Figure 5: Results.

References

[1] Global consequences of land use Science, 309 (2005), pp. 570-574 J.A. Foley, R. DeFries, G.P. Asner, C. Barford, G. Bonan, S.R. Carpenter, F.S. Chapin, M.T. Coe, G.C. Daily, H.K. Gibbs, J.H. Helkowski, T. Holloway, E.A. Howard, C.J. Kucharik, C. Monfreda, J.A. Patz, I.C. Prentice, N. Ramankutty, P.K. Snyder

[2] Biological invasions as global environmental change P.M. Vitousek, C.M. DAntonio, L.L. Loope, R. Westbrooks Am. Sci., 84 1996, pp. 468478

[3] Map of risk de la Xylella, IVIA,2 019, https://www.xylella.es/wp-content/uploads/mapa_xylella_CTexto.png

[4] Almond leaf scorch: leafhopper and spittlebug vectors, Purcell, AH, Journal of Economic Entomology, volume=73, number=6, pages=834–838,1980,Oxford University Press Oxford, UK.

[5] Diversity of vectors and their role in the spread of Xylella fastidiosa in olive orchards of Southeastern Brazil João R. S. Lopes and Joyce A. Froza Luiz de Queiroz College of Agriculture (Esalq), University of São Paulo (USP), Piracicaba, SP, Brazil

[6] New insights on Xylella fastidiosa subsp. pauca vector transmission to olive plants Bodino N.1, Cavalieri V.2 , Almeida R.P.P. 3, Saponari M.2, Dongiovanni C. 4, and Bosco D. 1,5 1 IPSP-CNR, Torino, Italy – 2 IPSP-CNR, Bari, Italy – 3UC Berkeley, US 4CRSFA Locorotondo (Bari) Italy, - 5University of Torino, Italy

[7] Defined a set of integrated tools recommended for IPM strategy to control spittlebugs Dongiovanni C.1, Fumarola G.1, Di Carolo M.1, Altamura G.2, Cavalieri V.2

[8] Climate dates, obtained from www.worldclime.net.

[9] Satellite maps date, obtained from www.efsa.europa.eu.

[10] Spatial Bayesian Modeling Applied to the Surveys of Xylella fastidiosa in Alicante *Spain* and Apulia *Italy* Martina Cendoya , Joaquín Martinez-Minaya, Vicente Dalmau, Amparo Ferrer, Maria Saponari, David Conesa, Antonio López-Quílez and Antonio Vicent 2020

13

New tools for linguistic pattern analysis and specialized text translation: hypergraphs and its derivatives

A. Criado-Alonso[♭],[1] D.Aleja[♮], M.Romance[♮] and R.Criado[♮]

(♭) Grupo LIyNMEDIA, Paseo de los Artilleros 38, 28032-Madrid.
(♮) DCNC Sciences-URJC, Departamento MACIMTE, Universidad Rey Juan Carlos,
C/ Tulipán s/n, 28933-Móstoles (Madrid) .

1 Introduction

One of the main problems in the analysis of complex networks is to find the most important entities (nodes and edges) from the relationships and interactions between them [2, 3]. The existence of interactions of different nature and simultaneous interactions between nodes and edges (v.g., group collaborations, chemical reactions in which more than two components interact, ...) have shown that hypergraphs and multilayer networks are very suitable structures for this type of studies [1, 2]. On the other hand, the latest advances in modern linguistics are based on the treatment of a language as a system or a complex network, to which mathematical tools, statistical measures and procedures of this branch of science can be applied to obtain a new, efficient and effective approach to the study of language [3, 5, 8, 9]. Additionally, the search for linguistic patterns, stylometry and forensic linguistics have in the theory of complex networks, their structures and associated mathematical tools, allies with which to model and analyze texts [4, 5]. Other linguistic aspects and elements, such as the specific terminology of a specialty language and the different combinations of words that give rise to new meanings (called ”collocations”) have also been successfully modeled using an appropriate methodology within the scope of this model [5]. Our idea is to look for a new methodology supported by several mathematical structures in order to analyze the relationships between words, sentences, paragraphs, chapters and texts, focusing on a quantitative concept of dependency between these elements that will be of singular help both to detect the style and level of knowledge of a language by an author, and to create new tools to detect plagiarism. Our motivation is the identification of the main structures and language level in written texts and specialized languages, and as a main motivation, the characterization of the level of competence of the language of a written text and the detection of elements that allow characterizing the style of an author. So, the questions we address are the following:

- How to characterize the competence level of language used in a written text?
- Can an author’s style be determined using specific parameters of a linguistic network associated with one of his written texts?
- What is the combination of words most frequently used in a corpus beyond locating the most relevant individual words?

[1] angeles.criado@urjc.es

- How to determine the most representative words of a text (not necessarily the most frequent)?
- Is it possible to associate a numerical measurement index and a mathematical structure that characterizes and identifies the author?

2 Methods and results

Many questions arise when trying to quantify and to associate a numerical measure index to a text or to a part of such tex. In fact, a breakthrough in linguistics was to be able to associate a complex network structure to a text in order to identify patterns, linguistic units, syntactic and semantic structure ..., so in order to solve this question we will set our sights on some new mathematical structures related to graph theory and higher order networks that will allow us to dissect and analyze this kind of linguistic structures. Therefore, our efforts will be focused on associating a mathematical structure to a written text as if it were a kind of seal of identity of that text. Having in mind that the latest advances in modern linguistics are based on the treatment of a language as a system or a complex network, we will extended these ideas to a general context in which the PageRank algorithm, the mathematical structure of hypergraph and a suitable multilayer line graph will be considered. A linguistic corpus can be mathematically modeled as a multilayer complex network $G = (X, E)$, in such a way that such that each node $X = \{1, \ldots, n\}$ is a word that appears in any of the texts that make up the linguistic corpus, and a direct link is established between two words that appear consecutively [5]. Within the multilayer network, four layers are distinguished (lexical layer, verbal layer, linking layer and remainder layer). The color indicates the type of node (layer). The unit of analysis considered is the sentence, that is, the words enclosed between two periods [7]. Also, it is important to note that commas and other punctuation marks within the sentence have been removed for the analysis done, as well as a previous analysis carried out by linguistic experts to distribute the words into the different layers of the network.

In this study we will focus on the lexical layer. For this purpose, we first take into account that a linguistic corpus can be mathematically modeled as a hypergraph, in which the nodes are the lexical words of the corpus, and each hyper-edge, corresponds to a sentence of a text of the corpus (set of lexical words between two points) and is formed by the lexical words that are part of it. So, the next step is to progress from lexical words to consider mesoscale structures: sentences, paragraphs, chapters,...:

To carry out this study a linguistic corpus composed by 86 extended abstracts and papers (volumes 1-6 of the International Journal of Complex Systems in Science (IJCSS), published between April 2011 and November 2016 (http://www.ij-css.org)), giving a total amount of $25,210$ sentences and $147,637$ words (of which 2,203 are lexical words) have been considered and analyzed.

Let's remember that a hypergraph (or higher order network) is a pair $\mathcal{H} = (X, \varepsilon)$ where:

- X is a finite set (the set of nodes), $X = \{1, ..., N\}$, and
- $\varepsilon = \{h_1, h_2, \ldots, h_n\}$ a collection of subsets of X such that $h_i \neq \emptyset \quad (i = 1, 2, \ldots, n)$ and $X = \bigcup_{i=1}^{n} h_i$. .

Each of these subsets $h_i \neq \emptyset \quad (i = 1, 2, \ldots, n)$ is called a hyperedge of $\mathcal{H} = (X, \varepsilon)$.

In this way, hypergraphs appeared as the natural extensions of complex networks to describe group interactions. In our model, we consider a higher order network (hypergraph) built on the set of lexical words (mainly adjectives and nouns). The hyper-edges may be sentences, paragraphs, chapters....

As we have said, throughout this study, we will consider that the nodes of the hypergraph are the lexical words of our written texts (or corpus) and the hyper-edges are the sentences, i.e., each sentence is the set of lexical words located between two periods. Now, in order to measuring similarity and dissimilarity of texts, first we consider how to analyze the similarity of sets. The basic Jaccard index to compare the degree of coincidence or similarity between two sets A and B can be obtained from the formula

$$\mathcal{J}(\mathcal{A},\mathcal{B}) = \frac{|A \cap B|}{|A \cup B|} \quad 0 \leq \mathcal{J}(\mathcal{A},\mathcal{B}) \leq 1.$$

In this regard, different generalizations of Jaccard index have been introduced in the literature (including the overlap index):

$$\mathcal{J}_n(A,B) = \frac{|A \cap B|^n}{|A \cup B|},\ \mathcal{I}(A,B) = \frac{|A \cap B|}{min\{|A|,|B|\}}$$

and $\mathcal{C}(A,B) = \mathcal{J}(\mathcal{A},\mathcal{B}) \cdot \mathcal{I}(\mathcal{A},\mathcal{B})$. Observe that $0 \leq \mathcal{C}(\mathcal{A},\mathcal{B}) \leq 1$.

Let us now recall some other definitions of the different elements and structures involved in our model. Another fundamental ingredient for carrying out this study is the PageRank algorithm. We consider in our model the PageRank algorithm and the Random Walkers' Heuristics. In that sense, it is clear that if we surf at random on a network, the more frequently we pass over a node or an edge, the more relevant that element is. So we consider a Markov chain on the network such that

- At time $t = 0$ we choose one node at random.
- If at time t the Markov chain is at node j, then we move at random (uniformly) and at time $t+1$ we move to one of the neighbors of j, i.e. the probability of moving from j to i is
$$p_{ji} = \frac{a_{ji}}{gr_{out}(j)}.$$
- In order to avoid sink nodes, at each step there is a small probability to make a random jump outside of the neighborhood.

This Markov chain has the following transition matrix

$$\phi_{ij} = (1-q)\frac{a_{ji}}{gr_{out}(j)} + \frac{q}{n},$$

where the frequency of reaching node i is the i-component of the stationary state of ϕ_{ij}, the stationary state is a positive eigenvector of ϕ_{ij} and, since $\phi_{ij} > 0$, this eigenvector is unique, up to normalization (Perron-Frobenius theorem). According to all of this, we consider the PageRank centrality, defined as the stationary state of the Markov chain with transition matrix

$$\phi_{ij} = (1-q)\frac{a_{ji}}{gr_{out}(j)} + qu_j,$$

with $u = (u_1, \cdots, u_n) \geq 0$ such that $\|u\|_1 = 1$, is a celebrated centrality measure used in many applications, such as the web pages ranking (Google), bibliographic and scientific collaboration networks, transportation systems, cybersecurity and many others. Finally, another ingredients related to this study are the concepts of of linegraph and dual graph of a hypergraph. If $\mathcal{H} = (X, \varepsilon)$ is a hypergraph, the linegraph associated to $\mathcal{H}$ is the graph $L(\mathcal{H}) = (\varepsilon, E')$, where if $h_i, h_j \in \varepsilon$, then

$$\{h_i, h_j\} \in E' \Leftrightarrow h_i \cap h_j \neq \emptyset.$$

It is also notorious that the linegraph $L(\mathcal{H})$ of a hypergraph $\mathcal{H}$ is a graph even though $\mathcal{H}$ is a hypergraph. Note that this concept is a particular case of the concept of intersection graph. If

$\mathcal{H} = (X, \varepsilon)$ is a hypergraph, the dual hipergraph associated with $\mathcal{H}$ is the hypergraph $\mathcal{H}^* = (\varepsilon, X')$ in such a way that if $X = \{1, ..., N\}$, then $X' = \{v_1, ..., v_N\}$ where $v_i = \{h_j | i \in h_j\}$, $i = 1, ..., N$. It is not difficult to verify that $(\mathcal{H}^*)^* = \mathcal{H}$. So, for any hypergraph $\mathcal{H}$ we have that $\Pi_2(\mathcal{H}^*) = L(\mathcal{H})$. Moreover, if I is the incidence matrix of $\mathcal{H}$, then its transpose matrix I^t is the incidence matrix of $\mathcal{H}^*$. In fact:

$$(I(\mathcal{H})^t) \cdot (I(\mathcal{H})) = \widetilde{A(\mathcal{H})} = (\widetilde{a_{ij}}) \in \mathbb{R}^{|\varepsilon| \times |\varepsilon|}$$

and

$$(I(\mathcal{H})) \cdot (I(\mathcal{H})^t) = A(\mathcal{H}) = (a_{ij}) \in \mathbb{R}^{N \times N},$$

where

$$a_{ij} = \begin{cases} |\{h \in \varepsilon \mid \ i \in h\}| & \text{if} \quad i = j, \\ |\{h \in \varepsilon \mid \ i, j \in h\}| & \text{if} \quad i \neq j. \end{cases} \tag{1}$$

It is important to highlight that nowadays there is a growing interest in all these structures.

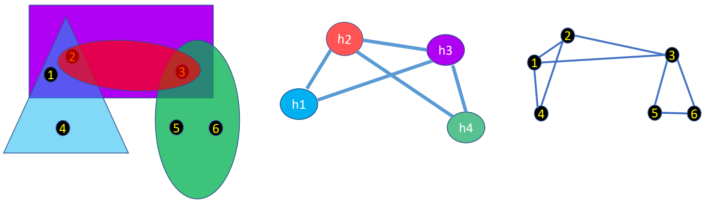


Figure 1: An example: $\mathcal{H}$, $L(\mathcal{H}) = \Pi_2(\mathcal{H}^*)$ and $\Pi_2(\mathcal{H})$.

At this point, we can give the following definition:

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*

$$\frac{\partial \mathcal{H}}{\partial\{i,j\}} = \frac{a_i - a_{ij} + a_j - a_{ij}}{a_{ij}} = \frac{a_i - 2a_{ij} + a_j}{a_{ij}}. \tag{2}$$

So, looking at Figure 1 we get $\frac{\partial \mathcal{H}}{\partial\{2,3\}} = \frac{3-2+3-2}{2} = 1$, $\frac{\partial \mathcal{H}}{\partial\{5,6\}} = 0$, $\frac{\partial \mathcal{H}}{\partial\{1,2\}} = \frac{1}{2}$, $\frac{\partial \mathcal{H}}{\partial\{1,3\}} = 3$, $\frac{\partial \mathcal{H}}{\partial\{2,4\}} = 2$, $\frac{\partial \mathcal{H}}{\partial\{1,4\}} = 1$, $\frac{\partial \mathcal{H}}{\partial\{3,5\}} = 2$, and $\frac{\partial \mathcal{H}}{\partial\{3,4\}} = +\infty = \frac{\partial \mathcal{H}}{\partial\{4,5\}} = \frac{\partial \mathcal{H}}{\partial\{2,6\}}$.

From these values, the derivative graph $\partial\mathcal{H}$ and the homogeneity graph $HG(\mathcal{H})$ can be constructed [6] (see Figure 2), having in mind that $\partial\mathcal{H}$ is the weighted graph obtained by considering the derivative of $\mathcal{H}$ with respect all the pairs of nodes $i, j \in X$, and by setting $\forall i, j \in X$ the corresponding numerical value of $\frac{\partial \mathcal{H}}{\partial\{i,j\}}$ on the edge $\{i, j\}$, in such a way that if $\frac{\partial \mathcal{H}}{\partial\{i,j\}} = 0$, then the nodes i and j collapse into a single node (ij), and having in mind that if $\frac{\partial \mathcal{H}}{\partial\{i,j\}} = \infty$, then the edge $\{i, j\}$ does not exist in the derivative graph, and the homogeneity graph $HG(\mathcal{H})$ of $\mathcal{H}$ is the weighted graph with the same nodes and edges as $\partial\mathcal{H}$, but considering as the weight of each edge the inverse value of the weight corresponding to the derived graph $\partial\mathcal{H}$.

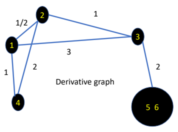


Figure 2: Derivative graph $\partial\mathcal{H}$ and Homogeneity graph $HG(\mathcal{H})$ of the analyzed example.

For the same reason, when analyzing the hypergraph $\mathcal{H}$ formed by all the sentences of the corpus under study, we obtained 82 pairs of words that appear in exactly the same sentences. Thus, for example

$$\frac{\partial\mathcal{H}}{\partial\{Monte, Carlo\}} = \frac{\partial\mathcal{H}}{\partial\{differential, RungeKuttta\}} = \frac{\partial\mathcal{H}}{\partial\{oscillatory, asynchronous\}} = 0.$$

$$\frac{\partial\mathcal{H}}{\partial\{network, system\}} = 16.33, \frac{\partial\mathcal{H}}{\partial\{language, formal\}} = \frac{\partial\mathcal{H}}{\partial\{connectance, asymetry\}} = 2.$$

Some computations have been made on our selected corpus composed of 25210 sentences (hyperedges). The numerical experiments were run on a iMac18,3 with 4,2 GHz Intel Core i7 and RAM 16 GB by using a Python 3.7 implementation with machine precision $\varepsilon \approx 2.22 \times 10^{-16}$. In fact, we consider three different rankings to be applied on the set of lexical words of the considered corpus:

To calculate **Ranking 1**. we first have built a graph on which to apply the PageRank algorithm. In order to do that, we convert each hyperedge of $\mathcal{H}$ into a clique to obtain the projection graph $\Pi_2(\mathcal{H})$. After this, taking into account that the average number of words of a sentence within the corpus under study is 5.97 and that, therefore, the local lexical density is 5.97 ,we can deduce that the damping factor corresponding to this configuration is 0.85, since

$$\begin{aligned} 5.97 =& \mathbb{E}(\ell) = \sum_{k=0}^{\infty} k \cdot \mathbb{P}(\ell = k) = \sum_{k=1}^{\infty} k \cdot (1-q) \cdot q^k \\ =& (1-q) \cdot q \sum_{k=1}^{\infty} k \cdot q^{k-1} = \frac{q}{1-q}. \end{aligned}$$

To calculate the **Ranking 2**, we will apply the PageRank algorithm considered on the network

$$\Pi_2(\mathcal{H}^*) = L(\mathcal{H})$$

so that, once the numerical value attributed to each phrase has been obtained, this value is distributed proportionally among the words that make up that sentence. Now, using the same reasoning as in the previous case, and having in mind that the average number of sentences of a paper included in the corpus under study is 27.12, in this context, the damping factor corresponding to this configuration is 0.96. Finally, to calculate **Ranking 3**, we apply the PageRank algorithm considered on the weighted graph $HG(\mathcal{H})$. Taking into account that the average number of words of a sentence is 5.63 (since, after collapsing words pairs $\{i, j\}$ such that $\frac{\partial\mathcal{H}}{\partial\{i,j\}} = 0$, the average length of sentences decreases, albeit slightly), the damping factor corresponding to this configuration is 0.84. And these are the rankings produced using these three criteria:

	Ranking 1	Ranking 2	Ranking 3
1st	network	network	network
2nd	system	system	system
3nd	model	model	model
4th	complex	complex	complex
5th	process	number	graph
6th	number	process	process
7th	information	structure	structure
8th	graph	new	information
9th	new	information	number
10th	structure	distribution	new

Table 1: Rankings of lexical words

3 Conclusions

By means of conclusions, we can point out the following:

- We introduce and study the derivative of a hypergraph and the homogeneity graph of a hypergraph as new and useful structures that can be used to study the degree of independence of the nodes of a hypergraph as well as to obtain a ranking of the most representative nodes of the hypergraph.
- The associate parameters to the homogeneity graph of a hypergraph allow us to obtain technical characteristics related to the style of different authors and the language competence level of any text written in English.
- These new concepts are developing and being implemented to configurate a tool for searching similarities and differences in a set of texts. The automatic classification of texts according to their differences or similarities, as well as its different applications to text classification, text summarization, stylometry and authorship detection are some of the possible applications of the methodology underlying this model.

References

[1] Benson, A., Three Hypergraph Eigenvector Centralities, SIAM J. MATH D.S., 1(2): 293–312, 2019.

[2] Boccaletti S., Bianconi G., Criado R., Del Genio C.I., Gómez-Gardeñes J., Romance M., Sendiña-Nadal I., Wang Z., Zanin, M., The structure and dynamics of multilayer networks *Phys. Rep.* , 544 (1):1–122, 2014.

[3] Cong, J. Liu, H : Approaching human language with complex networks *Phys. of Life Rev.*, 11–4, 2014.

[4] Criado-Alonso, A., Battaner-Moro, E., Aleja, D., Romance, M., Criado, R.: Using complex networks to identify patterns in specialty mathematical language: a new approach. *Social Network Analysis and Mining* 10 (1), 1-10, 2020.

[5] Criado-Alonso, A., Battaner-Moro, E., Aleja, D., Romance, M., Criado, R. Enriched line graph: A new structure for searching language collocations *Chaos Solitons & Fractals*, 142(1):110509, 2021.

[6] Criado-Alonso, A., Aleja, D., Romance, M., Criado, R. Derivative of a hypergraph as a tool for linguistic pattern analysis *https://arxiv.org/pdf/2207.09400.pdf*, 2022.

[7] Halliday, M.A.K., Matthiessen, C.M.I.M., Introduction to Functional Grammar (Third edition), Routledge, Taylor & Francis Group, London and New York (2004).

[8] Liu, H., Hu, F.: What role does syntax play in a language network? *EPL (Europhysics Letters)* 83, 18002, 2008.

[9] Mehler, A., Lűcking, A., Banisch, S., Blanchard, P., Frank-Job, B. (Eds.), Towards a Theoretical Framework for Analyzing complex Linguistics Networks, Springer-Verlag, 2016.

14

On an accurate method to compute the matrix logarithm

E. Defez$^{\star 1}$, J.J. Ibáñez*, J. M. Alonso$^{\flat}$ and J.R. Herráiz*

($\star$) Instituto de Matemática Multidisciplinar,
($\flat$) Instituto de Instrumentación para Imagen Molecular,
Universitat Politècnica de València,
Camino de Vera, s/n, 46022 Valencia, Spain.

1 Introduction

The problem of computing a matrix function is a problem interesting and difficult at the same time. On the one hand, it is difficult due to the need to use matrices, which can be of large dimension, where accurate results are to be achieved with the lowest computational cost. On the other hand, it is interesting because the great variety of different matrix functions (exponential, sign, trigonometric, hyperbolic, etc.) to be computed, each of which has its own problematic, and its multiple applications in different areas of applied mathematics, science and engineering. Among these matrix functions, the matrix exponential highlights, together with logarithm function.

Given a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ satisfying that $\sigma(A) \subset \mathbb{C} - (-\infty, 0\ ]$, we denote as a matrix logarithm of A to any matrix $X \in \mathbb{C}^{n\times n}$ fulfilling the matrix equation

$$A = e^X, \tag{1}$$

where e^X is the matrix exponential of X. There are infinite solutions of equation (1), but we only will focus on the *principal matrix logarithm* or the standard branch of the logarithm, denoted by $\log(A)$, which is the unique one of matrix A, see Theorem 1.31 of [1], whose eigenvalues all lie in the strip $\{z \in \mathbb{C}; -\pi < Im(z) < \pi\}$. The matrix logarithm is used in engineering [2], optics [3], mechanics [4], the study of viscoelastic fluids [5], control theory [6], graph theory [7], statistics and data analysis [8], among other many areas.

Several algorithms have been provided for computing this matrix function. The proposed methods include, for example, the inverse scaling and squaring (ISS) technique [9], an algorithm based on the arithmetic-geometric mean iteration [10], or the use of contour integrals [11]. In [12], a method based on Schur–Fréchet algorithm was ed. In references [13–17], different authors studied the use of Padé approximations with distinct and interesting variants. Recently, a method based on the Taylor series was presented in [18]. Also, methods based on different quadrature formulas were proposed in [19, 20].

[1] edefez@imm.upv.es

2 The original method proposed

In [19], the following integral representation of the matrix logarithm is given:

$$\log{(A)} = (A - I) \int_0^1 ((A - I)t + I)^{-1} \, dt. \tag{2}$$

Carrying out the change of variable $t = \dfrac{1+u}{2}$, we can write (2) in the form:

$$\log{(A)} = (A - I) \int_{-1}^1 ((A - I)(1+u) + 2I)^{-1} \, du. \tag{3}$$

In [20], integral of (3) was approximated using the Gauss-Legendre method, that it is a well known form of Gaussian quadrature for approximating the definite integral of a function. Considering the integration interval $[-1, 1]$, the Gauss–Legendre rule takes the form:

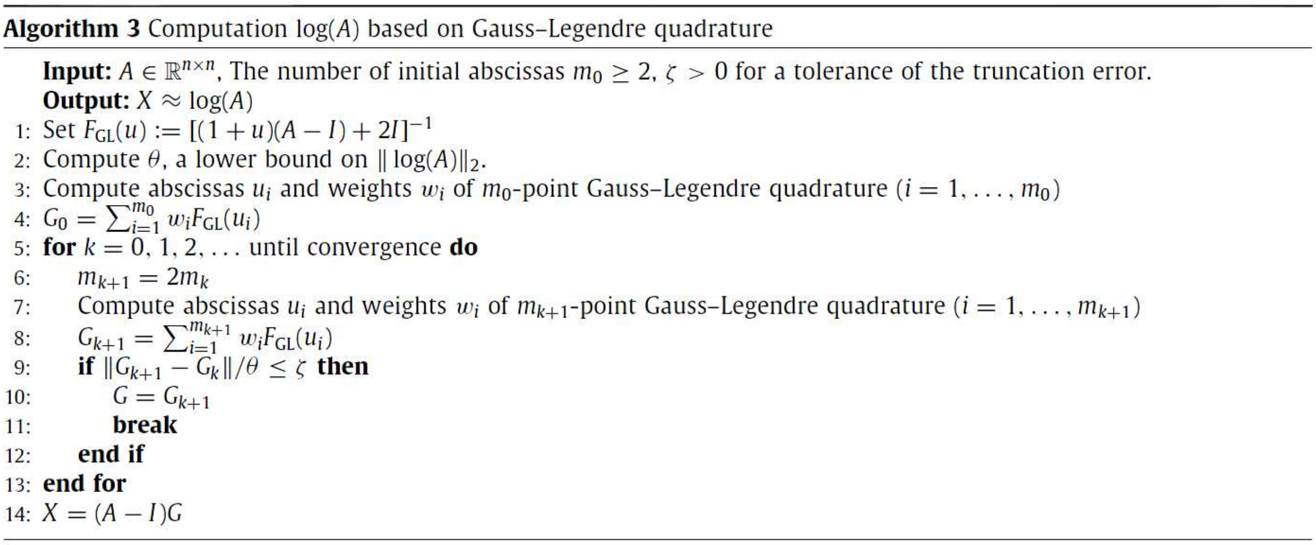
$$\int_{-1}^1 f(x)dx \approx \sum_{i=1}^n w_i f(u_i), \tag{4}$$

where n is the number of sample points used in the quadrature, u are the quadrature nodes, given by the roots of the n-th Legendre polynomial $P_n(u)$, and w are quadrature weights, defined by

$$w_i = \frac{2}{(1-u_i^2)\left(P_n'(u_i)\right)^2}, i = 1, 2, \ldots, n.$$

This choice of weights w and nodes u is the only one that allows the quadrature rule to integrate exactly polynomials of order $2n - 1$.

In [20], authors presented the Algorithm 3 designed to compute $\log{(A)}$ by automatically adding abscissas until the error is smaller than a given tolerance. This algorithm is reproduced in the Figure 1.



Algorithm 3 Computation log(A) based on Gauss–Legendre quadrature

```
Input: $A \in \mathbb{R}^{n \times n}$, The number of initial abscissas $m_0 \geq 2$, $\zeta > 0$ for a tolerance of the truncation error.
Output: $X \approx \log(A)$
Set $F_{\mathrm{GL}}(u) := [(1+u)(A-I)+2I]^{-1}$
Compute $\theta$, a lower bound on $\|\log(A)\|_2$.
Compute abscissas $u_i$ and weights $w_i$ of $m_0$-point Gauss–Legendre quadrature ($i = 1, \ldots, m_0$)
$G_0 = \sum_{i=1}^{m_0} w_i F_{\mathrm{GL}}(u_i)$
for $k = 0, 1, 2, \ldots$ until convergence do
    $m_{k+1} = 2m_k$
    Compute abscissas $u_i$ and weights $w_i$ of $m_{k+1}$-point Gauss–Legendre quadrature ($i = 1, \ldots, m_{k+1}$)
    $G_{k+1} = \sum_{i=1}^{m_{k+1}} w_i F_{\mathrm{GL}}(u_i)$
    if $\|G_{k+1} - G_k\|/\theta \leq \zeta$ then
        $G = G_{k+1}$
        break
    end if
end for
$X = (A - I)G$
```

Figure 1: Computation log(A) based on Gaus Legnedre quadrature [20, Algorithm 3].

The proposed algorithm is clearly based on the scalar one, and it has the problem that a matrix inverse must be calculated in each evaluation of the function to integrate. While in the scalar case the inverse is just a simple division, in the matrix case it is a very expensive operation, with a computational cost of $O(8r^3/3)$ flops, where r is the dimension of the matrix. Moreover, none of the previous calculations can be reused in the algorithm proposed.

A short trace is shown below to clarify the inefficiency of this Algorithm 3. Let us start it with $m_0 = 2$ and see how many inverses are needed if we simply perform 3 iterations:

- Step 1. To obtain G_0, the quadrature nodes $u_1 = -0.57735$, and $u_2 = 0.57735$ are needed. Thus, the matrices $F_{GL}\left(-0.57735\right)$, and $F_{GL}\left(0.57735\right)$ are evaluated, for which two inverse matrices are computed.

- Step 2. To evaluate G_1, the quadrature nodes $u_1 = -0.861136$, $u_2 = -0.339981$, $u_3 = 0.339981$, and $u_4 = 0.861136$ are employed. Note that these nodes are all different from the ones used in the previous step. Thus, the matrices $F_{GL}\left(-0.861136\right)$, $F_{GL}\left(-0.339981\right)$, $F_{GL}\left(0.339981\right)$, and $F_{GL}\left(0.861136\right)$ are evaluated by calculating four different inverse matrices more.

- Step 3. To provide G_2, we need the quadrature nodes $u_1 = -0.96029$, $u_2 = -0.796666$, $u_3 = -0.525532$, $u_4 = -0.183435$, $u_5 = 0.183435$, $u_6 = 0.525532$, $u_7 = 0.796666$, and $u_8 = 0.96029$, all of them distinct from the previous ones. Then, we compute the matrices $F_{GL}\left(-0.96029\right)$, $F_{GL}\left(-0.796666\right)$, $F_{GL}\left(-0.525532\right)$, $F_{GL}\left(-0.183435\right)$, $F_{GL}\left(0.183435\right)$, $F_{GL}\left(0.525532\right)$, $F_{GL}\left(0.796666\right)$, and $F_{GL}\left(0.96029\right)$, after working out eight additional inverse matrices.

- Step 4. To evaluate G_3, the quadrature nodes to be considered are $u_1 = -0.989401$, $u_2 = -0.944575$, $u_3 = -0.865631$, $u_4 = -0.755404$,$u_5 = -0.617876$, $u_6 = -0.458017$, $u_7 = -0.281604$, $u_8 = -0.0950125$, $u_9 = 0.0950125$, $u_{10} = 0.281604$, $u_{11} = 0.458017$, $u_{12} = 0.617876$, $u_{13} = 0.755404$, $u_{14} = 0.865631$, $u_{15} = 0.944575$, are $u_{16} = 0.989401$. Again no node is repeated. Thus, we evaluate the matrices $F_{GL}\left(-0.989401\right)$, $F_{GL}\left(-0.944575\right)$, $F_{GL}\left(-0.865631\right)$, $F_{GL}\left(-0.755404\right)$, $F_{GL}\left(-0.617876\right)$, $F_{GL}\left(-0.458017\right)$, $F_{GL}\left(-0.281604\right)$, $F_{GL}\left(-0.0950125\right)$, $F_{GL}\left(0.0950125\right)$, $F_{GL}\left(0.281604\right)$, $F_{GL}\left(0.458017\right)$, $F_{GL}\left(0.617876\right)$, $F_{GL}\left(0.755404\right)$, $F_{GL}\left(0.865631\right)$, $F_{GL}\left(0.944575\right)$, and $F_{GL}\left(0.989401\right)$ and sixteen different inverse matrices more will have been computed.

If we assume the given result in this Step 4 is sufficiently accurate to finish Algorithm 3, we would use the 16 computed matrices to calculate the matrix logarithm.

In this way, we can see that a total of 30 different inverse matrices have been calculated, but then only 16 have been used to calculate the sought matrix logarithm. Since slightly less than half of the calculated matrices are not used in the final approximation, we can conclude that this algorithm proposed in [20] is not computationally efficient.

In addition, it is usual in the matrix logarithm approximation to use a scaling factor s that follows from the well-known equality $\log\left(A\right) = 2^s \log\left(A^{2^{-s}}\right)$. This step has not been considered in the previous described algorithm, but if it is considered in most of the other ones proposed in the state–of–art. For example, *MATLAB* function *logm(A)* uses the algorithm described in [9,12], composed of three different stages. In the first phase, an integer s so that matrix $A^{2^{-s}}$ is "*near*" to the identity matrix I is calculated. To compute the matrix square roots, the scaled product form of the Denman–Beavers iteration [1, eq. (6.29)] is used. In the second phase, the value of $\log\left(A^{2^{-s}}\right)$ is approximated by $r_m(A^{2^{-s}})$, where $r_m(x)$ is the diagonal Padé approximant of function $\log\left(1+x\right)$

of degree m. Finally, in the last phase, $\log(A)$ is computed as $\log(A) = 2^s \log(A^{2^{-s}})$.

The purpose of this work is to introduce a new method, based on the Gauss–Legendre quadrature, that is more computationally efficient and that provides accurate results in the approximation of the matrix logarithm.

3 The proposed alternative algorithm

Algorithm 1 computes the logarithm of a matrix A by using the Gauss-Legendre method described in Section 2. The following is a description of the main stages that comprise the algorithm:

1. Determining an integer s so that matrix $A^{2^{-s}}$ is "*near*" to the identity matrix I. This is achieved in Steps from 1 to 7, where A is brought close to I by taking successive square roots of itself until its norm is lower than ε. To compute the required matrix square roots, the scaled Denman–Beavers iteration [1, eq. (6.29)] is used. This iteration is defined by:

$$\begin{aligned} X_{k+1} &= \frac{1}{2}(\mu_k X_k + \mu_k^{-1} Y_k^{-1}), \\ Y_{k+1} &= \frac{1}{2}(\mu_k Y_k + \mu_k^{-1} X_k^{-1}), \end{aligned} \tag{5}$$

   with $X_0 = A$, $Y_0 = I$ and $\mu_k = |\det(X_k)\det(Y_k)|^{-1/(2n)}$, and n is the dimension of matrix A. Although convergence is not guaranteed, if it does then X_k converges to $A^{1/2}$.

2. Computing $\log(A^{2^{-s}})$ by means of the Gauss-Legendre quadrature with a FIXED number of nodes (Step 8).

3. Recovering $\log(A) = 2^s \log(A^{2^{-s}})$ (Step 9).

Algorithm 1 Given a matrix $A \in \mathbb{C}^{r\times r}$, the number of nodes n, a tolerance ε, and a maximum number of iterations *maxiter*, this algorithm computes $L = \log(A)$ by the n-point Gauss-Legendre quadrature rule (4).

```
1: $s = 1$
2: $finish = ||A - I|| \le \varepsilon$
3: while not $finish$ and $s \le maxiter$ do
4:     $A \leftarrow A^{1/2}$, using scaled DB iteration (5)
5:     $finish = ||A - I|| \le \varepsilon$
6:     $s = s + 1$
7: end while
8: Compute $L$ by using (4).
9: $L \leftarrow 2^s L$
```

What is the optimal number n of nodes to be taken into account in the logarithm computation?. Let us remind that to guarantee the result of the quadrature is exact, the number n of nodes to use with Gauss-Legendre must be greater than or equal to $(m+1)/2$, where m is the degree of the polynomial function to be integrated. If we consider that m is the Taylor polynomial order that is used to approach the matrix logarithm, we could approximate and substitute the integrand by analogy with a polynomial of degree $m-1$. Thus, the number of nodes n could be estimated as $m/2$.

The Table 6 of Ref. [18] shows that the average degree of the Taylor polynomials used with the function *logm_pol* in the three numerical experiments described there is $42, 45$, and 39, which gives an arithmetic mean of 42. In this way, if we take $m = 42$, we get that $n = 42/2 = 21$. Moreover, it has been experimentally proven that $n = 20$ is the value that delivers the most accurate results in our executions, which practically coincides with the one previously provided. Therefore, $n = 20$ will be the fixed number of nodes in our implementation of the Gauss-Legendre quadrature formula.

The function *logm_gauss_sqr(A)* has been coded in MATLAB language, implementing our proposed Gauss-Legendre algorithm.

4 Numerical Experiments

Obviously, we need to check and compare the effectiveness, and accuracy of our algorithm, and its corresponding codification in the function *logm_gauss_sqr(A)*, with other existing implementations. For this purpose, the following state-of-the-art codes have been considered:

- *logm_new.* Initially, this function transforms the input matrix A to the Schur triangular form $A = QTQ^{\star}$. Then, the logarithm of the triangular matrix T is computed by using the inverse scaling and squaring technique and the Padé approximation. The square roots of matrix T are worked out by the Björck and Hammarling algorithm [1, eq. (6.3)]. It corresponds to the Algorithm 4.1 of [9].

- *logm_iss_full.* In this code, the matrix logarithm is computed by means of the transformation-free form of the inverse scaling and squaring method with the Padé approximation. In this case, the matrix square roots are calculated by the product form of the Denman–Beavers iteration. It corresponds to the Algorithm 5.2 of [9].

In our numerical experiments, a total of 225 heterogeneous matrices were employed, spread over three distinct sets respectively composed of 99 diagonalizable matrices (Set 1), 99 non-diagonalizable matrices (Set 2), and 18 matrices from the Matrix Computation Toolbox [22] and 9 from the Eigtool MATLAB Package [23] (Set 3). The size of all these matrices was 128 × 128, performing all calculations with MATLAB 2020b.

The results obtained in 3 numerical experiments, each corresponding to a distinct set of matrices, are discussed below. Firstly, for each code in comparison and respectively for each set of matrices, Figures 2, 3, and 4 show the graphs corresponding to the performance profile, the normwise relative error committed by each function, and the ratio of the relative error incurred by *logm_gauss_sqr(A)* with respect to *logm_new* and *logm_iss_full.* In terms of accuracy of results, our method exhibited a clear superiority over the other two ones for Sets 1 and 3, being more modest for Set 2. Next, Tables 1, 3, and 5 contain the percentage of cases in which the function *logm_gauss_sqr(A)* provided a relative error less than, greater than or equal to that of the functions *logm_new* and *logm_iss_full.* As can be seen, *logm_gauss_sqr(A)* outperformed the other 2 codes in a very considerable percentage of occasions, regardless of the set of matrices analyzed. Finally, Tables 2, 4, and 6 incorporate the elapsed times in the calculation of all the matrices that comprise each set. Depending on the set involved, the execution time of function *logm_gauss_sqr(A)* may be longer, shorter or intermediate to that of the other codes in comparison.

Numerical test 1			
$E(logm\_gauss\_sqr) < E(logm\_new)$	88.89%	$E(logm\_gauss\_sqr) < E(logm\_iss\_full)$	100.00%
$E(logm\_gauss\_sqr) > E(logm\_new)$	11.11%	$E(logm\_gauss\_sqr) > E(logm\_iss\_full)$	0.00%
$E(logm\_gauss\_sqr) = E(logm\_new)$	0.00%	$E(logm\_gauss\_sqr) = E(logm\_iss\_full)$	0.00%

Table 1: Improvement percentage in the normwise relative error committed by all the codes for Set 1.

Execution time for Set 1	
$logm\_gauss\_sqr$	7.54
$logm\_new$	12.24
$logm\_iss\_full$	4.24

Table 2: Execution time, in seconds, for matrices from Set 1.

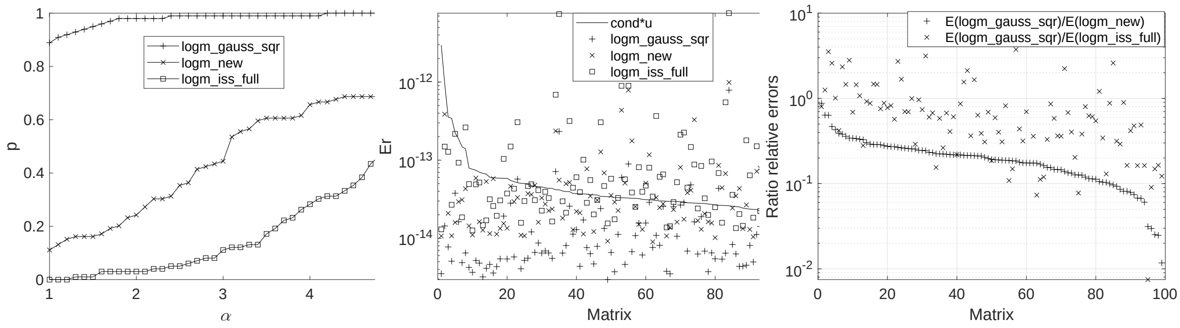


Figure 2: Profile, normwise relative errors, and ratio of relative errors for Set 1.

Numerical test 2			
$E(logm\_gauss\_sqr) < E(logm\_new)$	63.64%	$E(logm\_gauss\_sqr) < E(logm\_iss\_full)$	100.00%
$E(logm\_gauss\_sqr) > E(logm\_new)$	36.36%	$E(logm\_gauss\_sqr) > E(logm\_iss\_full)$	0.00%
$E(logm\_gauss\_sqr) = E(logm\_new)$	0.00%	$E(logm\_gauss\_sqr) = E(logm\_iss\_full)$	0.00%

Table 3: Improvement percentage in the normwise relative error committed by all the codes for Set 2.

Execution time for Set 2	
$logm\_gauss\_sqr$	14.44
$logm\_new$	11.21
$logm\_iss\_full$	8.56

Table 4: Execution time, in seconds, for matrices from Set 2.

5 Conclusions

In this work, a new algorithm based on the Gauss-Legendre quadrature formula has been presented to calculate the matrix logarithm. This algorithm have been tested on a battery of test

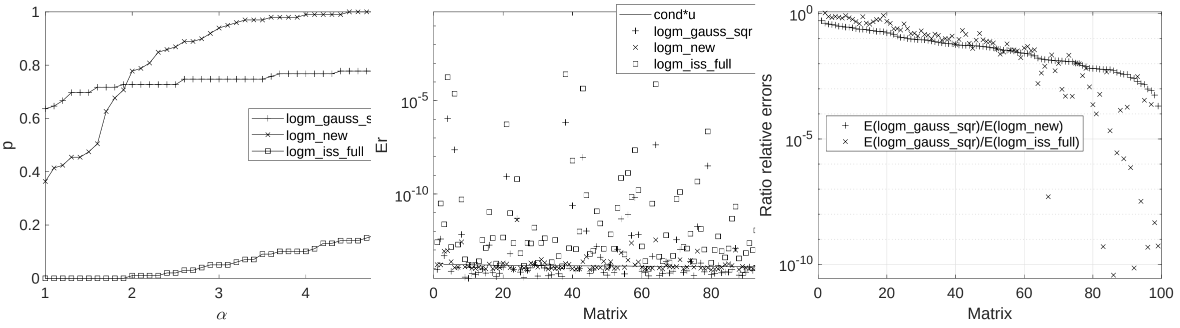


Figure 3: Profile, normwise relative errors, and ratio of relative errors for Set 2.

Numerical test 3			
$E(logm\_gauss\_sqr) < E(logm\_new)$	74.07%	$E(logm\_gauss\_sqr) < E(logm\_iss\_full)$	77.78%
$E(logm\_gauss\_sqr) > E(logm\_new)$	25.93%	$E(logm\_gauss\_sqr) > E(logm\_iss\_full)$	22.22%
$E(logm\_gauss\_sqr) = E(logm\_new)$	0.00%	$E(logm\_gauss\_sqr) = E(logm\_iss\_full)$	0.00%

Table 5: Improvement percentage in the normwise relative error committed by all the codes for Set 3.

Execution time for Set 3	
$logm\_gauss\_sqr$	0.02
$logm\_new$	0.07
$logm\_iss\_full$	0.10

Table 6: Execution time, in seconds, for matrices from Set 3.

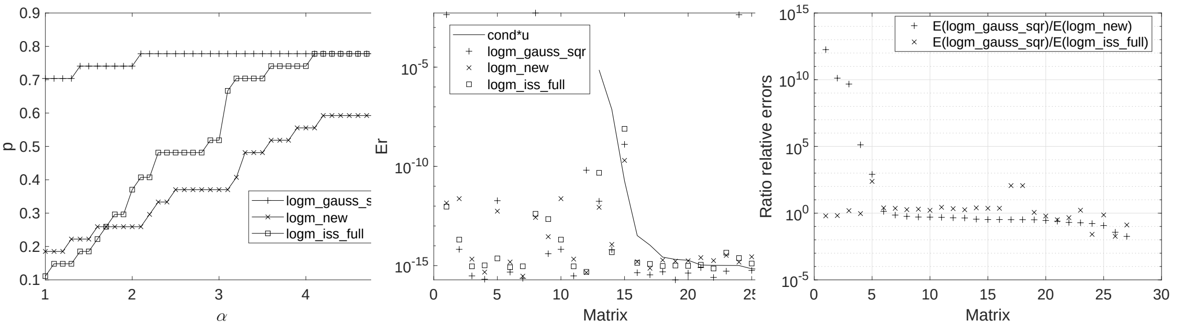


Figure 4: Profile, normwise relative errors, and ratio of relative errors for Set 3.

matrices in order to select to study their computational cost and accuracy. Taking into account the experimental results, it can concluded that:

- The normwise relative error figure indicates that the proposed algorithm is as numerically stable as the other known methods.
- In general, the proposed method *logm_gauss_sqr* has proven to be more accurate when

compared to other very well-known methods in the scientific literature, such as *logm_new* and *logm_iss_full*.

Acknowledgements

This work has been supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València (PAID-11-21).

References

[1] Higham, N. J., Functions of Matrices: Theory and Computation. Philadelphia, PA, USA, Society for Industrial and Applied Mathematics, 2008.

[2] Miyajima, S., Verified computation for the matrix principal logarithm *Linear Algebra and its Applications*, 569:38–61, 2019.

[3] Ossikovski, R., De Martino, A., Differential Mueller matrix of a depolarizing homogeneous medium and its relation to the Mueller matrix logarithm, *Journal of the Optical Society of America A*, 32 (2):343–348, 2015.

[4] Ramézani, H., Jeong, J., Non-linear elastic micro-dilatation theory: Matrix exponential function paradigm, *International Journal of Solids and Structures*, 67:1–26, 2015.

[5] Jafari, A., Fiétier, N., Deville, M. O., A new extended matrix logarithm formulation for the simulation of viscoelastic fluids by spectral elements, *Computers & fluids*, 39 (9):1425–1438, 2010.

[6] Lastman, G., Sinha, N., Infinite series for logarithm of matrix, applied to identification of linear continuous-time multivariable systems from discrete-time models, *Electronics Letters*, 27 (16):1468–147, 1991.

[7] Williams, P. M., Matrix logarithm parametrizations for neural network covariance models, *Neural Networks*, 12 (2): 299–308, 1999.

[8] Philip, L., Wang, X., Zhu, Y., High dimensional covariance matrix estimation by penalizing the matrix-logarithm transformed likelihood, *Computational Statistics & Data Analysis*, 114:12–25, 2017.

[9] Al-Mohy, A. H., Higham, N. J., Improved inverse scaling and squaring algorithms for the matrix logarithm, *SIAM Journal on Scientific Computing*, 34 (4): C153–C169. 2012.

[10] Cardoso, J. R., Ralha, R., Matrix arithmetic-geometric mean and the computation of the logarithm, *SIAM Journal on Matrix Analysis and Applications* 37 (2):719–743, 2016.

[11] Horenko, I., Schütte, C., Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics, *Multiscale Modeling & Simulation*, 7 (2):731–773, 2008.

[12] Kenney, C. S., Laub, A. J., A Schur–Fréchet algorithm for computing the logarithm and exponential of a matrix, *SIAM Journal on Matrix Analysis and Applications*, 19 (3):640–663, 1998.

[13] Cardoso, J. R., Leite, F. S., Theoretical and numerical considerations about Padé approximants for the matrix logarithm, *Linear Algebra and its Applications*, 330 (1-3):31–42, 2001.

[14] Dieci, L., Papini, A., Conditioning and Padé approximation of the logarithm of a matrix, *SIAM Journal on Matrix Analysis and Applications*, 21 (3):913–930, 2000.

[15] Cheng, S. H., Higham, N. J., Kenney, C. S., Laub, A. J., Approximating the logarithm of a matrix to specified accuracy, *SIAM Journal on Matrix Analysis and Applications*, 22 (4):1112–1125, 2001.

[16] Fasi, M., Higham, N. J., Multiprecision algorithms for computing the matrix logarithm, *SIAM Journal on Matrix Analysis and Applications*, 39 (1): 472–491, 2018.

[17] Fasi, M., Iannazzo, B., The dual inverse scaling and squaring algorithm for the matrix logarithm, *IMA Journal of Numerical Analysis*, 42 (3): 2829–2851, 2022.

[18] Ibáñez, J., Sastre, J., Ruiz, P., Alonso, J. M., Defez, E., An improved Taylor algorithm for computing the matrix logarithm, *Mathematics* 9 (17): 2018, 2021.

[19] Dieci, L., Morini, B., Papini, A., Computational techniques for real logarithms of matrices, *SIAM Journal on Matrix Analysis and Applications*, 17 (3): 570–593, 1996.

[20] Tatsuoka, F., Sogabe, T., Miyatake, Y., Zhang, S.-L., Algorithms for the computation of the matrix logarithm based on the double exponential formula, *Journal of Computational and Applied Mathematics*, 373:112396, 2020.

[21] Kenney, C., Laub, A. J., Condition estimates for matrix functions, *SIAM Journal on Matrix Analysis and Applications*, 10 (2):191–209, 1989.

[22] Higham, N. J., The Matrix Computation Toolbox (2002)
URL:https://www.maths.manchester.ac.uk/∼higham/mctoolbox/

[23] T. G. Wright, Eigtool, version 2.1 (2009).
web.comlab.ox.ac.uk/pseudospectra/eigtool.

A distribution rule for allocation problems with priority agents using least-squares method

J.C. Macías Ponce$^\flat$, A.E. Giles Flores$^\flat$ S.E. Delgadillo Alemán$^{\flat,1}$,
R.A. Kú Carrillo$^\flat$ and L.J.R. Esparza$^\natural$

($\flat$) Departamento de Matemáticas y Física,
Universidad Autónoma de Aguascalientes,
Av. Universidad 940, C.P. 20100, Aguascalientes, Mexico.
($\natural$) Departamento de Matemáticas y Física,
Cátedra CONACyT-Universidad Autónoma de Aguascalientes,
Av. Universidad 940, C.P. 20100, Aguascalientes, Mexico.

1 Introduction

In this work, we use the least-squares method for inconsistent systems with priority in some equations to find a new distribution law for shortage and surplus problems. These problems consist of distributing goods or resources among a set of agents. Notice that we can have three cases depending on the available resource, i.e., when the state is less than, greater than, or equal to the agents' demands, which correspond to the shortage, surplus, and the trivial one, respectively. In the literature, the most common shortage problem is the so-called Bankruptcy problem, characterized by non-negative solutions where no agent can obtain more than was demanded ([8], [9]). The allocation problem that we study is an interesting case since it considers different priorities to satisfy the demands of the agents, which are characteristics that agents may have or possess and that will influence the way the distribution is carried out. Such allocation may allow an agent to receive more than his demand or give from his resources, implying a redistribution of resources.

Let us define an allocation problem as the pair $(\boldsymbol{d}, E) \in \mathbb{R}^n \times \mathbb{R}$, where $\boldsymbol{d} = (d_1, \ldots, d_n)$ is called the demand vector, $d_i \geq 0$ denotes the demand of the agent i, for $i = 1, \ldots, n$, while that $E \geq 0$ represents the available amount of a perfectly divisible resource. Our approach will be to set a system of $n + 1$ linear equations and n unknowns as follows

$$\begin{aligned} x_i &= d_i, \quad i = 1, \ldots, n, && (1) \\ \sum_{i=1}^{n} x_i &= E, && (2) \end{aligned}$$

where equations (1) are called "demands" and equation (2) is called the "efficiency equation". In terms of this notation, the surplus case occurs when $\sum_{j=1}^{n} d_j < E$, while the shortage case occurs when $\sum_{j=1}^{n} d_j > E$. For both cases, the system given by equations (1) and (2) is inconsistent. For the trivial case, we have that $\sum_{j=1}^{n} d_j = E$. To solve (1) and (2) we use the least-squares method (LSM) for inconsistent systems which appears in different areas such as statistics [2], or

[1] elizabeth.delgadillo@edu.uaa.mx

linear algebra [6]. One problem with the LSM solution is that the efficiency equation could not be satisfied. Since it is desirable to divide the resources ultimately, we will apply the LSM to a reformulated linear system of equations that prioritizes the efficiency equation.

2 Methodology and Results

In this section we provide the definitions and main results of our method. Let us start by defining the M-inner product:

Definition 15.1. *Let $\boldsymbol{M}$ be a $m \times m$ real positive definite matrix, we define the M-inner product as $\langle \boldsymbol{v}, \boldsymbol{u} \rangle_M = \boldsymbol{v}^\top \boldsymbol{M} \boldsymbol{u}$, for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$.*

Next, we write the LSM in terms of the M-inner product.

Proposition 1. *Let $\boldsymbol{A}$ be a matrix of size $m \times n$ with real entries, $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{M}$ a positive definite matrix of size $m \times m$. A vector $\boldsymbol{x} \in \mathbb{R}^n$ is a least squares solution of the equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ with respect to the M-inner product if and only if it satisfies the following equation*

$$\boldsymbol{A}^\top \boldsymbol{M} \boldsymbol{A} \boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{M} \boldsymbol{b}. \tag{3}$$

Note that if $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is consistent, the least-squares solution is the usual one. We will call the M-Least Squares Method (M-LSM) the solution where the distance function is given by the norm induced by the M-inner product. The case where the M-LSM solution can be explicitly computed is the following.

Theorem 9. *In the setting of Proposition 1, the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, has a unique least-squares solution with respect to the M-inner product if and only if the columns of $\boldsymbol{A}$ are linearly independent. In this case, the M-LSM solution is given by*

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}^\top \boldsymbol{M} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{M} \boldsymbol{b}. \tag{4}$$

2.1 M-LSM and allocation problems

To solve the allocation problem we need to prioritize the efficiency equation. First, we apply the LSM to the equation system (1) and (2), but repeating the last equation $k-$times to give it a certain priority, i.e.,

$$x_i \quad = \quad d_i, \quad i = 1, \dots, n,$$

$$\left.\begin{array}{rcl} \sum_{i=1}^n x_i & = & E, \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_i & = & E. \end{array}\right\} \text{k} - \text{times} \tag{5}$$

In matrix notation, this system can be written as $\boldsymbol{A}_k \boldsymbol{x} = \boldsymbol{b}_k$, where

$$\boldsymbol{A}_k = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{e} \\ \vdots \\ \boldsymbol{e} \end{pmatrix}_{(n+k)\times n} \quad \text{and} \quad \boldsymbol{b}_k = (d_1, d_2, \dots, d_n, \underbrace{E, \dots, E}_{\text{k}-\text{times}})^\top. \tag{6}$$

The following result tell us what occurs when we take the limit as k to infinity.

Theorem 10. *Let $\boldsymbol{A}_k$ be a complete rank matrix as defined in* (6)*, $\boldsymbol{A}_k\boldsymbol{x}_k = \boldsymbol{b}_k$ the system of equations given by* (5)*, and $\hat{\boldsymbol{x}}_k$ the LSM solution for this system. Then, $\hat{\boldsymbol{x}}_k$ converges to $\boldsymbol{x}^*$ as $k \to \infty$, whose i-th element is given by*

$$[\boldsymbol{x}^*]_i = d_i + \frac{E - \sum_{j=1}^{n} d_j}{n}; \quad i = 1, \dots, n. \tag{7}$$

We will call this equation LSM rule.

The achieved solution can be interpreted as each agent receiving what demands plus an n-th of the deficit or surplus. Under certain conditions, this rule corresponds to the Constrained Equal Loss rule, [4, 5], which gives priority to the agents with higher demand. Now, we discuss the relationship between the repetition of the efficiency equation (2) and the M-inner product. Let us define the matrix $\boldsymbol{M}_k$ of dimension $(n+1) \times (n+1)$, as in Proposition 1,

$$\boldsymbol{M}_k = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0}^\top \\ \boldsymbol{0} & k \end{pmatrix}_{(n+1)\times(n+1)} \tag{8}$$

where k is the number of repetitions of the efficiency equation, $\boldsymbol{0}$ is a row zero vector of dimension n, and $\boldsymbol{I}$ is identity matrix of size $n \times n$. The M-LSM solution given by Theorem 9, equation (4) is given by

$$\hat{\boldsymbol{x}}_{M_k} = (\boldsymbol{A}^\top \boldsymbol{M_k} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{M_k} \boldsymbol{b}$$

which is obtained using the M-inner product, Definition 15.1. It is easy to see that

$$(\boldsymbol{A}^\top \boldsymbol{M}_k \boldsymbol{A})^{-1} = (\boldsymbol{A}_k^\top \boldsymbol{A}_k)^{-1} \quad \text{and} \quad \boldsymbol{A}^\top \boldsymbol{M}_k \boldsymbol{b} = \boldsymbol{A}_k^\top \boldsymbol{b}_k,$$

therefore $\hat{\boldsymbol{x}}_{M_k} = (\boldsymbol{A}^\top \boldsymbol{A_k})^{-1} \boldsymbol{A_k}^\top \boldsymbol{b_k} = \hat{\boldsymbol{x}}_k$. This changes a linear system of high dimension (large k) to modify the $[(n+1), (n+1)]$-th element of the matrix $\boldsymbol{M}_k$ given in (8). In particular, the M-LSM solution $\hat{\boldsymbol{x}}_{M_k}$ of Theorem 9 coincides with the solution $\hat{\boldsymbol{x}}_k$ of Theorem 10, and so we have that

$$\lim_{k\to\infty} \hat{\boldsymbol{x}}_{M_k} = \lim_{k\to\infty} \hat{\boldsymbol{x}}_k = \boldsymbol{x}^*.$$

2.2 Allocation problems with priority agents

Now, we can generalize the distribution rule by considering different priorities to satisfy the agents' demands, by applying Theorem 9 to the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ where $\boldsymbol{M}$ is given by

$$\boldsymbol{M}_k = \begin{pmatrix} \boldsymbol{p}\boldsymbol{I} & \boldsymbol{0}^\top \\ \boldsymbol{0} & k \end{pmatrix} \tag{9}$$

where $\boldsymbol{p} = (p_1, \dots, p_n)$, and $p_i \geq 1$ is the priority of the agent i, $i \in \{1, \dots, n\}$. Note that the priorities, p_i, are a generalization of the repetitions of equation i, but they are no longer required to be a natural number. Moreover, if $p_i > p_j$ we say that the agent i has more priority than the agent j. The following result provides each agent's allocated amount depending on its priority.

Corollary 1. *Let $\boldsymbol{A}\boldsymbol{x}_{k,p} = \boldsymbol{b}$ be the system of equations resulting of an allocation problem given in equations* (1) *and* (2)*, and let $\boldsymbol{M}_k$ as in* (9)*. Then, the i-th element of the M-LSM solution of this system, $\hat{\boldsymbol{x}}_{k,p}$, is given by*

$$[\hat{\boldsymbol{x}}_{k,p}]_i = d_i + \frac{E - \sum\limits_{j=1}^{n} d_j}{p_i \left(\frac{1}{k} + \sum\limits_{j=1}^{n} \frac{1}{p_j} \right)}; \quad i = 1, \dots, n. \tag{10}$$

If we request that the efficiency equation be satisfied, then we take the limit of the equation (10), as k tends to infinity. Therefore, $\hat{\boldsymbol{x}}_{k,p}$ converges to $\boldsymbol{x}_p^*$ as $k \to \infty$, whose i-th element is given by

$$[\boldsymbol{x}_p^*]_i = d_i + \frac{E - \sum_{j=1}^n d_j}{p_i \sum_{j=1}^n \frac{1}{p_j}}, \quad i = 1, \ldots, n. \tag{11}$$

Let us call this the Priority M-LSM rule. Moreover, we present a procedure to set priorities so that we recover different distribution rules reported in the literature [7].

3 Application of Priority M-LSM Rule

We present an application of the Priority M-LSM rule, equation (11), in order to propose a new distribution of the real police force for the states in Mexico[2], by considering the criminal incidence as a criterion to prioritize. Then, we select p_i as the total number of crimes committed by the population aged 18 and over, per 100,000 inhabitants, for each state i, according to the data from Public Safety and Justice 2018 from INEGI[3]. Also, we compare our distribution with the computed with other rules, such as the LSM rule, taking $p_i = 1$, for all $i = 1, \ldots, 32$ and the proportional rule, taking $p_i = \frac{\sum_j d_j}{d_i}$. In Figure 1, one can see the results obtained with these three allocation rules.

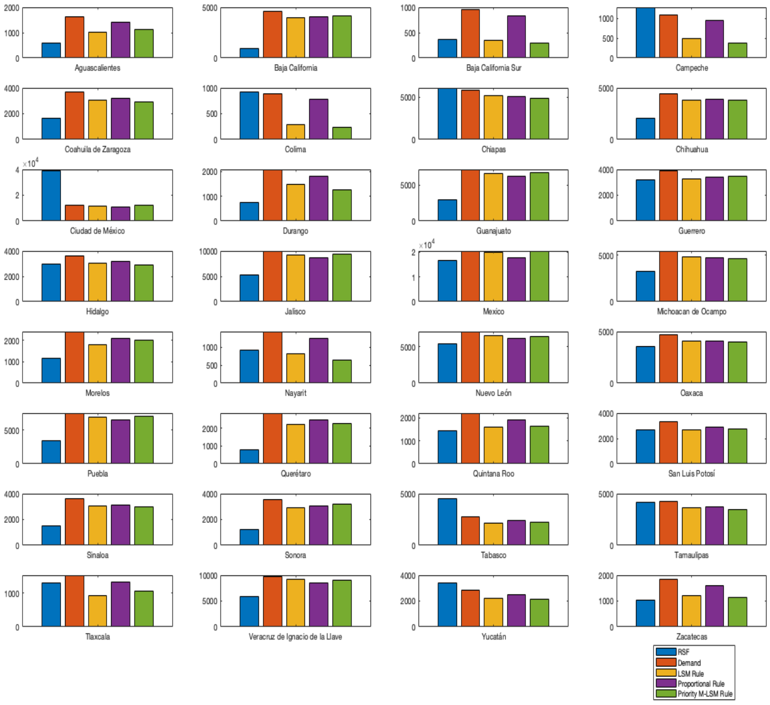


Figure 1: Graphs of the RSF, demand, and allocation solutions of police officers using three types of priorities for each state of Mexico.

Source: Own elaboration.

[2] https://www.eleconomista.com.mx/politica/Identifican-deficit-de-mas-de-100000-policias-en-el-pais-20210504-0014.html

[3] https://www.inegi.org.mx/temas/incidencia/

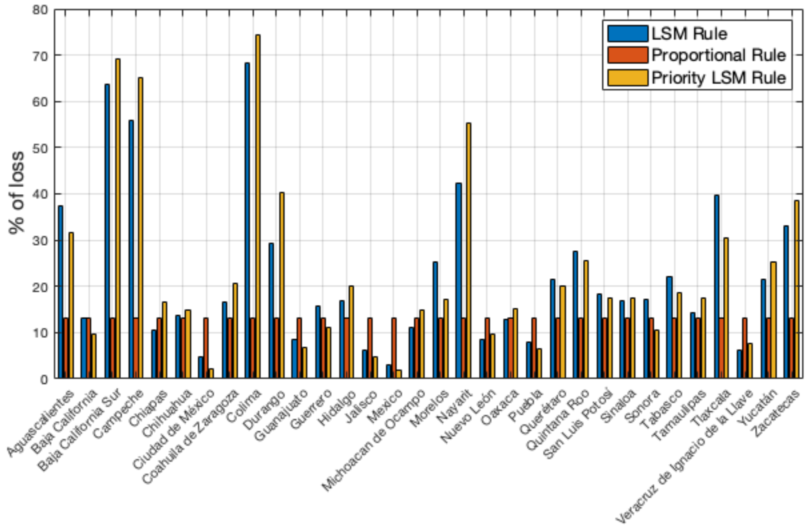


Figure 2: Percentage of a loss considering the LSM rule, proportional rule, and priority LSM rule for each state of Mexico.

Source: Own elaboration.

We can visualise more easily the difference among the rules, depending on the difference between the demand and the allocation, which is called loss. In Figure 1, we present the percentages of loss according with different rules, for each state. As expected, the Proportional rule always gives the same percentage of loss to all the agents (states), approximately 13%. In contrast, the percentage of loss for the LSM and the Priority M-LSM can vary abruptly from one state to the other. In this figure, we can also observe that the difference between the percentage of loss is significant for these two rules for most of the states.

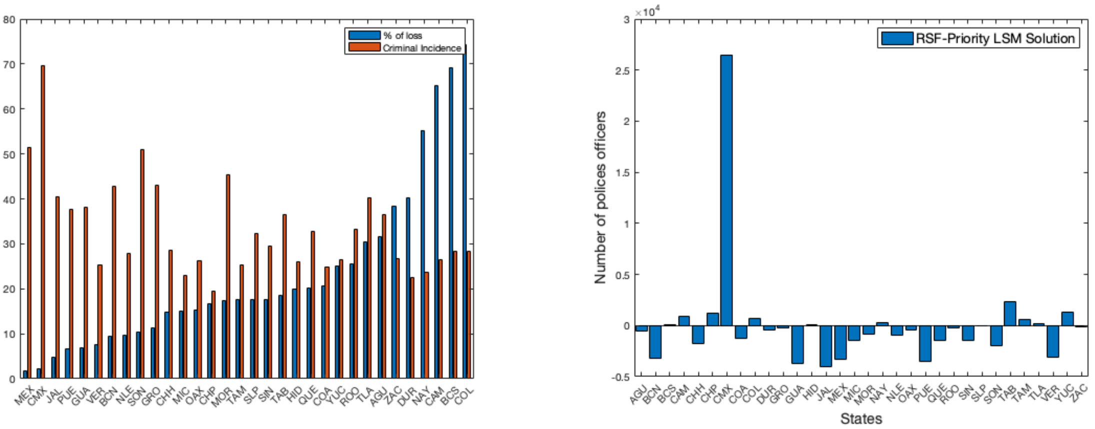


Figure 3: a) Percentage of loss of the demanded number of police officers and criminal incidence for each state of Mexico, b) Surplus (positive) and deficit (negative) of police officers of RSF according to Priority M-LSM solution in each state of Mexico.

Source: Own elaboration.

In Figure 3 a), we present the percentage of loss of the demanded number of police officers in ascending order and the criminal incidence for each state of Mexico, using only the M-LSM rule. Observe that the loss percentage is low for a high criminal incidence because the priority was established using this data. However, this does not occur for some states, since the police officer allocation also depends on the demand. In Figure 3 b), we show the surplus and deficit of police officers of RSF according to the Priority M-LSM solution in each state. Notice that Mexico City has the highest excess of police officers according to our rule. Moreover, the surpluses, mainly from Mexico City, are allocated among the states with deficit.

Acknowledgments

This work has been partially supported by grants PIM20-3, PIM21-1, PIM21-3, PIM21-5, and PIM21-7 from Universidad Autónoma de Aguascalientes. Esparza expresses her gratitude to the CONACYT for financing this work through a Cátedra CONACYT.

References

[1] Casas, F. M. G., Ramos, M. Á. H., & Sánchez, F. J. S. (2006). Teoría de Juegos Aplicada a Problemas de Bancarrota. Contribuciones a la Economía, (2006-02).

[2] Casella, G., & Berger, R. L. (2021). Statistical inference. Cengage Learning.

[3] Secretaría de Seguridad y Protección Ciudadana. (SA). (2021). Modelo Nacional de Policía y Justicia Cívica. Diario Oficial. https://www.gob.mx/cms/uploads/attachment/file/542605/DOC_1._MODELO_NACIONAL_DE_POLIC_A_Y_JC.pdf.
Extended version in: www.dof.gob.mx/2021/SSPC/SEGURIDADyPC_260121.pdf

[4] Herrero, C. (2003). Equal awards vs. equal losses: duality in bankruptcy. In Advances in economic design (pp. 413-426). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-662-05611-0_22

[5] Lorenzo, L. (2010). The constrained equal loss rule in problems with constraints and claims. Optimization, 59(5), 643-660. doi.org/10.1080/02331930802180301

[6] Margalit, D., & Rabinoff, J. (2017). Interactive Linear Algebra. Georgia Institute of Technology.

[7] Moulin, H. (2000). Priority rules and other asymmetric rationing methods. Econometrica, 68(3), 643-684. doi.org/10.1111/1468-0262.00126

[8] Olvera-Lopez, W., Sanchez-Sanchez, F., & Tellez-Tellez, I. (2014). Bankruptcy problem allocations and the core of convex games. Economics Research International, 2014. doi.org/10.1155/2014/517951

[9] O'Neill, B. (1982). A problem of rights arbitration from the Talmud. Mathematical social sciences, 2(4), 345-371. doi.org/10.1016/0165-4896(82)90029-4

A reaction-diffusion equation to model the population of Candida Auris in an Intensive Care Unit

Cristina Pérez-Diukina[♭,1], Juan Carlos Cortés López[♭] and Rafael-Jacinto Villanueva Micó[♭]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

In order to model a population of microorganisms in a given environment through time and space, ordinary differential equations (ODEs) and partial differential equations (PDEs) are, respectively, well studied and very valuable tools. In practice, exact solutions are few and so numerical solutions are often used to describe the dynamic behaviour of the population through time. In order to assert that the numerical solutions are modelling real world phenomena, it is important to calibrate these models with biological and physical data.

In this work, we have applied the Fisher Kolmogorov- Petrovsky–Piskunov (FKPP) equation to model the change in density trough time of Candida Auris (CA) inside an Intensive Care Unit (ICU). The multi-drug resistant yeast CA poses a global threat to the healthcare environment. This model allows us to evaluate the efficacy of well timed cleaning measures on CA population control in the ICU.

2 Methods

2.1 Numerical scheme

The 2-dimensional Fisher Kolmogorov–Petrovsky–Piskunov (Fisher–KPP) model is a reaction-diffusion system that is used to model population growth in a two dimensional coordinate space through time [8], given by the following equation:

$$\frac{\partial u}{\partial t} = D\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + ru(1-u), \tag{1}$$

where x and y are the coordinates of a point on a plane, u is the density of the population in $[a,b] \times [c,d],\ a,b,c,d \in \mathbb{R}$ at a given time t, $D > 0$ is the diffusion coefficient and $0 \leq r \leq 1$ is the growth rate.

[1] cperdiu@upv.es

In this model the variation of the population density u through time is guided by both the diffusion term $D\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)$ which models microorganism's expansion in a plane, and the population growth term $ru(1-u)$, which indicates the amount of CA at a given time, with restricted growth, as it is assumed that the environment has a limited amount of resources. However, the closed solution to the FKPP equation is not known. We thus have to solve it numerically.

We first solve the logistic growth equation $\frac{\partial u}{\partial t}=ru(1-u)$ with closed-form solution:

$$u(t)=\frac{u_0\exp(rt)}{u_0(\exp(rt)-1)+1}, \tag{2}$$

where u_0 is the initial normalized quantity of CA in the ICU. In order to estimate the growth rate, r, and the initial value, u_0, we fit the solution to the logistic growth model (2) to CA *in vitro* growth data [2] using least squares estimation method for non-linear functions in R programming language.

We then solve the FKPP model numerically using Matlab. We use explicit finite differences. We buid a 101×101 mesh-grid to model the plane representing the ICU where $x\in\{0,0.1,0.2,\dots,10\}$ and $y\in\{0,0.1,0.2,\dots,10\}$, which sets the spatial step to be $h=0.1$. We set Neumann, no flux boundary conditions, where the four corners of the ICU plane have no flux and are maintained to be 0.

We let the time go from 0 to 48 hours with a time step of

$$k=\min\{0.5,0.99\frac{h^2}{4D}\}$$

so that $k<\frac{h^2}{4D}$ in order to guarantee stability in the numerical scheme [5,6].

Our numerical scheme, in matrix form is given by:

$$\begin{aligned} u^{(j+1)} &= \frac{kD}{h^2}Au^{(j)}+u^{(j)}+kru^{(j)}\circ(1-u^{(j)}) \\ u^{(j)\prime} &= \Big[u_{1,1,j}=0,u_{2,1,j},u_{3,1,j},\dots,u_{N,1,j}=0,\dots,u_{1,N,j}=0,\dots,u_{N,N,j}=0\Big] \end{aligned} \tag{3}$$

for $j=1,\dots,T=48$, where A is a $N^2\times N^2$, $N=100$ matrix with all zeroes except at the main diagonal, D_0, off 1 diagonals, D_{-1},D_1 and off 10 diagonals,D_{-N},D_N. Here $\circ$ denotes the Hadamard product (also known as the element-wise product).

$$D_0=(\overbrace{0\ \dots\ 0}^{N+1}\quad\overbrace{-4\ \dots\ -4}^{N^2-2N+2}\quad\overbrace{0\ \dots\ 0}^{N+1})$$

$$D_1=D_{-1}=(\overbrace{0\ \dots\ 0}^{N}\quad\overbrace{1\ \dots\ 1}^{N^2-2N-1}\quad\overbrace{0\ \dots\ 0}^{N})$$

$$D_{-N}=(0\quad\overbrace{1\ \dots\ 1}^{N-2}\quad 0\ \ 0\quad\overbrace{1\ \dots\ 1}^{N-2}\quad\dots)$$

$$D_N=(\overbrace{0\ \dots\ 0}^{N+1}\quad\overbrace{1\ \dots\ 1}^{N-2}\quad 0\ \ 0\quad\overbrace{1\ \dots\ 1}^{N-2}\quad\dots)$$

This numerical solution has error $O(h+k^2)$.

To numerically solve the Fisher-KPP model, we define our initial condition, so that it both agrees with previous results and literature on microorganism populations, the dynamics of which in biofilms have been shown to share structural aspects with urbanizations [7]:

$$u(x, y, 0) = \exp\{3\big(- (x-5)^2 - (y-5)^2\big)\}. \tag{4}$$

We then applied Particle Swarm Optimization (PSO), a bio-inspired optimization algorithm, fist introduced by Kennedy and Eberhar in 1995 [3], to calibrate both the diffusion coefficient D and the growth coefficient r simultaneously. We define as the objective function for the PSO to minimize, the Symmetric Mean Absolute Percentage Error (SMAPE), which has been shown to be a better measure of relative error [9]. SMAPE is defined as:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{|z_i - f(t_i)|}{\frac{|z_i| - |f(t_i)|}{2}} \tag{5}$$

where $\{t_1, t_2, \ldots, t_n\}$ is the set of times where CA's $A600nm$ (absorbance at $600nm$ wavelength) was measured [2].

2.2 Introduction of a cleaning factor

The ICU is cleaned regularly by trained staff in order to minimize the spread of any harmful microorganism. To model this and in order to estimate the efficacy of cleaning measures on CA population control and determine potential recommendations, we introduce a cleaning factor into our model. We assume that the ICU is cleaned in a uniform manner. So at times separated by equal intervals, the amount of CA is reduced by some percent at every point (x, y) on the plane. Then when this is introduced to the model, the amount of CA in the ICU at every point (x, y) on the plane and time t is

$$u(x, y, t) = \begin{cases} p\hat{u}(x, y, t), & \text{if cleaning happend at time } t, \\ \hat{u}(x, y, t), & \text{otherwise.} \end{cases} \tag{6}$$

So we periodically reduce the population of CA present in the ICU by a percentage in a homogeneous way.

We then compare the values of

$$M = \frac{max_{i=1,\ldots,n}[\text{total amount of CA at time point i}]}{1,0197 \times 10^2} \tag{7}$$

for different combination of time intervals between cleaning (TI) and cleaning efficacy (CE). This represents the maximum quantity of CA present in the ICU relative to the worse case scenario where the whole ICU is infected $(1,0197 \times 10^2)$.

We choose to vary TI from 2 hours to 10 hours with a one hour increase. Cleaning more often than every $2h$ seems unrealistic and expensive, whereas cleaning less than every $10h$ appears to be too low for a healthcare environment with gravely ill patients who are very susceptible to any sort of infection.

3 Results

The PSO results for SMAPE return the parameter estimates $\hat{D} = 0.4141$, $\hat{r} = 0.3539$. Figure 2 shows the dispersion of CA through the ICU and Figure 1 shows that our total amount of CA at any given time point follows the real data rather well. Using these parameters we then introduced

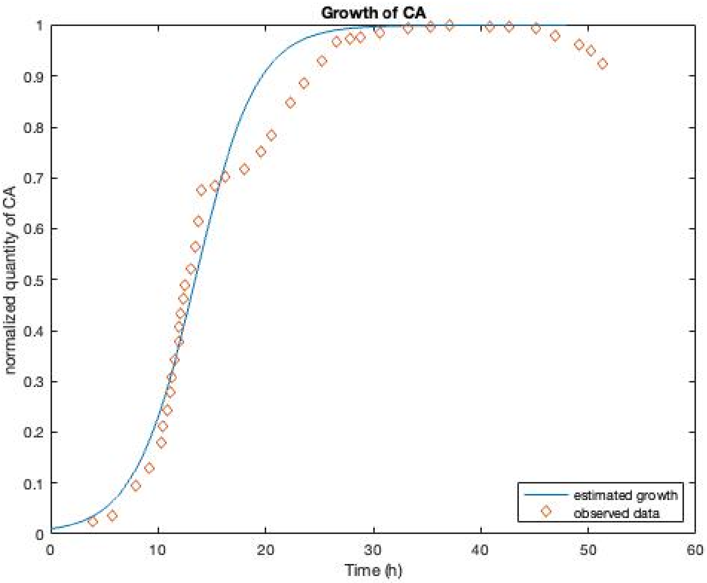


Figure 1: Observed *in vitro* growth of CA (points in orange) and total amount of CA at any given time point (blue line) estimated using PSO and SMAPE error.

the cleaning factor.

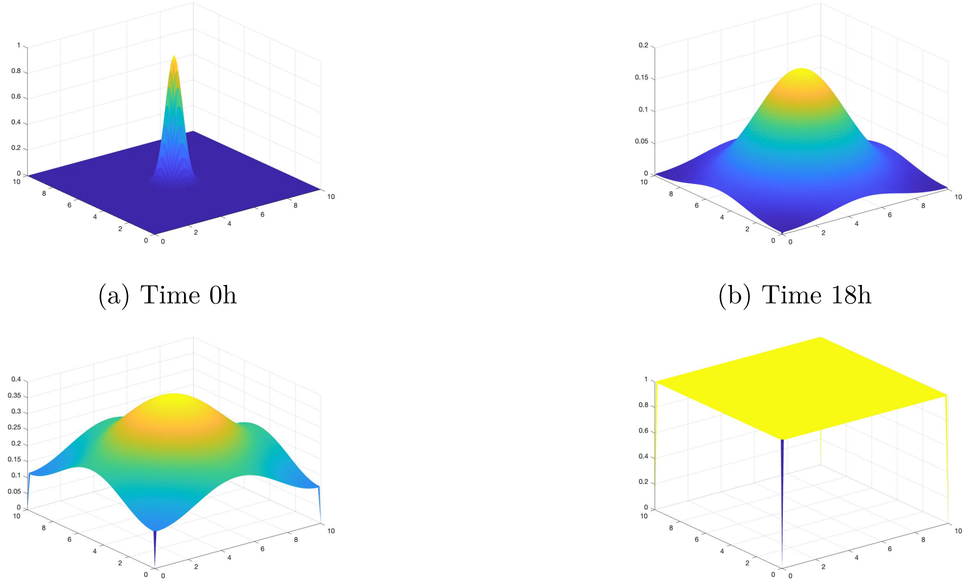
Table 1: Normalized maximum total amount of CA present between 0 and 48 hours (M (7)) for different combinations of TI and CE

		TI (hours)								
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CE (%)	**96.6**	0.0103	0.0103	0.0103	0.0103	0.0103	0.0103	0.0103	0.0103	0.0185
	90	0.0103	0.0103	0.0103	0.0103	0.0103	0.0265	0.1591	0.3573	0.4875
	80	0.0103	0.0103	0.0103	0.0362	0.2644	0.4727	0.7144	0.8023	0.8476
	70	0.0103	0.0103	0.0865	0.3949	0.6540	0.7570	0.8493	0.8951	0.9226
	60	0.0103	0.0702	0.4587	0.6695	0.7918	0.8557	0.9049	0.9341	0.9525
	50	0.0122	0.4161	0.6707	0.7901	0.8631	0.9062	0.9369	0.9564	0.9689

Table 1 compares values of M (7) and shows us that in order to control CA population and keep its amount always under the initial amount present in the outbreak, cleaning of the ICU should be performed at most every $3h$. If cleaning is performed every $3h$ the cleaning efficacy of the agent should be at least 70%. If it is affordable to perform cleaning every $2h$, then the cleaning agent should have an efficacy of at least 60%. For cleaning efficacies of 50% and under, we do not achieve the desired population control.

4 Conclusions

We have built a FKPP reaction-diffusion model, which we solved numerically and then calibrated to real-life data of CA growth. In our simulations, we showed that if cleaning is done often enough, and at a high enough frequency, the maximum amount of CA present in the ICU does not surpass



(a) Time 0h

(b) Time 18h

(c) Time 25h

(d) Time 48h

Figure 2: 3D plots of the FKPP numeric solutions when $\hat{r} = 0.3539, \; \hat{D} = 0.4141$

the initial amount of the outbreak. If cleaning is performed every 3 hours the cleaning efficacy of the agent should be at least 70%. If cleaning is performed every 2 hours, then the cleaning agent should have an efficacy of at least 60%.

There are, however, great limitations in our model and the introduction of the cleaning factor that need to be addressed.

First, our model is calibrated with in-vitro growth of CA which is not equivalent to microbial growth in a health care environment [2]. The ICU is kept at a temperature of 23^oC, whereas the strains in the observed data were incubated at 37^oC, which is a lot closer to the optimal temperature for CA growth. Therefore, we can expect CA growth to be slower in an ICU, approaching a full contamination level at around 48 hours, as mentioned by ICU personnel, rather than 25 hours.

Secondly, our model does not account for the ecological pressures present in the microbial environment such as local extinction and colonization processes as different species compete for resources [4]. It is important to think about how cleaning can affect the environmental competition. Cleaning could perhaps reduce in a more significant way a microorganism that is highly competitive with CA, which would then allow CA access to more resources and accelerate its growth. Some cleaning agents such as H_2O_2 vapor have been shown to be more effective at killing non CA species [1].

Homogeneous cleaning is also an unreasonable assumption. Some of the ICU material cannot be cleaned as in depth as a simple plastic surface. Keyboards, screens, tubes, etc. cannot be always cleaned with all cleaning agents and in a in depth manner.

This model, however, indicates that the efficacy of the cleaning agent and how often the ICU room is cleaned greatly affects how well the yeast can be controlled.

References

[1] Alireza Abdolrasouli et al. "In vitro efficacy of disinfectants utilised for skin decolonisation and environmental decontamination during a hospital outbreak withiCandida auris/i". In: Mycoses 60.11 (Sept. 2017), pp. 758–763. doi: 10.1111/myc.12699.

[2] Leiwen Fu et al. "Study on growth characteristics of Candida auris under different conditions in vitro and its in vivo toxicity". In: Zhejiang da Xue Xue Bao, Journal of Zhejiang University. Medical Sciences 40.7 (2020), pp. 1049–1055

[3] J. Kennedy and R. Eberhart. "Particle swarm optimization". In: Proceedings of ICNN'95-International Conference on Neural Networks. IEEE. doi:

[4] Juan E. Keymer et al. "Bacterial metapopulations in nanofabricated landscapes". In: Proceedings of the National Academy of Sciences 103.46 (Nov. 2006), pp. 17290–17295. doi: 10.1073/pnas.0607971103.

[5] Patrick Lascaux. Lectures on Numerical Methods for Time Dependent Equations: Applications to Fluid Flow Problems. Vol. 52. Tata Institute of Fundamental Research, 1976.

[6] Jordan Lee and Mr Sungwoo Jeong. Stability of Finite Difference Schemes on the Diffusion Equation with Discontinuous Coefficients. (2017.

[7] Amauri J. Paula, Geelsu Hwang and Hyun Koo. "Dynamics of bacterial population growth in biofilms resemble spatial and structural aspects of urbanization". In: Nature Communications 11.1 (Mar. 2020). doi: 10.1038/s41467-020-15165-4.

[8] Vladimir M Tikhomirov. Selected Works of AN Kolmogorov: Volume I: Mathematics and Mechanics. Vol. 25. Springer Science & Business Media, 1991, pp. 242–245.

[9] Tofallis, C "A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation", In: Journal of the Operational Research Society, 66(8),1352-1362 (2015).

17

Relative research contribution towards railways superstructure quality determination from the vehicles inertial response

E. Gómez[♭,1], J. H. Alcañiz[♭], G. Alandí[♮] and F. E. Arriaga[♢]

(♭) Universidad Católica San Antonio de Murcia,
Av. de los Jerónimos 135, Guadalupe de Maciascoque, Murcia, Spain.
(♮) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(♢)Idvia 2020 Horizonte 2020 SL.
Av. d'Aragó 30, València, Spain.

1 Introduction

In recent years, rail transport has become the most important strategic mode of transport for both freight and passengers [1]. In this context, the maintenance of track quality, as well as the rapid intervention of its defects, is one of the challenges on which railway administrations focus their efforts.

Consequently, this publication deals with the mathematical treatment of the inertial signals experienced by the track user vehicle, from the bogie room model [2], to determine the presence of alterations in the track superstructure geometry affecting the track quality, from a data acquisition system installed on commercial trains in operational conditions.

In this way, firstly, the appropriate hardware was designed to record the necessary data for the application of the mathematical model proposed, which is basically made up of triaxial accelerometers placed on the unsprung masses of the bogie, a GPS-type positioning system, a data acquisition system, a communication system and a power supply system. Secondly, algorithms were developed to process the signal obtained in each case, including the selection of the optimum frequency filter. Finally, the system was technologically packaged and functional prototype tests were carried out on track 1 of MetroValencia, in the section between the stops Torrent and Paiporta, where it was possible to verify the importance of signal filtering, a subject on which the thesis is currently focusing.

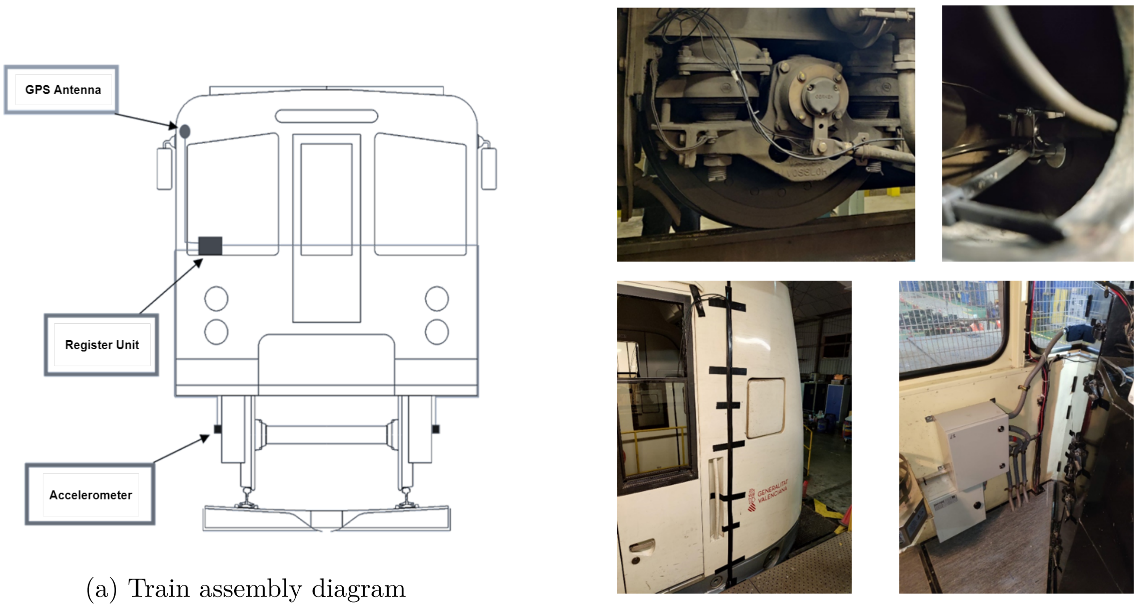
2 Methods

This section presents in detail, the methodology followed for the indirect assessment of rail defects and irregularities, which can be divided in two phases: i) data acquisition; and ii) data processing.

[1] egomez0@alu.ucam.edu

2.1 Data acquisition

In this regard, the time history of accelerations is registered on the vehicle body by means of tri-axial accelerometers; while its position on the track is constantly recorded with a GPS. The acquisition system was installed on a locomotive wagon of MetroValencia (Valencia, Spain) as shown in Figure 1. The tests consisted of two passes over the section between the Torrent and Paiporta stops, which is 4 km long.



(a) Train assembly diagram

(b) Photographs of the data acquisition system set-up

Figure 1: Assembly of data acquisition system

2.2 Data processing

. Then, the measurements are processed and transformed into 3D rail geometry data along the railway line, according to the algorithm described below and presented in Figure 2.

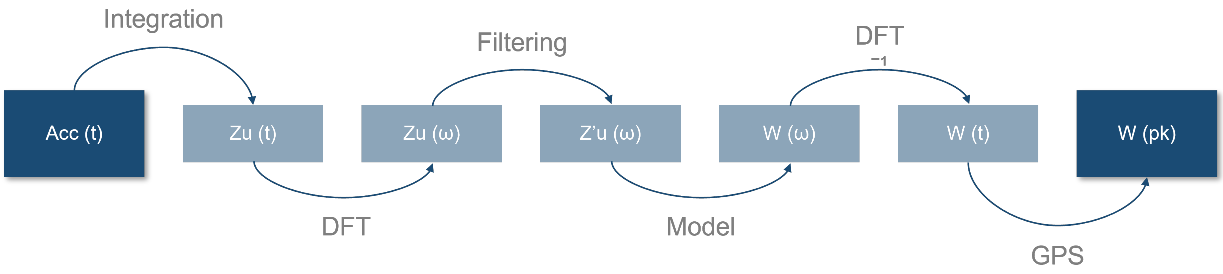


Figure 2: Scheme of the data processing algorithm

First, the recorded time history of accelerations is double-integrated over time through eq. (1) in order to obtain the time history of vehicle displacements.

$$Z_u(t) = W_0 + V_0 t + \int \int_0^t Acc(t)\, dt\, dt \tag{1}$$

Where x(t) and v(t) are the displacement and acceleration time histories, respectively; and x0, v0 are the initial values of vehicle displacements and velocities. Once the vehicle displacements have been calculated in the time domain, they shall be transformed into the frequency domain in order to allow a subsequent filtering of the data. This process is carried out by means of the Fourier transform in a discrete form (DFT, (2)).

$$Z_u(w) = \sum_{r=1}^{n} Z_u(t)\mathring{u}e^{\frac{2\pi i(r-1)(s-1)}{n}} \tag{2}$$

Where coefficients r and s vary from 1 to n; and n is the total number of points in the data series. A high-pass filtering of the signal is then performed with the aim of removing low frequencies (i.e., long wavelength components), since they do not correspond to rail irregularities or alignment defects, but to track geometry regular variations, such as cant and slope changes.

The fourth step of the algorithm deals with the vehicle-track interaction as shown in Figure 3. In this phase, the displacements on the wheel-rail interface are calculated by solving the system of differential equations provided by the two-masses model of the vehicle - eq. (3) -. The model explicitly accounts for the effect of the unsprung mass (i.e., wheelset) and a combination of sprung (i.e., car body) and semi-sprung (i.e., bogie) masses.

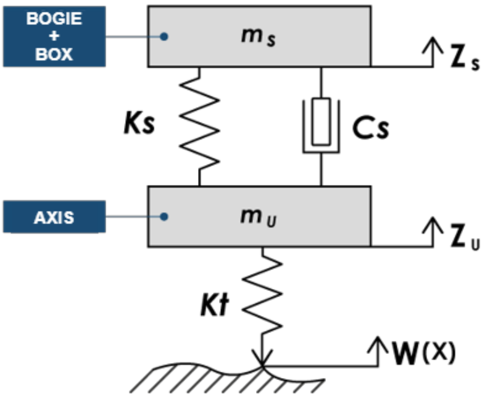


Figure 3: Quarter bogie model

$$\begin{gathered}
m_s\ddot{Z}_s + C_s(\dot{Z}_s - \dot{Z}_u) + K_s(Z_s - Z_u) = 0 \\
m_u\ddot{Z}_1 - C_s\dot{Z}_s + C_s\dot{Z}_u - K_sZ_s + (K_t + K_s)Z_u - K_tW = 0 \\
W = Z_u + \frac{m_s\ddot{Z}_s + m_u\ddot{Z}_u}{K_t}
\end{gathered} \tag{3}$$

Where m_u is the unsprung mass; m_s is the combination of sprung and semi-sprung masses; k1 is the track stiffness; and K_s, C_s are the stiffness and damping coefficients of the primary suspension, respectively. The recorded data have been transformed in the previous step into the frequency domain for filtering; and thus, eq.(3) is transformed into eq.(4) by means of the Fourier transform properties [3] and solved in such domain.

$$\begin{gathered}
-m_sw^2Z_s + iC_sw(Z_s - Z_u) + K_s(Z_s - Z_u) = 0 \\
-m_uw^2Z_u - iC_swZ_s + iC_swZ_u - K_sZ_s + (K_t + K_s)Z_u - K_tW = 0
\end{gathered} \tag{4}$$

Once the rail geometry is known in the frequency domain, the results shall be transformed back to the time domain, which is performed by means of the inverse discrete Fourier transform shown in eq. (5).

$$W(t) = \frac{1}{n}\sum_{r=1}^{n} W(w)\mathring{u}e^{-2\pi i\frac{(r-1)(s-1)}{n}} \tag{5}$$

Finally, the rail geometry dataset in the time domain is transformed into the space domain by correlating its values with the locations provided by the GPS.In this way, the position W for each rail point is obtained, which can be analysed and compared with the standard and track geometry defects can be discovered through its geometrical parameters: levelling, alignment, superelevation, track gauge and warping.

3 Results

The tests have been carried out on 4 km of MetroValencia track between the Torrent and Paiporta stops. In these tests, the geometric parameters, divided into vertical and horizontal, were analysed by applying the methodology. Vertical parameters include levelling, cant and warping, while alignment and track gauge are horizontal parameters.

For the sake of conciseness, only the results for superelevation are shown. Figure 4 shows the superelevation obtained and the GPS trajectory after the analysis, superimposed on the limits set by the UNE-EN 13868-1 standard.

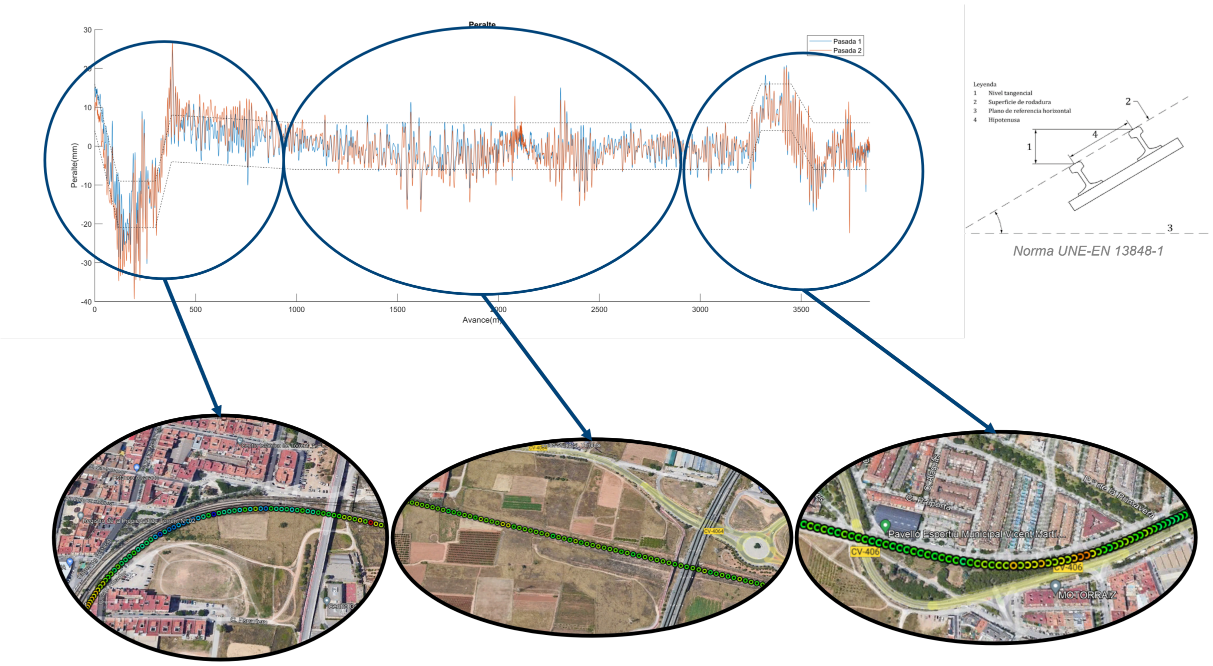


Figure 4: Experimental results for superelevation parameter

Then, Table 1 shows the results obtained for the first pass and Table 2 shows the results obtained for the second pass. Both tables show the number of out-of-tolerance points, the percentage of out-of-tolerance points in relation to the total, and the equivalent in linear metres out of tolerance that this percentage represents.

Results 1				
Parameters	Points out of tolerance	Points analysed	Track % out of tolerance	Meters out of tolerance
Leveling L	7204	534009	1.35	51.8
Leveling R	5348	534009	1	38.45
Alignment L	8115	534009	1.52	58.35
Alignment R	7266	534009	1.36	52.25
Gauge	0	534009	0	0
Superelevation	72848	534009	13.64	523.8
Warping	187	534009	0.04	1.34

Table 1: Results obtained for the first pass.

Results 2				
Parameters	Points out of tolerance	Points analysed	Track % out of tolerance	Meters out of tolerance
Leveling L	7858	544000	1.44	55.33
Leveling R	5261	544000	0.97	37.04
Alignment L	8134	544000	1.5	57.27
Alignment R	7027	544000	1.29	49.48
Gauge	0	544000	0	0
Superelevation	100174	544000	18.41	705.33
Warping	171	544000	0.03	1.2

Table 2: Results obtained for the second pass.

4 Conclusions and Future Work

The experiment carried out in MetroValencia has served to collect the necessary data to apply the methodology proposed in this paper. It has been observed that it is possible to analyse the quality of the railway superstructure through the inertial response of the vehicle and, through the mathematical treatment of the signal, to obtain the geometric parameters of the track.

In the comparison between the results obtained for the two passes (see Table 1 and Table 2), a good correlation of data was observed for all geometric parameters except camber, where there is a 5 % difference in the number of defects detected. In this sense, this result indicates that the signal filtering algorithm for the superelevation will be revised.

As for future work, in order to guarantee maximum precision on the measurements, the following steps will be carried out:

- A comparison of the records and the thresholds established according to the current regulations.
- Development and adjustment of adaptive filters with the speed.
- Adaptation of the analysis to the direction of circulation of the railway vehicles.

References

[1] A Elkhoury, N., Hitihamillage, L., Moridpour, S., Robert, D., Degradation prediction of rail tracks: A review of the existing literature. *The Open Transportation Journal*, 8:88–104, 2018.

[2] Pehlivan, F., Mizrak, C., Esen, I., Modelling and validation of 2-DOF rail vehicle model based on electro–mechanical analogy theory using theoretical and experimental methods *Engineering, Technology & Applied Science Research*, 8:3603-3608, 2018.

[3] Melis, M., Apuntes de Introducción a la Dinámica Vertical de la Vía y a las Señales Digitales en Ferrocarriles. Madrid, INGENIERÍA DE FERROCARRILES, METROS Y TÚNELES, 2008.

Computational Tools in Cosmology

Màrius Josep Fullana i Alfonso$^{\flat,1}$ and Josep Vicent Arnau i Córdoba$^{\flat}$

($\flat$) Institut Universitari de Matemàtica Multidisciplinària,
Universitat Politècnica de València,
Camí de Vera s/n, València 46022, Spain.

1 Introduction

Following the line started in [1], some interesting numerical and computational techniques used in part of our research in Cosmology are now presented. As remarked in [1] (with other algorithms we built), we also think the tools described in this conference presentation may work in different tasks concerning our research fields. Moreover, they may even be extended in Science and Technology in general. The approaches showed here have been developed in N-body algorithms applied when describing the CMB (Cosmic Miicrowave Background) anisotropies (see [2–6]). In our methods, innovations in CMB maps treatments are performed. Such novelties may be extended to other studies.

2 Methods

Along more than a decade, we have presented the advance of our CMB anisotropy computations using N-body codes. The codes with more resolution and precision we used were the N-body Hydra ones (see [3]). All versions were designed by members of the Hydra Consortium. We used 1) Codes without baryons. 1.a) Sequential versions. 1.b) Parallel ones. With both of them we computed the weak lensing (WL) and the Rees-Sciama (RS) contributions to the CMB angular power spectrum.

Using our numerical techniques, we reported a higher contribution –to lensing– than previous approaches. Our CMB anisotropies computations on every step of the run allowed less interpolations and approximations. This could be the explanation of our excess of power in lensing computations. Our higher resolution could also contribute to this excess.

Afterwards, we also performed computations with baryons (see [6]). This version allowed us to compute Sunyaev-Zel'dovich (SZ) contribution to the CMB angular power spectrum too.

An appropriate ray-tracing procedure through N-body simulations was proposed in the following basic references: [7, 8]. In these papers it was explained how to chose a preferred direction (PD) to cross the N-body simulated boxes. Such directions wee chosen to reach the initial position after passing through 16 boxes. For a box size $L = 512h^{-1}Mpc$ the distance between points entering and leaving each box was $\sim 104h^{-1}Mpc$. So, one had independent regions from redshift $z \sim 6$ ($\sim 5900h^{-1}Mpc$). For our computations, starting in $z \sim 6$ was sufficient. There was no need to start computations at higher redshifts. Applications based on our ray-tracing methods through PM simulations can be seen, for instance, in [2, 9].

[1] mfullana@mat.upv.es

Some important computational details of the map construction are now detailed.

For weak lensing (see [3]), small, unlensed maps of CMB temperature contrasts ($\Delta = \delta T/T$) were constructed to be subsequently deformed by lensing. In order to deform the unlensed maps, the lens deviations corresponding to a set of directions, covering an appropriate region of the sky, were calculated. These deviations corresponded to the quantities:

$$\vec{\delta} = -2 \int_{\lambda_e}^{\lambda_0} W(\lambda) \vec{\nabla}_{\perp} \phi \, d\lambda \ , \tag{1}$$

where $\vec{\nabla}_{\perp}\phi = -\vec{n} \wedge \vec{n} \wedge \vec{\nabla}\phi$ is the transverse gradient of the peculiar gravitational potential ϕ, and $W(\lambda) = (\lambda_e - \lambda)/\lambda_e$. The variable λ is:

$$\lambda(a) = H_0^{-1} \int_a^1 \frac{db}{(\Omega_{m0} b + \Omega_\Lambda b^4)^{1/2}} \ . \tag{2}$$

Once the deviations were calculated, they could be easily used to get the lensed maps from the unlensed ones. This was achieved using the relation:

$$\Delta_{_L}(\vec{n}) = \Delta_{_U}(\vec{n} + \vec{\delta}) \ , \tag{3}$$

where $\Delta_{_L}$ and $\Delta_{_U}$ are the temperature contrasts of the lensed and unlensed maps, respectively. The unit vector $\vec{n}$ defines the observation direction (line of sight).

Given the unlensed map $\Delta_{_U}$, and the map $\Delta_{_L}$ obtained from it after deformation by lensing (the lensed map), the chosen power spectrum estimator could be used to get the quantities $C_\ell(U)$ and $C_\ell(L)$, whose differences $C_\ell(LU) = C_\ell(L) - C_\ell(U)$ could be considered as an appropriate measure of the weak lensing effect on the CMB.

For the Rees-Sciama contribution we computed the integral (see [2,4]):

$$\frac{\Delta T}{T_B}(\vec{n}) = 2 \int_{\lambda_e}^{\lambda_0} W(\lambda) \, \frac{\partial \phi}{\partial \lambda} \, d\lambda \ , \tag{4}$$

where ϕ is the peculiar gravitational potential ϕ, $W(\lambda) = (\lambda_e - \lambda)/\lambda_e$ and λ is given in eq(2).

For the Sunyaev-Zel'dovich thermal contribution in the long wave regimes we computed the integral (see [6]):

$$\frac{\Delta T}{T_B}(\vec{n}) = -2 \, \frac{\sigma_T}{m_e c^2} \int_{\lambda_e}^{\lambda_0} n_e \, k \, T_e \, d\lambda \ , \tag{5}$$

where the subscript e refers to electrons.

Notice that we had to define different weak lensing regimes. Basically, this is the way to proceed (see [3]):

- *AWL* (A weak lensing), namely the effect due to scales $k > 2\pi/L_{max}$ (where $L_{max} = 42h^{-1}$ Mpc) at redshifts $z < 6$. This signal is dominated by strongly nonlinear scales (namely structures).

- *BWL*, the lensing signal due to scales $k < 2\pi/L_{max}$ which corresponds to modes that are always in the linear regime down to $z = 0$.

- *CWL*, the lensing signal due to scales $k \geq 2\pi/L_{max}$ but at redshifs $z > 6$.

- *RS*, the same regimes that one has for WL apply for RS.

- *SZ*, this distinction does not apply.

We now describe the main features of the algorithm designed to compute the physical quantities described above and that allowed to study the CMB anisotropies using the Hydra N-body codes (see [3]):

1. Decide upon the **direction** of the normal rays representing the geodesics.
2. Assuming the Born approximation and using the **photon step distance** Δ_{ps}, determine all the evaluation positions and times on the geodesics within the simulation volume from $z = 6$ down to the final redshift.
3. Associate **test particles** with each of these positions and times.
4. At each time-step of the N-body simulation (while it is running) determine which test particles require **force evaluations** (or other physical quantities depending on the CMB anisotropy effect to be computed).
5. At each test particle position evaluate the **force** (or corresponding physical quantity) on the test particle using the long-range **FFT** component and short-range **PP correction** as in the HYDRA algorithm.
6. During the FFT convolution for the test particles **eliminate** contributions from **scales larger than** $42h^{-1}$ **Mpc** by removing the signal from wavenumbers satisfying $k \leq 0.15h$ Mpc^{-1}.
7. If the evaluation time for a point on the geodesic lies between two time-steps calculate a **linear interpolation** of the two forces from the time-steps that straddle **the correct time**.
8. Resolve the **force** into its **transverse component** and hence recover the transverse component of the potential gradient. This applies for WL. For RS and SZ, see [2, 4] and [6], respectively, for the physical quantities to be computed (potential in Eq. (4) and electrons temperatures in Eq. (5), respectively).

3 Results

Simulations were performed in the framework of the concordance model with the following parameters: $h = 0.7$, $\Omega_b = 0.046$, $\Omega_d = 0.233$, $\Omega_\Lambda = 0.721$, optical depth $\tau = 0.084$ and $\sigma_8 = 0.817$. The power spectrum of the scalar (adiabatic) energy density perturbations was obtained with the CMBFAST code. No tensor modes were considered at all.

One of the numerical advances of our work was that the correlation function $\xi(r)$ extracted from one simulation (instead of 30 which was the usual method) was sufficient. This showed that our nonlinear algorithm was very robust. Besides, its form was that expected for the softening length and the box size. The code worked very well in spite of the modifications required by our CMB calculations.

For lens deformations (see [3]), the angular power spectra for one simulation was compared to the same simulation but where deflections were calculated by including an average over the 8 nearest geodesics. This reduced the resolution of the geodesic method, but maintained the same resolution in the gravitational solver. The resulting power spectrum was plotted in Figure 10 of [3] and showed a decaying signal at high ℓ which was similar to that found in earlier works (e.g. [10]). This showed that as we degraded the resolution of our ray-tracing method we indeed had close results to those of previous works with less resolution than ours. Therefore, the local averages, used in methods of other authors, might hide the highly nonlinear structures effects. Notice that these structures had a relative small size.

For RS (see [2, 4]) and SZ (see [6]), the results we obtained were of the order of magnitude or slightly greater than those obtained by other authors.

4 Conclusions

Our AP3M codes adapted to CMB calculations could be run for different values of the parameters defining the simulations; hence, this code allowed us to see how the resulting angular power spectra depended on the parameters defining both the N-body simulation and the ray-tracing procedure.

Simulations in boxes of $512h^{-1}$ Mpc led to good $C_{\ell}(LU)$ spectra for $1000 < \ell < 7000$. For $2000 < \ell < 7000$, all the simulations lied in a region of width $\sim$0.5 μK, indicating that the simulations gave consistent estimates of the signal in this range (see [3]). The signal in the range $4000 < \ell < 7000$ is 2.0 ± 0.4 μK, which is ~ 1.4 μK higher than that found elsewhere [10].

The values we obtained were compatible with studies based on the Millennium simulation (see [11]), where the authors reported a small contribution from nonlinearity at $\ell \simeq 4100$. However, the methods of [11] were designed to build all-sky lensed maps, and therefore did not have the necessary resolution to perform an accurate estimate of the weak lensing by strongly nonlinear structures in the ℓ-interval where we found our main effect.

Now we are working on the analysis and description of the numerical advances we have made in all the research described in the present paper. Such as the improvements we made on the resolution of N-body algorithms, on our FFT subroutines and on numerical parallelisation technics. Also our ameliorations on the numerical treatment of images necessary to extract the power spectrum of CMB. This work will be presented in the near future.

Acknowledgements

We dedicate this work to Professor Diego Pascual Sáez Milán, in memoriam.

We acknowledge the financial support of the Spanish Ministerio de Ciencia, Innovación y Universidades and the Fons Europeu de Desenvolupament 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, N., Fullana i Alfonso, M.J., Arnau, J.V., Sáez, D., On the Rees-Sciama effect: maps and statistics *Monthly Notices of the Royal Astronomical Society*, Volume 370, Issue 4, Pages 1849-1858, August 2006. https://doi.org/10.1111/j.1365-2966.2006.10607.x

[3] Fullana i Alfonso, M.J., Arnau, J.V., Thacker, R.J., Couchman, H. M. P., Sáez, D., Estimating Small Angular Scale Cosmic Microwave Background Anisotropy with High-Resolution N-Body Simulations: Weak Lensing *The Astrophysical Journal*, Volume 712, Number 1, 367-379, March 2010. https://doi.org/10.1088/0004-637X/712/1/367

[4] Fullana i Alfonso, M.J., Arnau, J.V., Thacker, R.J., Couchman, H. M. P., Sáez, D., On the estimation and detection of the Rees-Sciama effect *Monthly Notices of the Royal Astronomical Society*, Volume 464, Issue 4, Pages 3784-3795, February 2017. https://doi.org/10.1093/mnras/stw2615

[5] Fullana i Alfonso, M.J., Josep Vicent Arnau i Córdoba, J.V., Puchades Colmenero, N., *Some advances in Relativistic Positioning Systems. In Modelling for Engineering & Human Be-*

haviour 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*

[6] Fullana i Alfonso, M.J., Arnau i Córdoba, J.V., Thacker, R.J., Couchman, H.M.P., Sáez , D.P., A New Numerical Approach to Estimate the Sunyaev-Zel'dovich Effect. In: García-Parrado, A., Mena, F., Moura, F., Vaz, E. (eds), *Progress in Mathematical Relativity, Gravitation and Cosmology. Springer Proceedings in Mathematics & Statistics, vol 60.* Berlin, Heidelberg, Springer, 2014. Doi: 10.1007/978-3-642-40157-2_38

[7] Cerdà-Durán, P., Quilis, V., Sáez, D., Non Gaussian Signatures in the Lens Deformations of the CMB Sky: A New Ray-Tracing Procedure *Phys. Rev.* 69D, 043002, 2004. *arXiv:astro-ph/0311431v1*

[8] Sáez, D., Puchades, N., Fullana, M.J., Arnau, J.V., Ray-Tracing through N-body simulations and CMB anisotropy estimations *Proceedings of Science: CMB and Physics of the Early Universe*, PoS CMB2006 pp.058, Ischia, 2006.

[9] Antón, L., Cerdà-Durán, P., V. Quilis, V., D. Sáez, D., Cosmic Microwave Background Maps Lensed by Cosmological Structures: Simulations and Statistical Analysis, *The Astrophysical Journal*, 628, 1, 2005. *arXiv:astro-ph/0504448v1*

[10] Das, S., Bode, P., A Large Sky Simulation of the Gravitational Lensing of the Cosmic Microwave Background *The Astrophysical Journal*, Volume 582, Number 1, 1-13, July 2008. Doi: 10.1086/589638

[11] Carbone, C., Springel, V., Baccigalupi, C.,Bartelmann, M., Matarrese, S., Full-sky maps for gravitational lensing of the cosmic microwave background *Monthly Notices of the Royal Astronomical Society*, Vol. 388, 1618-1626, 2008. Doi: 10.1111/j.1365-2966.2008.13544.x

Dynamical analysis of a family of Traub-type iterative methods for solving nonlinear problems

F.I. Chicharro$^\flat$, A. Cordero$^\flat$, N. Garrido$^{\flat,1}$ and J.R. Torregrosa$^\flat$

($\flat$) Instituto de Matemática Multidisciplinar, Universitat Politècnica de València,
Camí de Vera s/n, 46022 València, Spain.

1 Introduction

The problem of calculating the roots of nonlinear functions arises in any scientific and technological application. Due to the increasing volume of data available, the problems modeled by these applications are of larger dimensions and therefore more difficult to solve.

Most processes require the resolution of a system of nonlinear equations of the form $F(X) = 0$, where $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a sufficiently Fréchet differentiable function in D, being D an open convex set.

Very often the complexity of the system $F(X) = 0$ prevent its analytical resolution, and as a consequence its solution is approximated numerically using iterative methods. There is a large classical literature regarding this issue, especially dedicated to the scalar case of solving nonlinear equations [1,2]. For this case, a lot of iterative methods have been designed with very high orders of convergence that allow obtaining quality approximations to the solutions of the problems.

However, for multidimensional nonlinear problems it is important to take into account that the dimensions of the problem difficult the efficiency of iterative schemes to approximate their roots, so that it is more frequent to use methods that, although they have lower orders of convergence, require a lower computational cost and therefore are usually more efficient. In this sense, we recall Traub's iterative scheme [3] with cubic order of convergence and the following iterative expression:

$$\begin{aligned} y^{(k)} &= x^{(k)} - [F'(x^{(k)})]^{-1}F(x^{(k)}), \\ x^{(k+1)} &= y^{(k)} - [F'(x^{(k)})]^{-1}F(y^{(k)}), \end{aligned} \qquad k = 0, 1, 2, \dots, \tag{1}$$

where $F'(x^{(k)})$ is the Jacobian matrix of F in the k-th iteration.

In addition to the order of convergence, the methods can be compared in terms of their efficiency and computational cost. In this sense, Ostrowski introduced in [4] the efficiency index, defined by $EI = p^{1/d}$, where p is the order of convergence of the iterative method and d is the number of different functional evaluations performed on each iteration of the algorithm. Traub's method computes two evaluations of F at the points $x^{(k)}$ and $y^{(k)}$, that is $2n$ functional evaluations, and a Jacobian matrix, with n^2 functional evaluations. Therefore, its efficiency index is

$$EI = 3^{1/(n^2+2n)}.$$

[1] neugarsa@mat.upv.es

Furthermore, Cordero et al. in [5] introduced the computational efficiency index defined by $CI = p^{1/(d+op)}$, where op is the number of products-quotients required per iteration. In Traub's method we must take into account that each iteration requires solving two different linear systems with the same matrix of coefficients, $F'(x^{(k)})$. These kind of systems are solved by Gaussian elimination and need $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ products and quotients. As in Traub's method both systems have the same matrix of coefficients, the total number of products-quotients is only $\frac{1}{3}n^3 + 2n^2 - \frac{1}{3}n$. Then, we have

$$CI = 3^{1/(\frac{1}{3}n^3+3n^2+\frac{5}{3}n)}.$$

In this work, we propose to generalise Traub's method to a two-step family of iterative schemes for solving nonlinear problems with a real parameter in its iterative structure. To this end, in Section 2 we present the iterative structure of the family and we analyze its order of convergence. We devote Section 3 to check numerically the performance of some members of the family to solve two bidimensional systems of nonlinear equations. Finally, we end this paper with some conclusions.

2 Design and convergence of $TM\alpha$ family

Based on Traub's iterative method, in this section we propose a generalization of the scheme by including a real parameter $\alpha \neq 0$. Our aim is to keep the order of convergence and efficiency of Traub's scheme, but at the same time to design a family of methods from which to select the most stable ones for each problem. That is, to select from among the different methods of the family obtained when selecting different values of α those that best approximate the solution and even improve the performance of Traub's scheme.

From Traub's iterative expression in (1), we include the sequence of points

$$z^{(k)} = x^{(k)} + \alpha(y^{(k)} - x^{(k)}), \qquad k = 0, 1, 2, \ldots$$

with a parameter $\alpha \in \mathbb{R}$, $\alpha \neq 0$, and then by making some changes in the second step of (1), we obtain the following family of methods:

$$\begin{aligned} y^{(k)} &= x^{(k)} - [F'(x^{(k)})]^{-1}F(x^{(k)}), \\ z^{(k)} &= x^{(k)} + \alpha(y^{(k)} - x^{(k)}), \\ x^{(k+1)} &= y^{(k)} - \frac{1}{\alpha^2}[F'(x^{(k)})]^{-1}\left((\alpha-1)F(x^{(k)}) + F(z^{(k)})\right), \end{aligned} \qquad k = 0, 1, 2, \ldots \tag{2}$$

Let us note that for each value of $\alpha \in \mathbb{R}$ in (2) we obtain a different method belonging to the iterative family that is denoted by $TM\alpha$. In addition, Traub's method is obtained for $\alpha = 1$.

We can observe that each iteration of $TM\alpha$ family requires three different functional evaluations: two of the vectorial functions $F(x^{(k)})$ and $F(z^{(k)})$, and one of the Jacobian matrix $F'(x^{(k)})$. The following result shows the error equation of $TM\alpha$ family and the sufficient conditions to prove that it has cubic order of convergence for any value of the parameter.

Theorem 11. *Let $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a sufficiently differentiable function in an open convex set D and let us denote by $x^* \in D$ a solution of $F(x) = 0$, such that F' is continuous and nonsingular in x^*. Then, if the initial estimation $x^{(0)}$ is close enough to x^*, family $TM\alpha$ converges to x^* with order of convergence three for any value of $\alpha \neq 0$, being its error equation*

$$e^{(k+1)} = \left(2C_2^2 + (\alpha-1)C_3\right)e^{(k)^3} + \mathcal{O}(e^{(k)^4}),$$

where $e^{(k)} = x^{(k)} - x^*$ is the error in each iteration and $C_j = \frac{1}{j!}[F'(x^*)]^{-1}F^{(j)}(x^*)$, $j \geq 2$.

All the methods of $TM\alpha$ family converge cubically, so all of them have the same efficiency index and computational efficiency index than Traub's method.

After generalizing Traub's iterative scheme, next section is devoted to check numerically the performance of some members of the family. The results are also compared with Traub's method in order to verify that the proposed family includes methods that even improve the performance of Traub's one.

3 Numerical experiments

Next we consider two polynomial systems to test the performance of $TM\alpha$ family. For the numerical implementation we have selected three members of the family, corresponding to $\alpha = 1$ (Traub's method), $\alpha = 10$ and $\alpha = -10$.

As mentioned in Theorem 11, the initial estimate $x^{(0)}$ needs to be close to the solution of the problem to guarantee convergence. Since all the iterative methods require an initial estimate $x^{(0)}$, we have previously performed a study of the basins of attraction of the selected methods for each numerical example. This dynamical study is based on the multidimensional real dynamical tools described in [6, 7]. The basins of attraction are the initial estimates that converge to the roots of the nonlinear function. The study of these initial estimates allows us to compare the performance of the different methods of the family and to determine those that have a greater number of initial estimates converging to the solution we are trying to approximate. All numerical developments in this section have been performed in two dimensions, that is, two nonlinear equations with two variables (x_1 and x_2) in order to be able to represent the basins of attraction of the methods in the dynamical planes.

In the dynamical planes we represent each variable on the coordinate axes. We define a grid of points in the plane, so that each point represents an initial estimate to start calculating iterations of each method. When there is convergence to the root on the nonlinear function, the point is represented in a colour. In other case, the point is represented in black. Therefore, with the dynamical planes we can see the set of points that if taken as initial estimation will converge to the root and we can compare these sets for different methods.

Example 1

Let us consider the following polynomial system

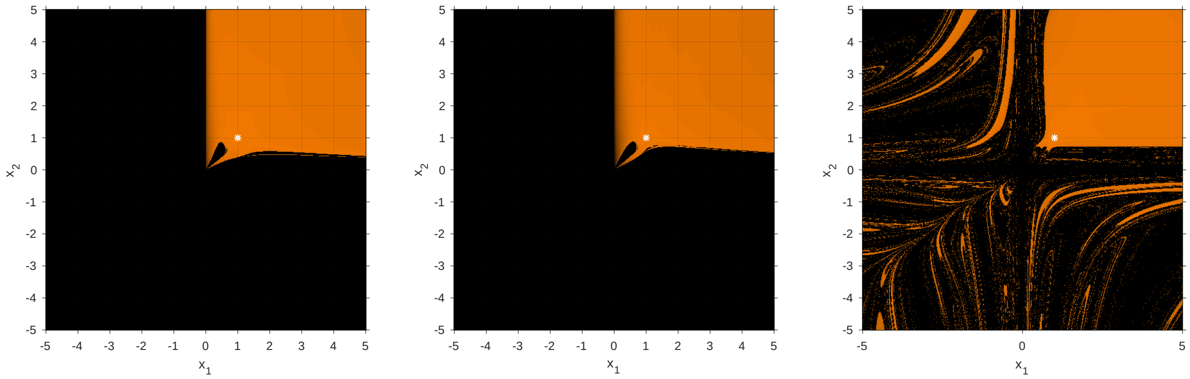
$$\begin{cases} x_1^2 x_2 &= 1, \\ x_2^2 x_1 &= 1, \end{cases}$$

whose only real root is $x^* = (1, 1)$. We are going to approximate x^* using methods of $TM\alpha$ family.

Figure 1 represents the dynamical planes of Example 1, corresponding to Traub's method, and the iterative schemes obtained for $\alpha = 10$ and $\alpha = -10$. These plots have been generated taking a mesh of 500×500 points. The convergence is set when $||x^{(k)} - x^*|| < 10^{-10}$ or the method reach a maximum of 50 iterations. When there is convergence to x^*, the point is depicted in orange. We also represent the root of the nonlinear system with a white star.

We can see in Figure 1 that the method for $\alpha = -10$ improves the stability of Traub's scheme and the corresponding to $\alpha = -10$ as the number of initial estimations converging to x^* (orange colour) is greater in Figure 1a than in Figures 1b and 1c.

Taking into account the basins of attraction of Figure 1, we have selected to test the methods some initial estimations taken from different regions in the plane. Table 1 shows the numerical results obtained for solving Example 1. The table includes the number of iterations until the



(a) $\alpha = 1$ (Traub)   (b) $\alpha = 10$ (c) $\alpha = -10$

Figure 1: Dynamical planes for Example 1.

stopping criteria is reached, the value of $||F(x^{(k+1)})||$ in the last iteration, the difference $||x^{(k+1)} - x^{(k)}||$ between the two last iterates and an approximation to the theoretical order of convergence of the methods by means of the ACOC, defined in [8] by

$$p \approx \frac{\ln\left(||x^{(k+1)} - x^{(k)}||/||x^{(k)} - x^{(k-1)}||\right)}{\ln\left(||x^{(k)} - x^{(k-1)}||/||x^{(k-1)} - x^{(k-2)}||\right)}, \qquad k = 2, 3, \ldots$$

The convergence is set when $||x^{(k+1)} - x^{(k)}|| < 10^{-10}$ or $||F(x^{(k+1)})|| < 10^{-10}$, and the iterations stop when there is no convergence to the roots after 50 iterations. We write nc to indicate this case.

$x^{(0)}$	α	iter	$\lvert\lvert F(x^{(k+1)})\rvert\rvert$	$\lvert\lvert x^{(k+1)} - x^{(k)}\rvert\rvert$	ACOC
$\begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}$	1	4	6.38446e-28	5.97037e-10	2.95792
	10	4	5.83313e-17	1.98129e-6	2.888
	-10	3	6.55344e-26	2.97059e-9	2.1447
$\begin{pmatrix} 4 \\ 3 \end{pmatrix}$	1	5	3.8879e-5	2.10075e-14	2.88947
	10	6	1.27539e-13	2.94634e-5	3.71231
	-10	4	5.01225e-12	9.9243e-5	2.91059
$\begin{pmatrix} -0.5 \\ 2 \end{pmatrix}$	1	nc	-	-	-
	10	nc	-	-	-
	-10	9	7.50067e-17	2.44021e-6	2.99672

Table 1: Numerical results for Example 1.

We can observe in Table 1 that the results agree with the dynamical planes (Figure 1). The best results approximating the solution of the problem are obtained for $\alpha = -10$, since it requires fewer iterations and in some cases it is the only method of the three considered that is convergent. In general, when the methods are convergent, we obtain approximations very close to the solution and with an ACOC close to the cubic order of convergence obtained in Section 2.

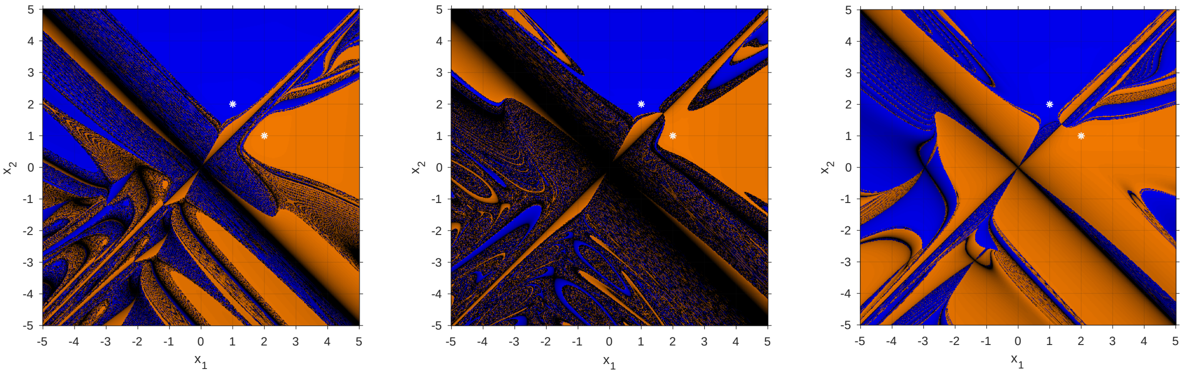
Example 2

Now we consider a polynomial system with two real roots:

$$\left\{\begin{array}{rcl} x_1^3 + x_2^3 & = & 9, \\ x_1^2 x_2 + x_1 x_2^2 & = & 6. \end{array}\right.$$

The real solutions of the system are denoted by $x_1^* = (1,2)$ and $x_2^* = (2,1)$.

For the implementation of the dynamical planes we follow the same criteria as in Figure 1, but for Example 2 we have denoted in blue or orange the convergence to x_1^* or x_2^*, respectively. Figure 2 shows the dynamical planes of Example 2 and the methods of $TM\alpha$ for $\alpha = \{1, 10, -10\}$. We can see again that the basins of attraction of Traub and $\alpha = 10$ are smaller since there are more points in the plane represented in black.



(a) $\alpha = 1$ (Traub)

(b) $\alpha = 10$

(c) $\alpha = -10$

Figure 2: Dynamical planes for Example 2.

As the considered polynomial system has two real roots, we show in different tables the numerical results obtained when we are trying to approximate each solution. In this sense, we have select from Figure 2 initial points represented in blue or orange for Tables 3 and 4 in order to analyze the convergence to x_1^* or x_2^*, respectively. Similar results to those in Table 1 are obtained, since the method obtained for $\alpha = -10$ is always convergent to the corresponding root even in cases where the other methods are not convergent.

$x^{(0)}$	α	iter	$\|\|F(x^{(k+1)})\|\|$	$\|\|x^{(k+1)} - x^{(k)}\|\|$	ACOC
$\begin{pmatrix} 1.5 \\ 3 \end{pmatrix}$	1	4	4.88318e-27	9.43998e-10	2.95792
	10	4	4.46148e-16	3.13269e-6	2.888
	-10	3	5.01241e-25	4.69692e-9	2.1447
$\begin{pmatrix} -4.5 \\ 1 \end{pmatrix}$	1	nc	-	-	-
	10	nc	-	-	-
	-10	10	9.60177e-18	1.33838e-6	2.1109
$\begin{pmatrix} 4.5 \\ -1 \end{pmatrix}$	1	nc	-	-	-
	10	nc	-	-	-
	-10	10	1.87412e-21	6.53728e-8	2.88154

Table 2: Numerical results for Example 2 approximating $x_1^* = (1,2)$.

$x^{(0)}$	α	iter	$\|\|F(x^{(k+1)})\|\|$	$\|\|x^{(k+1)} - x^{(k)}\|\|$	ACOC
$\begin{pmatrix}4\\1\end{pmatrix}$	1	4	6.06358e-11	2.08324e-4	2.68829
	10	5	3.8871e-17	1.79614e-6	3.02901
	-10	4	1.89885e-22	3.04763e-8	2.71486
$\begin{pmatrix}2\\-3.5\end{pmatrix}$	1	7	3.50169e-11	1.73502e-4	2.59589
	10	nc	-	-	-
	-10	10	5.54428e-22	4.35592e-8	2.88959
$\begin{pmatrix}-1.5\\-4\end{pmatrix}$	1	nc	-	-	-
	10	nc	-	-	-
	-10	10	4.33411e-28	4.43146e-10	2.88404

Table 3: Numerical results for Example 2 approximating $x_2^* = (2, 1)$.

4 Conclusions

In this paper, we propose a family of iterative methods with a real parameter α. This family is a generalization of Traub's iterative scheme, obtained for $\alpha = 1$. After determining the sufficient conditions to obtain a cubic order of convergence for any value of the parameter, a numerical analysis of some schemes of the family when applied to approximate the solution of systems of nonlinear equations is carried out. This study allows verifying that the theoretical developments are correct and that the family includes methods that improve the stability of the classical Traub's iterative scheme.

Acknowledgements: This research was partially supported by Spanish Ministerio de Ciencia, Innovación y Universidades PGC2018-095896-B-C22 and by the internal research project MICoCo of Universidad Internacional de La Rioja (UNIR).

References

[1] Amat S., Busquier S. Advances in Iterative Methods for Nonlinear Equations. Springer, 2016.

[2] Petković, M.S., Neta B., Petković L.D., Džunić J. Multipoint Methods for Solving Nonlinear Equations. Elsevier, 2013.

[3] Traub, J.F. Iterative Methods for the Solution of Equations. Prentice-Hall, New York, 1964.

[4] Ostrowski, A. Solution of Equations and Systems of Equations. Prentice-Hall, Englewood Cliffs, NJ, 1964.

[5] Cordero, A., Hueso, J.L., Martínez, E., Torregrosa, J.R. A modified Newton-Jarratt's composition *Numerical Algorithms*, 55:87–99, 2010.

[6] Campos, B., Cordero, A., Torregrosa, J.R., Vindel, P. A multidimensional dynamical approach to iterative methods with memory *Applied Mathematics and Computation*, 271:701–715, 2015.

[7] 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.

[8] Cordero, A., Torregrosa, J.R. Variants of Newton's method using fifth order quadrature formulas *Applied Mathematics and Computation*, 190:686-698, 2007.

20

Multidimensional extension of conformable fractional iterative methods for solving nonlinear problems

Giro Candelario[♭,1], Alicia Cordero[♮], Juan R. Torregrosa[♮] and María P. Vassileva[♭]

(♭) Área de Ciencias Básicas y Ambientales, Instituto Tecnológico de Santo Domingo
Avenida de Los Próceres #49, Los Jardines del Norte 10602,
Santo Domingo, Dominican Republic.

(♮) Departmento de Matemática Aplicada, Universitat Politècnica de València
Camino de Vera, s/n 46022 Valencia, Spain.

1 Introduction

Fractional calculus is an extension of classical calculus, and the theoretical aspects from this are held. Many problems can be described by using fractional calculus, because of the higher degree of freedom compared to classical calculus [1, 2].

In recent years, some Newton-type methods for solving nonlinear equations have been proposed by using the Riemann-Liouville, Caputo and conformable fractional derivatives [3–5].

First, let us introduce some preliminary concepts related to conformable derivative. The left conformable fractional derivative of a function $f : [a, \infty) \longrightarrow \mathbb{R}$, starting from a, of order $\alpha \in (0, 1]$, $\alpha, a, x \in \mathbb{R}$, $a < x$, is [6]

$$(T^a_\alpha f)(x) = \lim_{\varepsilon \longrightarrow 0} \frac{f(x + \varepsilon(x - a)^{1-\alpha}) - f(x)}{\varepsilon}. \tag{1}$$

If the limit exists, f is α-differentiable. If f is also differentiable, $(T^a_\alpha f)(x) = (x - a)^{1-\alpha} f'(x)$. If f is α-differentiable in (a, b), for some $b \in \mathbb{R}$, $(T^a_\alpha f)(a) = \lim_{x \to a^+} (T^a_\alpha f)(x)$. Note that $T^a_\alpha C = 0$, where C is a constant.

Newly, a Newton-type method by using conformable derivative was designed for solving nonlinear equations in [5] with the following iterative expression:

$$x_{k+1} = a + \left((x_k - a)^\alpha - \alpha \frac{f(x_k)}{(T^a_\alpha f)(x_k)} \right)^{1/\alpha}, \quad k = 0, 1, 2, \ldots \tag{2}$$

where $(T^a_\alpha f)(x_k)$ is the left conformable fractional derivative of order α, $\alpha \in (0, 1]$, starting from a, $a < x_k$, $\forall k$. When $\alpha = 1$, we obtain the classical Newton-Raphson method. The quadratic convergence of this method is stated in [5] by using an appropriate conformable Taylor series [7].

The method proposed in [5], as seen in equation (1), can be used only to solve scalar problems. To design a conformable vectorial Newton-type method in order to find the solution $\bar{x} \in \mathbb{R}^n$ of

[1] giro.candelario@intec.edu.do

a nonlinear system $F(x) = \hat{0}$, with coordinate functions $f_1, \ldots, f_n$, where $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a sufficiently Fréchet-differentiable function in an open convex set D, we have to introduce the necessary existing concepts and results.

We can find in [8] a definition of conformable partial derivative:

Definition 1. *Let f be a function in n variables, $x_1, \ldots, x_n$, the conformable partial derivative of f of order $\alpha \in (0, 1]$ in $x_i > a = 0$ is defined as:*

$$\frac{\partial_0^\alpha}{\partial x_i^\alpha} f(x_1, \ldots, x_n) = \lim_{\epsilon \to 0} \frac{f(x_1, \ldots, x_i + \epsilon x_i^{1-\alpha}, \ldots, x_n) - f(x_1, \ldots, x_n)}{\epsilon}, \tag{3}$$

In [8] we can also find a definition of conformable Jacobian matrix:

Definition 2. *Let f, g be functions in 2 variables x and y, and the respective partial derivatives exist and are continuous. Being $x > a_1$ and $y > a_2$, where $a = (a_1, a_2) = (0, 0) = \hat{0}$, then the conformable Jacobian matrix is:*

$$F_{\hat{0}}^{\alpha(1)}(x) = \begin{pmatrix} \dfrac{\partial_0^\alpha f}{\partial x^\alpha} & \dfrac{\partial_0^\alpha f}{\partial y^\alpha} \\ \dfrac{\partial_0^\alpha g}{\partial x^\alpha} & \dfrac{\partial_0^\alpha g}{\partial y^\alpha} \end{pmatrix} = \begin{pmatrix} x^{1-\alpha} \dfrac{\partial f}{\partial x} & y^{1-\alpha} \dfrac{\partial f}{\partial y} \\ x^{1-\alpha} \dfrac{\partial g}{\partial x} & y^{1-\alpha} \dfrac{\partial g}{\partial y} \end{pmatrix}. \tag{4}$$

This concept can be directly extended to higher dimensions.

The new concepts and results required to design a conformable vectorial Newton-type method are stated in the next Section.

2 Methods

2.1 New concepts and results

Considering that in equation (3), $x_i \in (0, \infty)$, it can be defined the conformable partial derivative in $x_i \in (a, \infty)$:

Definition 3. *Let f be a function in n variables, $x_1, \ldots, x_n$, the conformable partial derivative of f of order $0 < \alpha \leq 1$ in $x_i \in (a, \infty)$ is*

$$\frac{\partial_a^\alpha}{\partial x_i^\alpha} f(x_1, \ldots, x_n) = \lim_{\epsilon \to 0} \frac{f(x_1, \ldots, x_i + \epsilon (x_i - a)^{1-\alpha}, \ldots, x_n) - f(x_1, \ldots, x_n)}{\epsilon}. \tag{5}$$

When $x_i = a$, $\dfrac{\partial_a^\alpha}{\partial x_i^\alpha} f(x_1, \ldots, a, \ldots, x_n) = \lim\limits_{x_i \to a^+} \dfrac{\partial_a^\alpha}{\partial x_i^\alpha} f(x_1, \ldots, x_i, \ldots, x_n)$.

This conformable partial derivative is linear, and the product, quotient and chain rules are held, like conformable derivative in [6].

It can also be stated the definition of conformable Jacobian matrix for $x_1 \in (a_1, \infty)$ and $x_2 \in (a_2, \infty)$, being $x = (x_1, x_2)$ and $a = (a_1, a_2)$:

Definition 4. *Let f and g be the coordinate functions of a vector valued function $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ with variables $x_1 > a_1$ and $x_2 > a_2$, being $x = (x_1, x_2)$ and $a = (a_1, a_2)$, such that the respective partial derivatives exist and are continuous. The conformable Jacobian matrix is*

$$F_a^{\alpha(1)}(x) = \begin{pmatrix} \dfrac{\partial_{a_1}^\alpha f}{\partial x_1^\alpha} & \dfrac{\partial_{a_2}^\alpha f}{\partial x_2^\alpha} \\ \dfrac{\partial_{a_1}^\alpha g}{\partial x_1^\alpha} & \dfrac{\partial_{a_2}^\alpha g}{\partial x_2^\alpha} \end{pmatrix} = \begin{pmatrix} (x_1 - a_1)^{1-\alpha} \dfrac{\partial f}{\partial x_1} & (x_2 - a_2)^{1-\alpha} \dfrac{\partial f}{\partial x_2} \\ (x_1 - a_1)^{1-\alpha} \dfrac{\partial g}{\partial x_1} & (x_2 - a_2)^{1-\alpha} \dfrac{\partial g}{\partial x_2} \end{pmatrix}. \tag{6}$$

This concept can also be directly extended to higher dimensions.

Likewise in [7] (Theorem 4.1), we can get a new Taylor series, where the conformable derivatives start from some point $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ distinct from another point $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$ where they are being evaluated:

Theorem 1. *Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be an infinitely α-differentiable vector valued function, for $\alpha \in (0,1]$, around some point $b_i \in (a_i, \infty)$, $\forall i = 1, \ldots, n$, being $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$. Then, F has the conformable Taylor power series*

$$F(t) = F(b) + \frac{F_a^{\alpha(1)}(b)}{\alpha}\Delta + \frac{F_a^{\alpha(2)}(b)}{2!\alpha^2}\Delta^2 + \cdots, \tag{7}$$

being $\Delta = H^{\odot\alpha} - L^{\odot\alpha}$; $H = t - a$, $L = b - a$, where $\odot$ is the Hadamard power.

In addition, in order to perform the convergence analysis, another concept has to be introduced.

Theorem 2. *Let $x, y \in \mathbb{R}^n$, $r \in \mathbb{R}$, and be $\odot$ the Hadamard product/power. The Newton's binomial theorem for fractional power and vector values is*

$$(x+y)^{\odot r} = \sum_{k=0}^{\infty} \binom{r}{k} x^{\odot(r-k)} \odot y^{\odot k}, \quad k \in \{0\} \cup \mathbb{N}, \tag{8}$$

where the generalized binomial coefficient (see [9]) is

$$\binom{r}{k} = \frac{\Gamma(r+1)}{k!\Gamma(r-k+1)}, \quad k \in \{0\} \cup \mathbb{N}. \tag{9}$$

Now, we set the design of conformable Newton-type method for solving nonlinear systems.

2.2 Design and convergence analysis

As we can see in [5], let us consider the approximation of a function F with the Taylor power series (7) up to order one, evaluated at the solution $\bar{x}$ of $F(x) = \hat{0}$:

$$F(x) \approx F(\bar{x}) + \frac{F_a^{\alpha(1)}(\bar{x})}{\alpha}\Delta. \tag{10}$$

Since $F(\bar{x}) = \hat{0}$, and $\Delta = H^{\odot\alpha} - L^{\odot\alpha}$; $H = x - a$, $L = \bar{x} - a$,

$$F(x) \approx \frac{F_a^{\alpha(1)}(\bar{x})}{\alpha}\left[(x-a)^{\odot\alpha} - (\bar{x}-a)^{\odot\alpha}\right]. \tag{11}$$

Multiplying both sides of (11), by $\alpha\left[F_a^{\alpha(1)}(\bar{x})\right]^{-1}$ from the left,

$$\alpha\left[F_a^{\alpha(1)}(\bar{x})\right]^{-1} F(x) \approx (x-a)^{\odot\alpha} - (\bar{x}-a)^{\odot\alpha}. \tag{12}$$

From $(\bar{x}-a)^{\odot\alpha}$, we can get $\bar{x}$, so

$$\bar{x} \approx a + \left((x-a)^{\odot\alpha} - \alpha\left[F_a^{\alpha(1)}(\bar{x})\right]^{-1} F(x)\right)^{\odot 1/\alpha}. \tag{13}$$

Considering iterates $x^{(k)}$ and $x^{(k+1)}$ as approximations of solution $\bar{x}$, we get the conformable Newton-type method for nonlinear systems:

$$x^{(k+1)} = a + \left[\left(x^{(k)} - a\right)^{\odot\alpha} - \alpha\left[F_a^{\alpha(1)}\left(x^{(k)}\right)\right]^{-1} F\left(x^{(k)}\right)\right]^{\odot 1/\alpha}, \; k = 0, 1, 2, \ldots \tag{14}$$

Next result shows that quadratic convergence of vectorial Newton-type method (14) by using the conformable Taylor series (7) is obtained.

Theorem 3. *Let $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a continuous function in an open convex set $D \subseteq \mathbb{R}^n$ holding a zero $\bar{x} \in \mathbb{R}^n$ of a vector valued function $F(x)$. Let $F_a^{\alpha(1)}(x)$ be the conformable Jacobian matrix of F starting at $a \in \mathbb{R}^n$, of order α, for any $\alpha \in (0,1]$. Let us suppose that $F_a^{\alpha(1)}(x)$ is continuous and non-singular at $\bar{x}$. If an starting point $x^{(0)} \in \mathbb{R}^n$ is quite close to $\bar{x}$, then the local order of convergence of conformable vectorial Newton's method*

$$x^{(k+1)} = a + \left[\left(x^{(k)} - a\right)^{\odot\alpha} - \alpha\left[F_a^{\alpha(1)}\left(x^{(k)}\right)\right]^{-1} F\left(x^{(k)}\right)\right]^{\odot 1/\alpha}, \quad k = 0,1,2,\ldots$$

is at least 2, and the error equation is

$$e^{(k+1)} = \alpha C_2(\bar{x} - a)^{\odot(\alpha-1)} e^{(k)^2} + O\left(e^{(k)^3}\right), \tag{15}$$

where $C_j = \dfrac{1}{j!\alpha^{j-1}}\left[F_a^{\alpha(1)}(\bar{x})\right]^{-1} F_a^{\alpha(j)}(\bar{x})$, $j = 2,3,4,\ldots$, *such that* $a < x^{(k)}$, $\forall k$.

Next, some numerical tests with some nonlinear systems of equations are made. We comment that, in all tests, a comparison with classical Newton-Raphson's method (when $\alpha = 1$) is being made. The dependence on initial estimates of both methods is also analyzed.

3 Results

The numerical tests are made by using Matlab R2020a with double precision arithmetic, $\|F(x^{(k+1)})\| < 10^{-8}$ or $\|x^{(k+1)} - x^{(k)}\| < 10^{-8}$ as stopping criterium, and a maximum of 500 iterations. We used $a = (a_1, \ldots, a_n) = (-10, \ldots, -10)$ for each test to be sure that $a_i < x_i$, $\forall i = 1, \ldots, n$, as seen in Definitions 3 and 4, and $a < x^{(k)}$, $\forall k$, according to Theorem 3. We also use the Approximated Computational Order of Convergence (ACOC)

$$ACOC = \frac{\ln(\|x^{(k+1)} - x^{(k)}\|/\|x^{(k)} - x^{(k-1)}\|)}{\ln(\|x^{(k)} - x^{(k-1)}\|/\|x^{(k-1)} - x^{(k-2)}\|)}, \quad k = 0,1,2,\ldots,$$

introduced in [10], to verify the theoretical order of convergence is got in practice, and $\alpha \in (0,1]$.

Our test function to vector values is $F(x,y) = (x^2 - 2x - y + 0.5, x^2 + 4y^2 - 4)^T$ with real and complex roots $\bar{x}_1 \approx (-0.2222, 0.9938)^T$, $\bar{x}_2 \approx (1.9007, 0.3112)^T$ and $\bar{x}_3 \approx (1.1608 - 0.6545i, -0.9025 - 0.2104i)^T$. The conformable Jacobian matrix of $F(x,y)$ is

$$F_a^{\alpha(1)}(x,y) = \begin{pmatrix} (x-a_1)^{1-\alpha}(2x-2) & (y-a_2)^{1-\alpha}(-1) \\ (x-a_1)^{1-\alpha}(2x) & (y-a_2)^{1-\alpha}(8y) \end{pmatrix},$$

where $a = (a_1, a_2) = (-10, -10)$.

α	$\bar{x}$	$\|F(x^{(k+1)})\|$	$\|x^{(k+1)} - x^{(k)}\|$	iter	ACOC
1	-	-	-	> 500	-
0.9	$\bar{x}_3$	5.40×10^{-11}	5.86×10^{-6}	54	2.00
0.8	$\bar{x}_3$	9.77×10^{-9}	7.87×10^{-5}	86	2.00
0.7	$\bar{x}_3$	2.27×10^{-14}	4.75×10^{-8}	36	1.98
0.6	$\bar{x}_2$	2.95×10^{-10}	1.16×10^{-5}	23	2.05
0.5	$\bar{x}_2$	4.89×10^{-10}	1.48×10^{-5}	122	2.05
0.4	$\bar{x}_2$	5.34×10^{-13}	5.01×10^{-7}	86	2.04
0.3	$\bar{x}_2$	4.94×10^{-10}	1.79×10^{-5}	35	2.03
0.2	$\bar{x}_2$	1.16×10^{-14}	6.76×10^{-8}	21	1.98
0.1	$\bar{x}_2$	2.16×10^{-10}	1.08×10^{-5}	39	2.06

Table 1: Results for $F(x,y) = \hat{0}$ with initial estimation $x^{(0)} = (-2, -1.5)^T$

α	$\bar{x}$	$\lVert F(x^{(k+1)})\rVert$	$\lVert x^{(k+1)}-x^{(k)}\rVert$	iter	ACOC
1	$\bar{x}_1$	8.31×10^{-11}	7.29×10^{-6}	5	2.00
0.9	$\bar{x}_1$	5.94×10^{-11}	6.15×10^{-6}	5	2.00
0.8	$\bar{x}_1$	4.21×10^{-11}	5.17×10^{-6}	5	2.00
0.7	$\bar{x}_1$	2.97×10^{-11}	4.33×10^{-6}	5	2.00
0.6	$\bar{x}_1$	2.09×10^{-11}	3.62×10^{-6}	5	2.00
0.5	$\bar{x}_1$	1.45×10^{-11}	3.01×10^{-6}	5	2.00
0.4	$\bar{x}_1$	1.01×10^{-11}	2.49×10^{-6}	5	2.00
0.3	$\bar{x}_1$	6.97×10^{-12}	2.06×10^{-6}	5	2.00
0.2	$\bar{x}_1$	4.80×10^{-12}	1.69×10^{-6}	5	2.00
0.1	$\bar{x}_1$	3.30×10^{-12}	1.39×10^{-6}	5	2.00

Table 2: Results for $F(x,y)=\hat{0}$ with initial estimation $x^{(0)}=(-2,1.5)^T$

In Table 1, we can see for $F(x,y)$ that classical Newton's method (when $\alpha=1$) does not get any solution in 500 iterations, while conformable vectorial Newton's procedure converges. We observe also that ACOC may be even slightly greater than 2 when $\alpha\neq 1$. Note also that complex root $\bar{x}_3$ is found with real initial estimate.

In Table 2, we can observe for $F(x,y)$, with a different initial estimation, that classical Newton's scheme and conformable Newton's method have a similar behaviour, regarding the number of iterations and the ACOC. Once again, the quadratic convergence of conformable Newton's method is held for any $\alpha\in(0,1]$.

In order to study the stability of conformable vectorial Newton's method tested above, we analyze the dependence on initial estimates by observing convergence planes, which is defined in [11], and is also employed in [3–5].

For constructing convergence planes we consider from initial estimates (x_0,y_0), the points x_0 in the horizontal axis, and values of $\alpha\in(0,1]$ in the vertical axis. Each one of 2 planes in the figure is performing a distinct value of y_0 from initial estimates (x_0,y_0). Each color in the planes represents different solutions, and when it is painted in black no solution was found in 500 iterations. Each plane is generated by a 400×400 grid, with a maximum of 500 iterations, and a tolerance of 0.001.

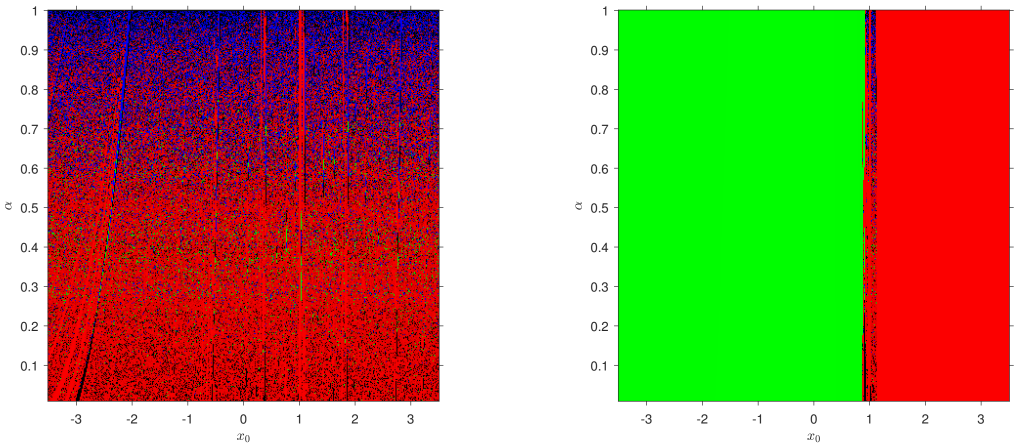
In Figure 1, it can be observed for $F(x,y)$ that in (b) is obtained almost 100% of convergence, while in (a) is obtained around 86% of convergence. This method converges to all roots for each case, even to complex root with real initial estimates.

We can also see, in general, it is possible to get several solutions with the same initial estimate by choosing different values of α.

4 Conclusions

The first conformable fractional Newton-type iterative method for solving nonlinear systems has been designed. We have introduced the analytical implements needed for the construction of this method. The convergence analysis has been made, and quadratic convergence of classical Newton's method is held. Numerical tests have been made, and the dependence on initial estimates was analyzed, sustaining the theory. We could see that conformable vectorial Newton-type method shows, in some cases, a better numerical behaviour than classical one in terms of number of iterations, ACOC, and wideness of basins of attractions of the roots. We could also see that complex roots may be found with real initial estimates, and several roots may be obtained with the same initial estimate by choosing distinct values for α.

Acknowledgement: This research was partially supported by PGC2018-095896-B-C22 (MCIU /AEI/10.13039/501100011033/FEDER Una manera de hacer Europa, UE) and by Dominican Re-



(a) $y_0 = -1.5$, 85.99% of convergence (b) $y_0 = 1.5$, 99.62% of convergence

Figure 1: Convergence planes of $F(x, y)$. $\bar{x}_1$: green, $\bar{x}_2$: red, $\bar{x}_3$: blue

public FONDOCYT 2018-2019-1D2-140.

References

[1] K.S. Miller, An Introduction to Fractional Calculus and Fractional Differential Equations. New York, J. Wiley and Sons, 1993.

[2] I. Podlubny, Fractional Differential Equations. New York, Academic Press, 1999.

[3] A. Akgül, A. Cordero, J.R. Torregrosa, A fractional Newton method with 2αth-order of convergence and its stability *Appl. Math. Letters*, 98: 344–351, 2019.

[4] G. Candelario, A. Cordero, J.R. Torregrosa, Multipoint Fractional Iterative Methods with $(2\alpha + 1)$th-Order of Convergence for Solving Nonlinear Problems *Mathematics*, 452: 2020, https://doi.org/10.3390/math8030452.

[5] G. Candelario, A. Cordero, J.R. Torregrosa, M.P. Vassileva, An optimal and low computational cost fractional Newton-type method for solving nonlinear equations *Appl. Math. Letters*, 124: 2022, https://doi.org/10.1016/j.aml.2021.107650.

[6] T. Abdeljawad, On conformable fractional calculus *Comput. Appl. Math.*, 279: 57–66, 2015.

[7] Ş. Toprakseven, Numerical Solutions of Conformable Fractional Differential Equations by Taylor and Finite Difference Methods *Natural Appl. Sci.*, 23: 850–863, 2019.

[8] A. Atangana, D. Baleanu, A. Alsaedi, New properties of conformable derivative *Open Math.*, 13: 889–898, 2015.

[9] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions. Washington, D.C., Dover, 1970.

[10] A. Cordero, J.R. Torregrosa, Variants of Newton's method using fifth order quadrature formulas *Appl. Math. Comput.*, 190: 686–698, 2007.

[11] A.Á. Magreñán, A new tool to study real dynamics: The convergence plane *Appl. Math. Comput.*, 248: 215–224, 2014.

21

Application of Data Envelopment Analysis to the evaluation of biotechnological companies

B. Latorre-Scilingo[♭], S. González-de-Julián[♮,1] and I. Barrachina-Martínez[♭]

(♭) Research Centre for Economics Engineering,
Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain.

1 Introduction

Nasdaq Biotechnology (NBI) is a stock market index that includes companies related to biotechnology sector which are listed on Nasdaq Market. It was created in 1993 with base 200 and it's currently composed by 373 companies that meet the inclusion's criteria. Biotechnological sector has grown in recent years, and biotechnological companies are characterized by low income, many intangible assets (from product patents), high expenses in research and development and many of them present losses [1–3]. The companies included in NBI reveal differences in terms of size, where there are few companies with many employees and many small companies with a reduced number of employees [1].

In addition to financial analysis, investors have recently begun to consider environmental, social and governance factors (ESG) in their decisions that affect value creation and risk of companies. According to a report by the Organization for Economic Cooperation and Development, only 25% of listed companies in the USA have sustainability rating, compared to 10% in Europe and 5% in Japan. However, companies rated represent 95% of total market capitalization in USA while in Europe and Japan, that percentage is reduced to 89% and 78%, respectively [4]. Companies that are not rated have lower trading volume [5].

Nowadays, there are more than 125 agencies dedicated to rate companies in terms of sustainability, some of them are: Bloomberg ESG Data Service, ISS, MSCI ESG Risk, S&P Sustainability, Sustainalytics y Refinitiv. Sustainalitycs' ESG risk rating turns out to be the most stable over time and considers all sustainable factors as a whole [6]. To do so, Sustainalitycs evaluates firms according to their exposure to risk management and assigns them a score that reflects their level of ESG risk level. The higher the score, the worse their rating since the company is considered riskier.

Another important aspect that investors consider is efficiency. It indicates the optimization of resources used to obtain results [7]. Efficiency is the relationship between inputs and outputs of a productive system. Data Envelopment Analysis (DEA) is used to study the efficiency of several sectors and stock markets, highlighting sectors such as: banking, energy, tourism, among others. It is a non-parametrical deterministic technique, that uses mathematical programming, and an efficient frontier is obtained as a result [8].

The objective of this study is to evaluate the efficiency of the companies that are included in

[1] silgonde@upv.es

Nasdaq Biotechnology based on financial and stock market data, and its relationship with ESG risk rating carried out by Sustainalitycs.

2 Methods

For each of the 373 companies include in NBI, the followings variables are collected from website www.investing.com at 31st December 2021: number of indexes where the company is listed, year of initial public offering, employees, shares outstanding, price per share, assets, liabilities, equity, revenue, operating income and net income.

From the website www.sustainalytics.com, the sustainability rating of the companies is obtained. Those with an ESG risk score less than 30 points are classified as "sustainable", on the contrary, those that exceed this value are considered "unsustainable".

Based on the data collected, the following variables are calculated: return on equity (ROE), return on assets (ROA), market capitalization, age of the value listing in the market and level of debt.

For the study, those companies with risk rating and positive net worth are included. A factorial analysis is carried out with the principal component factor extraction method to determine the dimensions that best explain the variability of this sample of companies. According to the results of this analysis, each company is assigned the score obtained in each of the resulting factors. And finally, the companies that turn out to have a similar size from the previous analysis are considered.

To evaluate the efficiency, Data Envelopment Analysis (DEA) is used, in which each company constitutes a Decision-Making Unit (DMU). The following inputs are included: total assets, number of employees, level of debt, and the number of years that the company is listed (non-controllable input). As outputs are considered: ROE, ROA, the annual variation in market capitalization and the level of ESG risk (undesirable output).

The methodology with input and output orientation and variable returns to scale is used. This analysis is expressed in its input orientation according to equation (1.1), and output orientation equation (1.2):

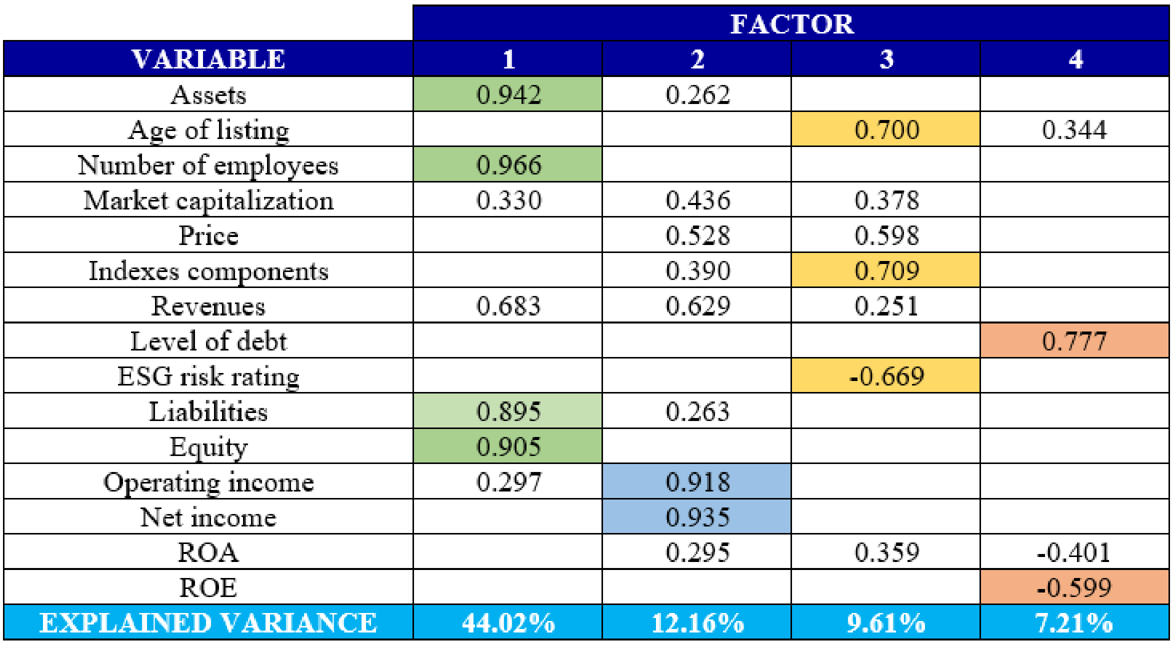
$$\begin{gathered} min\ \theta - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) \\ subject\ to \\ \sum_{j=1}^{n}\lambda_j x_{ij} + s_i^- = \theta x_{io}\ i=1,2,...,m;\ \sum_{j=1}^{n}\lambda_j y_{rj} + s_r^+ = y_{ro}\ r=1,2,...,s; \\ \lambda_j \geq 0\ j=1,2,...,n;\ \sum_{j=1}^{n}\lambda_j = 1 \end{gathered} \tag{1}$$

$$\begin{gathered} max\ \phi - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) \\ subject\ to \\ \sum_{j=1}^{n}\lambda_j x_{ij} + s_i^- = x_{io}\ i=1,2,...,m;\ \sum_{j=1}^{n}\lambda_j y_{rj} + s_r^+ = \phi y_{ro}\ r=1,2,...,s; \\ \lambda_j \geq 0\ j=1,2,...,n;\ \sum_{j=1}^{n}\lambda_j = 1 \end{gathered} \tag{2}$$

Where x_{ij} ($x_{ij} \geq 0$) represents the quantities of input i (i = 1, 2,..., m) consumed by the j-th unit and y_{rj} ($y_{ij} \geq 0$) represents the observed quantities of output r (r = 1, 2,...,s) produced by the j-th unit.

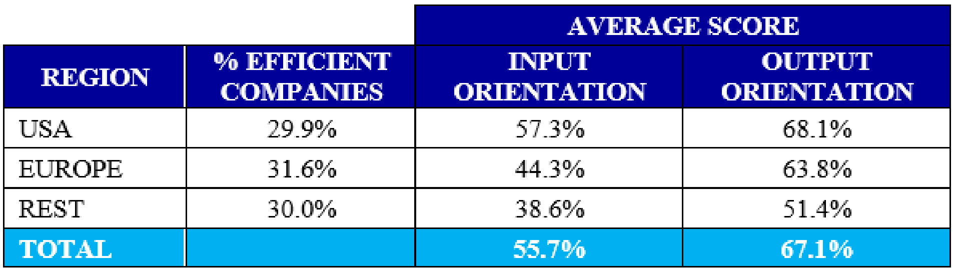
3 Results

Of the 373 companies that are included in the NBI, 268 meet the required criteria of having ESG risk rating, at least 1 year being listed and positive net worth. When performing the factorial analysis, figure 1 shows that 4 factors are obtained, which explain 73% of the variance. Factor 1 is the one that explains the most variability (44%) and is related to the dimension or size of the company. Factor 2 refers to financial results. Factor 3 is related to age, indexes and sustainability performance. And factor 4 identifies those companies that are more indebted and perform worse results in terms of financial profitability.



	FACTOR			
VARIABLE	1	2	3	4
Assets	0.942	0.262		
Age of listing			0.700	0.344
Number of employees	0.966			
Market capitalization	0.330	0.436	0.378	
Price		0.528	0.598	
Indexes components		0.390	0.709	
Revenues	0.683	0.629	0.251	
Level of debt				0.777
ESG risk rating			-0.669	
Liabilities	0.895	0.263		
Equity	0.905			
Operating income	0.297	0.918		
Net income		0.935		
ROA		0.295	0.359	-0.401
ROE				-0.599
EXPLAINED VARIANCE	**44.02%**	**12.16%**	**9.61%**	**7.21%**

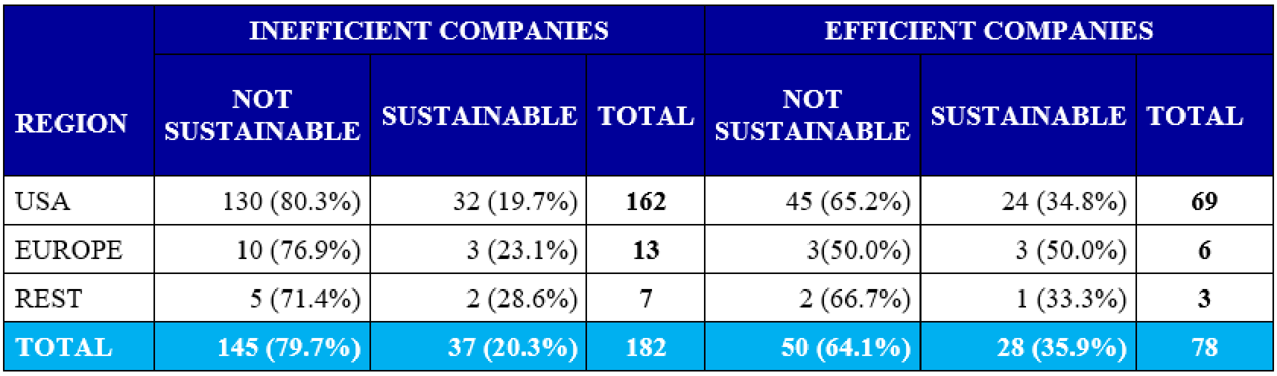
Figure 1: Factorial analysis results



		AVERAGE SCORE	
REGION	% EFFICIENT COMPANIES	INPUT ORIENTATION	OUTPUT ORIENTATION
USA	29.9%	57.3%	68.1%
EUROPE	31.6%	44.3%	63.8%
REST	30.0%	38.6%	51.4%
TOTAL		**55.7%**	**67.1%**

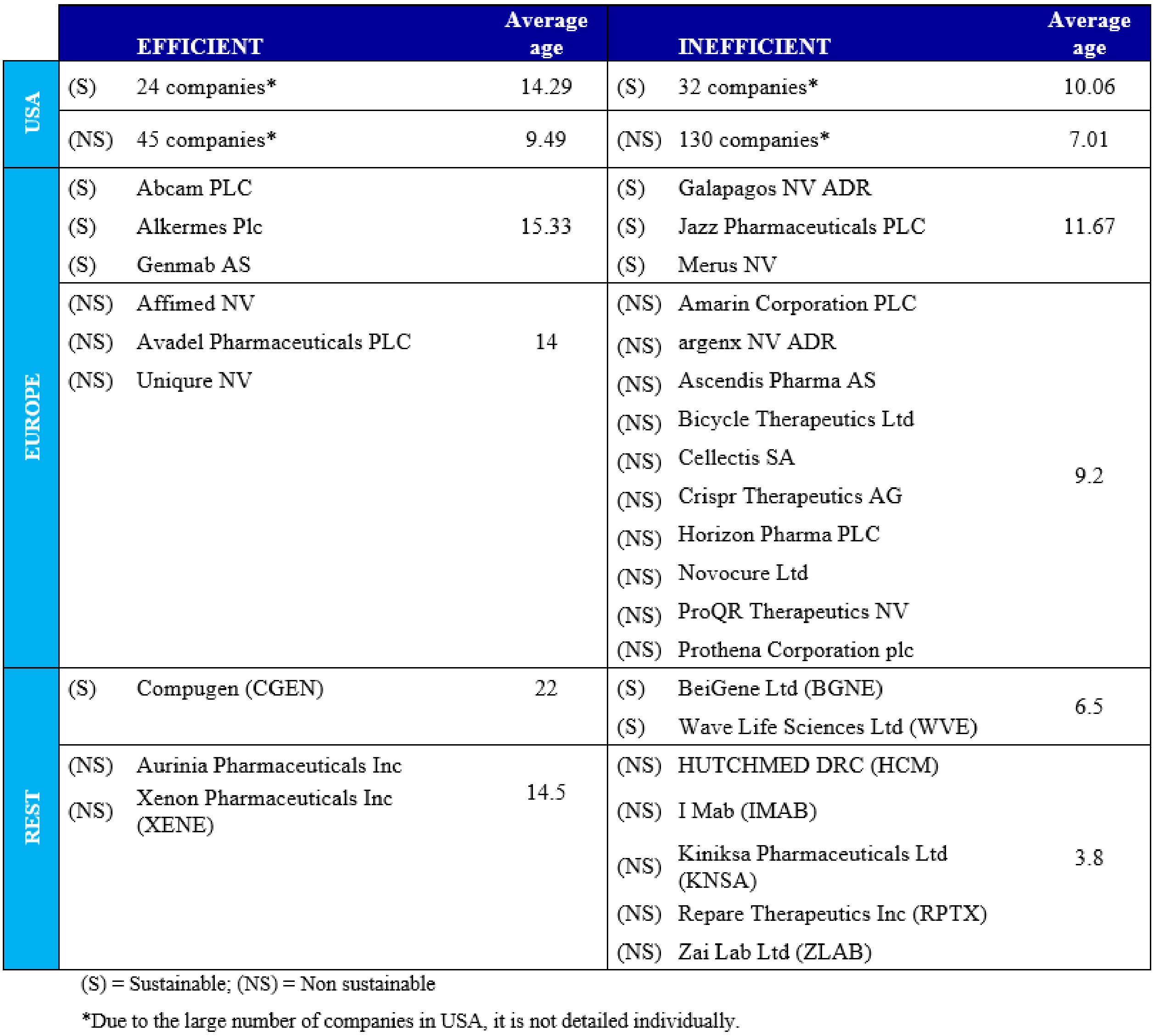
Figure 2: Efficiency by region

To assess their efficiency, 260 companies are finally included whose factorial score of factor 1 is between -1 and 1. Out of these 260 companies, 90% belong to the USA, 7% to Europe and 3% to the rest of the world. 65 of these companies (25%) are sustainable.



REGION	INEFFICIENT COMPANIES			EFFICIENT COMPANIES		
	NOT SUSTAINABLE	SUSTAINABLE	TOTAL	NOT SUSTAINABLE	SUSTAINABLE	TOTAL
USA	130 (80.3%)	32 (19.7%)	**162**	45 (65.2%)	24 (34.8%)	**69**
EUROPE	10 (76.9%)	3 (23.1%)	**13**	3(50.0%)	3 (50.0%)	**6**
REST	5 (71.4%)	2 (28.6%)	**7**	2 (66.7%)	1 (33.3%)	**3**
TOTAL	**145 (79.7%)**	**37 (20.3%)**	**182**	**50 (64.1%)**	**28 (35.9%)**	**78**

Figure 3: Relationship between efficiency companies and sustainability



		EFFICIENT	Average age		INEFFICIENT	Average age
USA	(S)	24 companies*	14.29	(S)	32 companies*	10.06
	(NS)	45 companies*	9.49	(NS)	130 companies*	7.01
EUROPE	(S)	Abcam PLC	15.33	(S)	Galapagos NV ADR	11.67
	(S)	Alkermes Plc		(S)	Jazz Pharmaceuticals PLC	
	(S)	Genmab AS		(S)	Merus NV	
	(NS)	Affimed NV	14	(NS)	Amarin Corporation PLC	9.2
	(NS)	Avadel Pharmaceuticals PLC		(NS)	argenx NV ADR	
	(NS)	Uniqure NV		(NS)	Ascendis Pharma AS	
				(NS)	Bicycle Therapeutics Ltd	
				(NS)	Cellectis SA	
				(NS)	Crispr Therapeutics AG	
				(NS)	Horizon Pharma PLC	
				(NS)	Novocure Ltd	
				(NS)	ProQR Therapeutics NV	
				(NS)	Prothena Corporation plc	
REST	(S)	Compugen (CGEN)	22	(S)	BeiGene Ltd (BGNE)	6.5
				(S)	Wave Life Sciences Ltd (WVE)	
	(NS)	Aurinia Pharmaceuticals Inc	14.5	(NS)	HUTCHMED DRC (HCM)	3.8
	(NS)	Xenon Pharmaceuticals Inc (XENE)		(NS)	I Mab (IMAB)	
				(NS)	Kiniksa Pharmaceuticals Ltd (KNSA)	
				(NS)	Repare Therapeutics Inc (RPTX)	
				(NS)	Zai Lab Ltd (ZLAB)	

(S) = Sustainable; (NS) = Non sustainable

*Due to the large number of companies in USA, it is not detailed individually.

Figure 4: Relationship between efficiency and age of listing

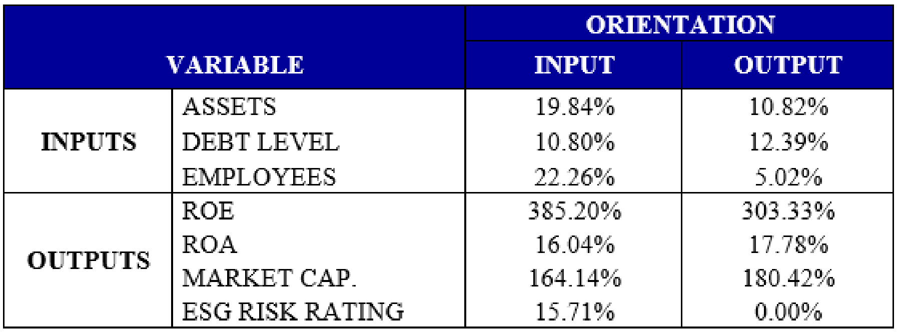
When carrying out DEA, 30% of the companies are efficient, as shown in figure 2, both in input and output orientation, being the same companies efficient in both orientations.

It is observed that under the output orientation the scores obtained are, on average, better than under input orientation. And USA companies obtained the higher results.

Of the efficient ones, 35% are sustainable (with an ESG risk index between low and medium) as figure 3 shows. No relationship is observed between sustainability and efficiency. In the USA, 10% of companies are both efficient and sustainable, while in Europe, that percentage increases to 18%, and 10% in the remaining region.

Regarding the age of listing, figure 4 shows that the efficient and sustainable companies are on average older than the non-efficient and non-sustainable ones.

Regarding the slacks, figure 5 shows that the inefficient companies must reduce their inputs or increase their outputs on average to reach the efficient frontier. It can be seen that the variables with the hightest slack in both orientations are the same: ROE and market capitalization.



VARIABLE		ORIENTATION	
		INPUT	OUTPUT
INPUTS	ASSETS	19.84%	10.82%
	DEBT LEVEL	10.80%	12.39%
	EMPLOYEES	22.26%	5.02%
OUTPUTS	ROE	385.20%	303.33%
	ROA	16.04%	17.78%
	MARKET CAP.	164.14%	180.42%
	ESG RISK RATING	15.71%	0.00%

Figure 5: Slacks

4 Conclusions

The companies included in the NBI are very different in terms of their size and economic-financial indicators. Many of them have a particular functioning given that their activity has been dedicated to the development of patents and pharmaceutical products, that is why they have a high volume of expenses, little income, and even negative net worth.

The factor analysis shows that the dimension that most differentiates the 373 companies is their size and explains 44% of their variability. In addition, factor 3 indicates that older companies are listed in a larger number of indices and obtain better sustainability results.

Out of the 260 companies finally included in the analysis, 90% of them belong to the USA and 25% are sustainable (65 companies). The efficiency evaluation shows that the same companies are located on the efficient frontier (30%), regardless of the orientation used. In the USA, 29.9% of companies are efficient, while in Europe 31.6% are efficient, and in the remaining region 30%. Inefficient companies must reduce the number of employees by an average of 22% (in input orientation) and increase their economic profitability by 385%. Furthermore, out of the 65 companies that are sustainable, 43% are efficient. Therefore, a relationship between sustainability and efficiency is not observed.

References

[1] Morrison, C., Lähteenmäki, R., Public biotech in 2016 - The numbers. *Nature Biotechnology*, 35(7): 623–9, 2017.

[2] Pisano, G.P., Can Science Be a Business? *Harvard Business Review*, 10: 1–12, 2006.

[3] Bianchi, M., Cavaliere, A., Chiaroni, D., Frattini, F., Chiesa, V., Organisational modes for Open Innovation in the bio-pharmaceutical industry: An exploratory analysis. *Technovation*, 31(1): 22–33, 2011.

[4] Boffo, R., Patalano, R., ESG Investing: Practices, Progress and Challenges. Paris, OCDE, 2020. www.oecd.org/finance/ESG-Investing- Practices-Progress-and-Challenges.pdf

[5] Zumente, I., Lāce, N., ESG rating—necessity for the investor or the company? *Sustainability*, 13(16):8940, 2021.

[6] Alcaide, M.Á., de la Poza, E., Guadalajara, N., Assessing the sustainability of high-value brands in the IT sector. *Sustainability*, 11(6):1598, 2019.

[7] Balseiro-Barrios, H.D., Luna-Amador, J.A., Maza-Ávila, F.J., Financial efficiency analysis of listed companies in the colombian stock market between 2012 and 2017. *Revista Finanzas y Politica Económica. Universidad Católica de Colombia*, Vol. 13: 19–41, 2021.

[8] Ji, Y.B., Lee, C., Data envelopment analysis.*The Stata Journal*, Vol. 10(2): 267–280, 2010.

22

An algorithm for solving Feedback Nash stochastic differential games with an application to the Psychology of love

Jorge Herrera de la Cruz[♮] and José-Manuel Rey[♭,1]

(♭) Department of Economic Analysis,
Complutense University of Madrid
28224 Pozuelo de Alarcón, Spain.
(♮) Department of Mathematics and Data Science,
CEU San Pablo
28003, Madrid.

1 Introduction

The purpose of this work is twofold. Firstly, we introduce a new algorithm to find feedback Nash equilibria of stochastic differential games and, secondly, we apply our methodology to a problem of significance in the social sciences, related to human behavior.

The numerical analysis of stochastic differential games (SDG) is currently a topic of growing interest (see [2–4, 12, 15, 16]). However, most contributions in the literature on differential games focus either on the study of the theoretical properties of certain classes of SDGs; or on using heuristic approximations to avoid solving the stochastic Hamilton-Jacobi-Bellman (HJB) equations of the problem. Our approach extends the idea in [9], where an algorithm- called RaBVItG (Radial Basis Value Iteration Game)- is introduced to solve the HJB system to find feedback Nash equilibria of deterministic differential games.

The second objective of this work is related to a research line followed by [1], [8], [17], [18] -among others- focused on mathematical models of love relationships. In this work, our goal is to evaluate the risk of rupture in a dyadic romantic relationship that is intended to last (see [10]). By analyzing the state variable (the *feeling*), being- in this model- a stochastic process, we can compute, as a key result, a measure of the ”probability of rupture”, at a certain moment of the relationship.

2 Methods

2.1 Mathematical Model

The model presented here is a stochastic two–person generalization of the optimal control model introduced in [17]. A deterministic version was also considered in [10]. The state variable at time

[1] j-man@ccee.ucm.es

$t \geq 0$ is described by $x(t)$ – called the *feeling* variable–, which is modeled by a stochastic process $\{x(t)\}_{t\geq 0}$, with $x : [0, \infty) \to X \subseteq \mathbb{R}^+$, being X the state space. The feeling evolves according to

$$\mathrm{d}x\,(t) = \left[-rx(t) + a_1 c_1\,(t) + a_2 c_2\,(t)\right]\mathrm{d}t + \sigma\,(x\,(t))\,\mathrm{d}w, \tag{1}$$

where $r, a_1, a_2 > 0$ and, for $i = 1, 2$, $c_i : [0, \infty) \to \mathbb{R}^+$ is a (piece-wise continuous) function that measures the effort put into the relationship by partner i at time t, and $w(t)$ is a Wiener process.

The total well-being W_i of each partner i is defined as

$$W_i\,(c_i) = \mathbb{E}\left(\int_0^\infty e^{-\rho_i t}\,(U_i\,(x\,(t)) - D_i\,(c_i\,(t)))\,\mathrm{d}t|x(0) = y\right),\ i = 1, 2, \tag{2}$$

where U_i and D_i are, respectively, the utility of feeling and disutility of effort and $\rho_i > 0$ is the individual rate of temporal preference. The underlying psychological rationale is explained in [17].

Given (1) and $x(0) = x_0$, the couple finds the effort trajectories $c_1^*(t), c_2^*(t)$ such that each individual's well-being integral (2) is maximal. Let us consider $c_i = S_i\,(x)$, being $S_i : X \to \mathbb{R}^+$ the *feedback* map that provides the effort by player i for the feeling x. We look for $\left(S_1^\heartsuit\,(\cdot), S_2^\heartsuit\,(\cdot)\right)$, such that $S_i^\heartsuit : X \to \mathbb{R}^+$ is a stationary feedback Nash equilibrium of the stochastic differential game. Indeed, this equilibrium is attained if $S_1^\heartsuit\,(x\,(t))$ solves

$$\max_{c_1(t)} \mathbb{E}\left(\int_0^\infty e^{-\rho_1 t}\,(U_1\,(x\,(t)) - D_1\,(c_1\,(t)))\,\mathrm{d}t|x(0) = y\right) \tag{3}$$

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

$$\max_{c_2(t)} \mathbb{E}\left(\int_0^\infty e^{-\rho_2 t}\,(U_2\,(x\,(t)) - D_2\,(c_2\,(t)))\,\mathrm{d}t|x(0) = y\right) \tag{4}$$

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 $y = x_0$, and $c_i(t) \in \mathbb{R}^+$ for $t \geq 0$.

Assuming the existence of a feedback Nash equilibrium $S^\heartsuit = \left(S_1^\heartsuit, S_2^\heartsuit\right)$ for the couple's problem, we define $v_i^\heartsuit : X \to \mathbb{R}$ as the *value function* of partner i,

$$v_i^\heartsuit\,(x_0) = W_i\left(S_i^\heartsuit\,(x\,(t))\right),\ i = 1, 2.$$

where $S_i^\heartsuit\,(x\,(t))$ is the optimal feedback control for partner i with initial state $x(0) = x_0$. The value functions $v_i^\heartsuit$ must satisfy the following stochastic (HJB) equations,

$$\begin{cases} \rho_1 v_1\,(x) = \max_{c_1 \in \mathbb{R}^+}\left\{U_1\,(x) - D_1\,(c_1) + v_1'\,(x)\left(-rx + a_1 c_1 + a_2 S_2^\heartsuit\,(x)\right) + \frac{1}{2} v_1''\,(x)\,\sigma^2\,(x)\right\}, \\ \rho_2 v_2\,(x) = \max_{c_2 \in \mathbb{R}^+}\left\{U_2\,(x) - D_2\,(c_2) + v_2'\,(x)\left(-rx + a_1 S_1^\heartsuit\,(x) + a_2 c_2\right) + \frac{1}{2} v_2''\,(x)\,\sigma^2\,(x)\right\}. \end{cases} \tag{5}$$

And the solution of (5) gives the stochastic feedback maps as

$$\begin{cases} S_1^\heartsuit\,(x) = arg\max_{c_1 \in \mathbb{R}^+}\left\{U_1\,(x) - D_1\,(c_1) + v_1'\,(x)\left(-rx + a_1 c_1 + a_2 S_2^\heartsuit\,(x)\right) + \frac{1}{2} v_1''\,(x)\,\sigma^2\,(x)\right\}, \\ S_2^\heartsuit\,(x) = arg\max_{c_2 \in \mathbb{R}^+}\left\{U_2\,(x) - D_2\,(c_2) + v_2'\,(x)\left(-rx + a_1 S_1^\heartsuit\,(x) + a_2 c_2\right) + \frac{1}{2} v_2''\,(x)\,\sigma^2\,(x)\right\}, \end{cases} \tag{6}$$

obtaining the feedback Nash stochastic equilibirum of the problem. By inserting $S_i^\heartsuit(x(t))$, $i = 1, 2$, into (1), we obtain

$$\mathrm{d}x\,(t) = \left[-rx(t) + a_1 S_1^\heartsuit\,(x\,(t)) + a_2 S_2^\heartsuit\,(x\,(t))\right]\mathrm{d}t + \sigma\,(x\,(t))\,\mathrm{d}w,$$

with $\left\{x^\heartsuit(t)\right\}_{t\geq 0}$, the optimal stochastic process of *feeling* for the couple with x_0.

2.2 Discretization and Pseudocode

Although we present here a discretization and computational implementation of the model introduced above, it can be easily extended to N- players with more than one control and state variables per player. The presented model is discretized in a Semi-Lagrangian way (see, for instance, [5]). Thus, the discrete version of (2) is

$$W_i^h\left(c_i^h\right)=\mathbb{E}\left\{h\sum_{k=0}^{\infty}e^{-\rho_i k}\left(U_i\left(x_k\right)-D_i\left(c_{i,k}\right)\right)|x_0=y\right\},\ i=1.2, \tag{7}$$

where $c_i^h=\{c_{i,k}\}_{k\geq 0}$ is a sequence of (feasible) controls for partner i, defined by the piece-wise constant function $c_i^h\left(\tau\right)=c_{i,k},\ \tau\in[t_k,t_{k+1})$, where $t_k=hk, k\in\mathbb{N}\cup\{0\}$. The sequence $x_k=x(t_k)$ is obtained by discretization of (1) using the Euler-Maruyama scheme [11],

$$x_{k+1}=x_k+hf\left(x_k,c_{1,k},c_{2,k}\right)+\sigma\left(x_k\right)\xi_k, \tag{8}$$

with $f(x,c_1,c_2)=-rx+a_1c_1+a_2c_2$, $x_0=y$, and ξ_k as the increment of a standard Brownian motion $w(t)$ in the interval $[t_k,t_{k+1})$. The discrete value function for partner $i=1,2$ is given by

$$v_i^h\left(y\right)=\max_{c_i^h}W_i^h\left(c_i^h\right).$$

According to [14], the Gaussian variable ξ_k, can be replaced by a discrete variable with probability distribution $\mathbb{P}\left(\xi_k=\sqrt{h}\right)=\mathbb{P}\left(\xi_k=-\sqrt{h}\right)=\frac{1}{2}$. Therefore, we can redefine (8) as a set of two displacements,

$$x_{k+1}=x_k+\delta_d,\ d=1,2,$$

where, for the sake of simplicity, we will write $\delta_d=hf\left(x_k,c_{1,k},c_{2,k}\right)+\sigma\left(x_k\right)\left(\pm\sqrt{h}\right)$.

Then, we arrive to the Dynamic Programming Principle in discrete time

$$\begin{cases}v_1^h\left(y\right)=\max_{c_1\in\mathbb{R}^+}\left\{h\left(U_1\left(y\right)-D_1\left(c_1\right)\right)+\frac{(1-\rho_1 h)}{2}\sum_{d=1}^{2}v_1^h\left(y+\delta_d\left(y,c_1,S_2^h\left(y\right)\right)\right)\right\},\\ v_2^h\left(y\right)=\max_{c_2\in\mathbb{R}^+}\left\{h\left(U_2\left(y\right)-D_2\left(c_2\right)\right)+\frac{(1-\rho_2 h)}{2}\sum_{d=1}^{2}v_2^h\left(y+\delta_d\left(y,S_1^h\left(y\right),c_2\right)\right)\right\},\end{cases} \tag{9}$$

together with the corresponding discrete version of (6), namely

$$\begin{cases}S_1^h\left(y\right)=arg\max_{c_1\in\mathbb{R}^+}\left\{h\left(U_1\left(y\right)-D_1\left(c_1\right)\right)+\frac{(1-\rho_1 h)}{2}\sum_{d=1}^{2}v_1^h\left(y+\delta_d\left(y,c_1,S_2^h\left(y\right)\right)\right)\right\},\\ S_2^h\left(y\right)=arg\max_{c_2\in\mathbb{R}^+}\left\{h\left(U_2\left(y\right)-D_2\left(c_2\right)\right)+\frac{(1-\rho_2 h)}{2}\sum_{d=1}^{2}v_2^h\left(y+\delta_d\left(y,S_1^h\left(y\right),c_2\right)\right)\right\}.\end{cases} \tag{10}$$

We consider an approximation of the functions v_i^h, satisfying (9), by a spatial discretization of the state space. Defining $\tilde{X}=\{y_j\}_{j=1,\ldots,Q}\subset X$ as a set of arbitrary Q points, in general, the points $y_i^\sharp=y_j+\delta_d\left(y_j,c_1,c_2\right)$ in (9) do not belong to $\tilde{X}$. To find approximate values $\tilde{v}_i^h\left(y_j\right)$ of $v_i^h\left(y_j\right)$ for $y_j\in\tilde{X}, i=1,2$, the values $v_i^h\left(y^\sharp\right)$ in (9) are calculated through a collocation mesh-free algorithm using the set of scattered nodes $\tilde{X}$, by means of

$$\begin{cases}\tilde{v}_1^h\left(y_j\right)=\max_{c_1\in\mathbb{R}^+}\left\{h\left(U_1\left(y_j\right)-D_1\left(c_1\right)\right)+(1-\rho_1 h)\overline{RBF}\left[V_1\right]\left(y_1^{\#}\right)\right\},\\ \tilde{v}_2^h\left(y_j\right)=\max_{c_2\in\mathbb{R}^+}\left\{h\left(U_2\left(y_j\right)-D_2\left(c_2\right)\right)+(1-\rho_2 h)\overline{RBF}\left[V_2\right]\left(y_2^{\#}\right)\right\},\end{cases} \tag{11}$$

with discrete feedback strategies

$$\begin{cases}\tilde{S}_1^h\left(y_j\right)=arg\max_{c_1\in\mathbb{R}^+}\left\{h\left(U_1\left(y_j\right)-D_1\left(c_1\right)\right)+(1-\rho_1 h)\overline{RBF}\left[V_1\right]\left(y_1^{\#}\right)\right\},\\ \tilde{S}_2^h\left(y_j\right)=arg\max_{c_2\in\mathbb{R}^+}\left\{h\left(U_2\left(y_j\right)-D_2\left(c_2\right)\right)+(1-\rho_1 h)\overline{RBF}\left[V_1\right]\left(y_2^{\#}\right)\right\}.\end{cases} \tag{12}$$

where

$$\begin{cases} y_1^\sharp = y_j + hf\left(y_j, c_1, \tilde{S}_2^h\left(y_j\right)\right) + \sigma\left(y_j\right)\left(\pm\sqrt{h}\right), \\ y_2^\sharp = y_j + hf\left(y_j, \tilde{S}_1^h\left(y_j\right), c_2\right) + \sigma\left(y_j\right)\left(\pm\sqrt{h}\right), \end{cases}$$

and $\overline{RBF}[V_i]$, $i = 1,2$, denoting the average of the i-th value function's approximation by radial basis functions [6]. Specifically, for $y^\sharp$ that does not belong to $\tilde{X}$, we have, for $i = 1,2$,

$$\tilde{v}_i^h\left(y_{i,d}^\#\right) \approx \overline{RBF}\left[\mathrm{V}_i\right] = \frac{1}{2}\sum_{d=1}^{2}\sum_{j=1}^{Q}\lambda_{i,j}\Phi\left(\left\|y_{i,d}^\# - y_j\right\|\right),$$

where $y_{i,d}^\sharp = y_j + hf\left(y_j, [c_1, c_2]\right) \pm \sigma\left(y_j\right)\sqrt{h}$, and $\lambda_{i,j} \in \mathbb{R}$ are weighting coefficients, with $\Phi\left(\|y - y_j\|\right) = \exp\left(-\frac{\|y-y_j\|^2}{\sigma^2}\right)$, and $\sigma > 0$ (see [7]). In addition, for $i = 1,2$, and $j = 1,...,Q$, the parameters $\lambda_{i,j}$, are obtained by solving

$$\mathrm{A}\bar{\lambda}_i = \mathrm{V}_i,$$

where A is the matrix with entries $\mathrm{A}_{j,l} = \Phi\left(\|y_l - y_j\|\right)$ for $j = 1,...,Q$, and $\bar{\lambda}_i = \left[\lambda_{i,1},....,\lambda_{i,Q}\right]^T$.

The algorithm to produce a solution of the discretized problem is called RaBVItG, which refers to Radial Basis approximations, Value Iteration and Game Iteration. It essentially consists of two main loops: game iteration, to find a Nash Equilibrium for a given value function, and value iteration, to improve the approximation of the value function, given a previously obtained equilibrium. Let us define $\mathrm{V} = [\mathrm{V}_1, \mathrm{V}_2]$, $\mathrm{C} = [\mathrm{C}_1, \mathrm{C}_2]$, denoting the arrays storing the information for both partners, with $\mathrm{V}_i = \left[\tilde{v}_i^h\left(y_1\right),...,\tilde{v}_i^h\left(y_Q\right)\right]^T$, $\mathrm{C}_i = \left[\tilde{c}_i^h\left(y_1\right),...,\tilde{c}_i^h\left(y_Q\right)\right]^T$, $i = 1,2.$, evaluated at the points $y_j \in \tilde{X}$.

Let $T_i = \left[T_{i,1},...,T_{i,Q}\right] : \mathbb{R}^Q \to \mathbb{R}^Q$ and $G_i = \left[G_{i,1},...,G_{i,Q}\right] : \mathbb{R}^Q \to \mathbb{R}^Q$ be two operators defined component-wise by

$$T_{i,j}\left(\mathrm{V}_i\right) = h\left(U_i\left(y_j\right) - D_i\left(c_i\right)\right) + \left(1 - \rho_i h\right)\overline{RBF}\left[\mathrm{V}_i\right]\left(y_j + \delta_d\right),\ j = 1,...,Q,$$

and

$$G_{i,j}\left(\mathrm{V}_i\right) = arg\max_{c_i \in \mathbb{R}^+}\left\{h\left(U_i\left(y_j\right) - D_i\left(c_i\right)\right) + \left(1 - \rho_i h\right)\overline{RBF}\left[\mathrm{V}_i\right]\left(y_j + \delta_d\right)\right\},\ j = 1,...,Q, \tag{13}$$

with δ_d, $d = 1,2$, the set of displacements $\delta_d = hf\left(y_j, [c_1, c_2]\right) \pm \sigma\left(y_j\right)\sqrt{h}$. Then, we define

1. *Game Iteration.* For $s = 0,1,,...,$ generate C_i^{s+1} at step $s+1$ for partner i, as follows:

   $$\mathrm{C}_i^{s+1} = \theta\mathrm{C}_i^s + (1-\theta)\,G_i\left(\mathrm{C}^s, \mathrm{V}_i^r\right),\ i = 1,2,$$

   where G_i is defined in (13), $\theta \in (0,1)$ is a weighting coefficient –see [13], and V_i^r is defined below. The Game Iteration loop follows the scheme

   $$\begin{cases} \tilde{c}_{1,j}^{s+1} \equiv \theta\tilde{c}_{1,j}^s + (1-\theta)\,arg\max_{c_1 \in \mathbb{R}^+}\left\{h\left(U_1\left(y_j\right) - D_1\left(c_1\right)\right) + \left(1 - \rho_1 h\right)\overline{RBF}\left[\mathrm{V}_1^r\right]\left(y_1^\sharp\right)\right\}, \\ \tilde{c}_{2,j}^{s+1} \equiv \theta\tilde{c}_{2,j}^s + (1-\theta)\,arg\max_{c_2 \in \mathbb{R}^+}\left\{h\left(U_2\left(y_j\right) - D_2\left(c_2\right)\right) + \left(1 - \rho_2 h\right)\overline{RBF}\left[\mathrm{V}_2^r\right]\left(y_2^\sharp\right)\right\}, \\ y_1^\sharp = y_j + hf\left(y_j, \left[c_1, \tilde{c}_{2,j}^s\right]\right) \pm \sigma\left(y_j\right)\sqrt{h}, \\ y_2^\sharp = y_j + hf\left(y_j, \left[\tilde{c}_{1,j}^s, c_2\right]\right) \pm \sigma\left(y_j\right)\sqrt{h}, \end{cases}$$

   where $\tilde{c}_{i,j}^s \equiv \tilde{c}_{i,j}^s\left(y_j\right)$. Is iterated until a convergence criterion is satisfied, that is, $\left\|\mathrm{C}^{s+1} - \mathrm{C}^s\right\| < \epsilon_1$, for a given $\epsilon_1 > 0$ ($\|\cdot\|$ is the Euclidean norm). Given V_i^r, $i = 1,2$, a candidate for feedback Nash equilibrium is thus obtained: $\mathrm{C}^{s+1} = \left[\mathrm{C}_1^{s+1}, \mathrm{C}_2^{s+1}\right]$.

   This is the input for the next loop, to produce an estimate of the functions V_i^{r+1}, $i = 1,2$.

2. *Value Iteration.* Given C^{s+1}, obtained from the previous loop, the value functions at step $r+1$ are:
$$\mathrm{V}_i^{r+1} = T_i\left(\mathrm{V}_i^r;\mathrm{C}^{s+1}\right),\ i=1,2,$$
where, for $j=1,...,Q$,
$$\begin{cases} T_{1,j} \equiv h\left(U_1\left(y_j\right) - D_1\left(\tilde{c}_{1,j}^{s+1}\right)\right) + \left(1-\rho_1 h\right)\overline{RBF}\left[\mathrm{V}_1^r\right]\left(y_1^\sharp\right), \\ T_{2,j} \equiv h\left(U_2\left(y_j\right) - D_2\left(\tilde{c}_{2,j}^{s+1}\right)\right) + \left(1-\rho_2 h\right)\overline{RBF}\left[\mathrm{V}_2^r\right]\left(y_2^\sharp\right), \\ y_1^\sharp \equiv y_2^\sharp = y_j + hf\left(y_j, \left[\tilde{c}_{1,j}^{s+1}, \tilde{c}_{2,j}^{s+1}\right]\right) \pm \sigma\left(y_j\right)\sqrt{h}. \end{cases}$$
Is iterated until satisfying the convergence criterion $\left\|\mathrm{V}^{r+1} - \mathrm{V}^r\right\| < \epsilon_2$, with $\epsilon_2 > 0$ given. A candidate solution for the value functions are thus obtained: $\mathrm{V}^{r+1} = \left[\mathrm{V}_1^{r+1}, \mathrm{V}_2^{r+1}\right]$.

Once the convergence conditions are met, the algorithm generates the outputs $\mathrm{V}^\heartsuit = \left[\mathrm{V}_1^\heartsuit, \mathrm{V}_2^\heartsuit\right], \mathrm{C}^\heartsuit = \left[\mathrm{C}_1^\heartsuit, \mathrm{C}_2^\heartsuit\right]$ as the computational solutions for the value functions and control policies of the couple's problem. Notice that $\mathrm{V}^\heartsuit$ and $\mathrm{C}^\heartsuit$ is an approximate fixed-point of the numerical scheme
$$\begin{cases} \mathrm{C}^{s+1} = G(\mathrm{C}^s, \mathrm{V}^r, \Delta), \\ \mathrm{V}^{r+1} = T(\mathrm{V}^r, \mathrm{C}^{s+1}, \Delta), \end{cases}$$
where Δ denotes the chosen set of spatio-temporal discretization parameters. Once $\left(\mathrm{C}^\heartsuit, \mathrm{V}^\heartsuit\right)$ are obtained, we can recover the corresponding approximated feedback maps defined in (12). For the purpose of our model analysis below, we take $f\left(y, [c_1, c_2]\right) = -ry + a_1c_1 + a_2c_2$, and $\sigma\left(y_j\right) \equiv \sigma$ constant.

3 Results

Our analysis is based on some parameterization of the model from the literature (see [8] and [10]). We present a synthesized approach of how our algorithm can compute the probability of breakup for a couple, addressing the problem of estimating the change in the probability of breakup after a shock of the feeling occurring at time $k > 0$. Even though the stabilization mechanism obtained by the algorithm is working, the perturbed feeling trajectory may enter the zone of risk of rupture at a certain moment k. This "alarm zone" is called "Love at Risk (LaR)" here, and it is defined as follows:
$$\mathbb{P}\left(x_k^\heartsuit \leq x_{min}\right) = \alpha$$
where $x_k^\heartsuit$ is the (optimal) solution of the computational couple's problem previously defined, $\mathbb{P}$ is its probability function, so that x_{min} is the α-percentile of the distribution of $x_k^\heartsuit$.

To illustrate the probability estimation described above, consider the case where s_- consists of a large one-period shock (of size σ) taking place at time $k = 60$ (five years after the wedding). In Figure 1 (left) we show the percentile trajectories of the stochastic process steered by the approximated stabilization mechanism for a particular couple, and for different values of σ. Computing a large ensemble of trajectories, we produce an estimate of the distribution of the feeling values for the perturbed process over a whole year ($k = 60, ..., 72$) after the shock at $k = 60$. In Figure 1 (right) we show the empirical distributions of the feeling variable before the shock and over one year after the shock. As shown in Figure 1 (right), given that the shock at $k = 60$ has the size of the uncertainty σ, the probability of breakup over the year after the event increases as σ increases.

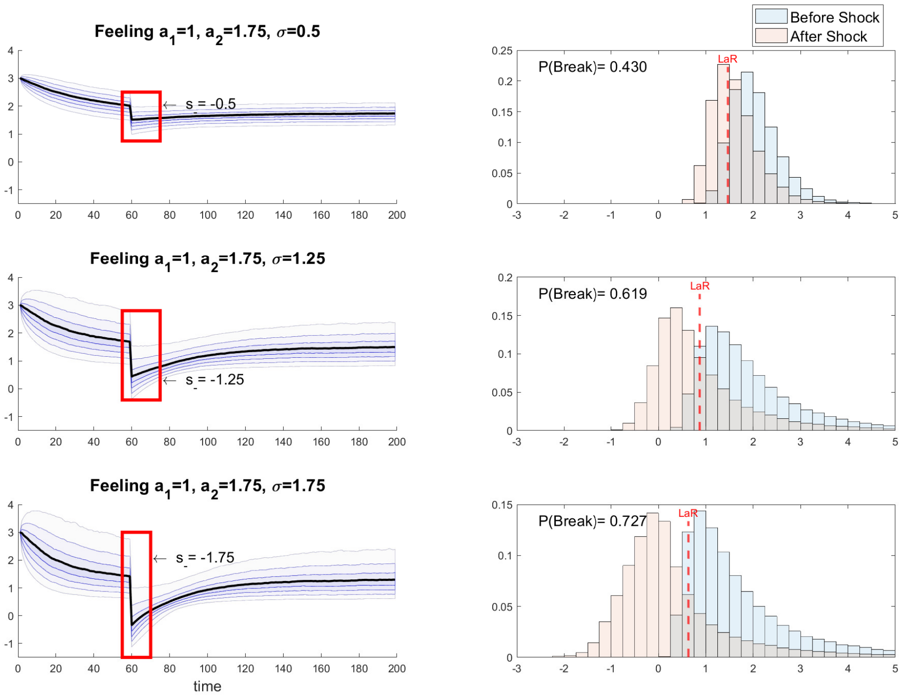


Figure 1: Left: Feedback response to a one-period shock of size s_-- proportional to σ five years after the wedding ($k = 60$) for a heterogamous couple with $a_1 = 1, a_2 = 1.75$, and for $\sigma = 0.5, 1.25, 1.75$. Feeling trajectories are obtained through the feedback-map. Right: Empirical distribution for the feeling values obtained from an sample of 10000 trajectories before and after the shock for the different values of σ. LaR values correspond to the unperturbed process.

We also analyze how the probability of breakup after a shock varies with respect to the size of the shock and the uncertainty of the feeling dynamics. For the same type of couple considered above, we also obtain that the probabilities of breakup for different values of σ- and different sizes of a one-period shock occurring five years after the wedding- for any level σ, the probability of breakup increases as the size of the shock increases. Also, a higher level of uncertainty entails a lower LaR level and, in addition, it makes more likely that the level of feeling remains in the secure zone (i.e. over x_{min}) for the relationship.

4 Conclusions

In this article we have introduced an algorithm to find feedback Nash equilibria for a class of stochastic differential games. The algorithm builds on a combination of two fixed point iterations, a first one to find the Nash equilibrium by fixing the value of the game, and a second iteration to find the value of the game given a Nash equilibrium. The algorithm can be applied to a general class of N-player infinite horizon stochastic games. We have also considered a substantial issue in the applied sciences, namely the design of a long-term rewarding romantic relationship. We formulate this problem as a two-person optimal control problem to steer the feeling of the relationship in a stochastic environment. The algorithm allows us to find approximate solutions of a computational version of the control problem for different stochastic dynamics. In particular, we have focused on estimating the risk of breakup of a long-term relationship at a certain time after the initial commitment. The computational model allows us to estimate the probability of breaking up in the face of an external shock. The analysis can be applied to different types of couples and different levels of stochasticity in the feeling dynamics.

References

[1] Dario Bauso, Dia Ben Mansour, Boualem Djehiche, Hamidou Tembine, and Raul Tempone. Mean-field games for marriage. PloS one, 9(5):e94933, 2014.

[2] Alain Bensoussan, Chi Chung Siu, Sheung Chi Phillip Yam, and Hailiang Yang. A class of non-zero-sum stochastic differential investment and reinsurance games. *Automatica*, 50(8):2025–2037, 2014.

[3] Chao Deng, Xudong Zeng, and Huiming Zhu. Non-zero-sum stochastic differential reinsurance and investment games with default risk. *European Journal of Operational Research*, 264(3):1144–1158, 2018.

[4] Jacob Engwerda. LQ dynamic optimization and differential games. John Wiley & Sons, 2005.

[5] Maurizio Falcone. Numerical methods for differential games based on partial differential equations. *International Game Theory Review*, 8(02):231–272, 2006.

[6] Gregory E. Fasshauer. Meshfree approximation methods with MATLAB, volume 6. *World Scientific*, 2007.

[7] Gregory E. Fasshauer and Jack G Zhang. On choosing "optimal" shape parameters for rbf approximation. *Numerical Algorithms*, 45(1-4):345–368, 2007.

[8] Thierry Goudon and Pauline Lafitte. The lovebirds problem: why solve hamilton-jacobi-bellman equations matters in love affairs. *Acta Applicandae Mathematicae*, 136(1):147–165, 2015.

[9] Jorge Herrera, Benjamin Ivorra, and Ángel M Ramos. An algorithm for solving a class of multiplayer feedback-nash differential games. *Mathematical Problems in Engineering 2019*, 2019.

[10] Jorge Herrera and José-Manuel Rey. Controlling forever love. *PloS one*, 16(12):e0260529, 2021.

[11] Desmond J Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, 43(3):525–546, 2001.

[12] Ricardo Josa-Fombellida and Juan Pablo Rincón-Zapatero. New approach to stochastic optimal control. *Journal of Optimization Theory and Applications*, 135(1):163–177, 2007.

[13] Jacek B Krawczyk and Stanislav Uryasev. Relaxation algorithms to find nash equilibria with economic applications. *Environmental Modeling & Assessment*, 5(1):63–73, 2000.

[14] Harold Kushner and Paul G Dupuis. Numerical methods for stochastic control problems in continuous time, volume 24. *Springer Science & Business Media*, 2013.

[15] Paola Mannucci. Nonzero-sum stochastic differential games with discontinuous feedback. *SIAM journal on control and optimization*, 43(4):1222–1233, 2004.

[16] Jesús Marín-Solano and Ekaterina V. Shevkoplyas. Non-constant discounting and differential games with random time horizon. *Automatica*, 47(12):2626–2638, 2011.

[17] José-Manuel Rey. A mathematical model of sentimental dynamics accounting for marital dissolution. *PloS one*, 5(3):e9881, 2010.

[18] Sergio Rinaldi, Fabio Della Rossa, Fabio Dercole, Alessandra Gragnani, and Pietro Landi. *Modeling love dynamics*, volume 89. World Scientific, 2015.

23

Detection of border communities using convolution techniques

José Miguel Montañana[♭,1], Antonio Hervás[♮], Samuel Morillas[†] and Alejandro Méndez[§]

(♭) inpeek GmbH, Darmstadt, Germany,
(♮) IUMM. Universitat Politècnica de València, València, Spain,
(†) IUMPA. Universitat Politècnica de València, València, Spain,
(§) ETSINF. Universitat Politècnica de València, València, Spain,

1 Introduction

Graph theory has become an essential element in many scientific areas, as its ability to model many types of relationships and processes has made it a powerful tool for the resolution of problems in areas such as physics, biology, or computer science. The study of graphs modelling real-world systems has shown that in these systems there exists a high level of order and structure in their vertices. This characteristic is what is called community structure. A community consists of a group of vertices with similar properties and functions in its graph. Numerous algorithms have been proposed for the detection of communities in the last few decades [1]. Nevertheless, none of the algorithms presented to date have shown good results in every kind of graph, due to the great variety of them that can be found in real systems. Classical methods, such as Girvan-Newman [2], Walktrap [3], or Fast Greedy [4], have shown good results in strongly connected and well-conditioned graphs; but little to none work has been done to develop good performing algorithms for detecting communities in directed weakly connected bad-conditioned graphs.

The objective of this work is to propose a method for the detection of communities in this specific complicated type of graph. In these, we can find edges whose weight is much lower compared to other edges in the graph, which can condition the resulting communities. These edges act like noise in the system, and can be eliminated under certain conditions, and the goal of this work is to find a method to eliminate that noise under adequate conditions.

Keywords: Graphs and networks applications. Complex networks. Noise filtering.

2 Methods

2.1 Edge detection in images

Edge detection is commonly used as one of the first steps when retrieving information from an image, as it provides data about the structure and properties of the objects in the scene. Edge detection techniques are commonly based on the use of convolution, a mathematical process that consists in adding its neighbors to each element of a matrix, weighted by a kernel (or filter). The

[1] jmmontanana@gmail.com;ahervas@imm.upv.es;almencar@inf.upv.es

election of this kernel will determine the result obtained and its use. An example of convolution over a image is shown in Figure 1.

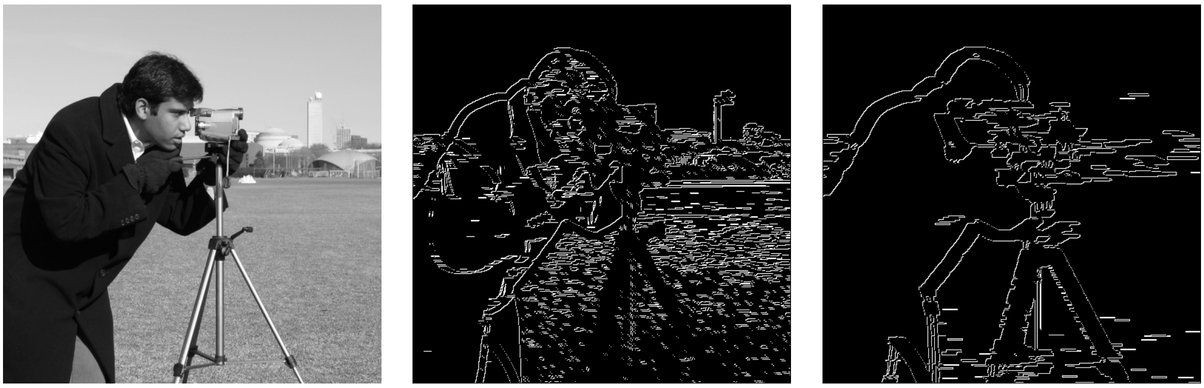
$$\begin{bmatrix} 3 & 1 & 4 & 1 & 5 & 9 \\ 2 & 6 & 5 & 3 & 5 & 8 \\ 9 & 7 & 9 & 3 & 2 & 3 \\ 8 & 4 & 6 & 2 & 6 & 4 \\ 3 & 3 & 8 & 3 & 2 & 7 \\ 9 & 5 & 0 & 2 & 8 & 8 \end{bmatrix} * \begin{bmatrix} -1 & 2 & -1 \\ 2 & -4 & 2 \\ -1 & 2 & -1 \end{bmatrix} = \begin{bmatrix} -19 & 12 & -4 & 0 \\ 7 & -9 & -2 & 9 \\ 3 & 6 & 7 & -20 \\ 5 & -21 & -4 & 24 \end{bmatrix}$$

Figure 1: Convolution using a 3 × 3 filter.

This work will be focused on the Laplacian of Gaussian filter (or *LoG*), an edge detection filter that combines both the Gaussian smoothing to reduce noise in the image with the second derivative computation with the Laplacian. This filter is defined by the following equation:

$$\nabla^2 g(x,y) = \Big(\frac{x^2+y^2-2\sigma^2}{\sigma^4}\Big) e^{\frac{-x^2+y^2}{2\sigma^2}} \tag{1}$$

where σ corresponds to a regularisation parameter that defines the amount of noise removed, as well as the potential loss of information. An example of edge detection in images can be seen in Figure 2, where the Laplacian of Gaussian is applied over the image on the left. The images in the figure are obtained using Python 3.10, the original one comes from the *scikit-image* package, and the results are obtained with the *gaussian_laplace* function from the *scipy* package. It can be appreciated how the regularisation parameter affects the final result: a lower value of σ does not remove all the noise in the image, while a larger value removes some information (this can be appreciated in the buildings in the background).



(a) Original image

(b) LoG with $\sigma = 2$

(c) LoG with $\sigma = 4$

Figure 2: Example of edge detection using Laplacian of Gaussian.

2.2 Convolution-based community detection algorithm

Recently, some algorithms have been proposed that use the convolution product to detect communities in graphs, in a similar fashion to what is done in the detection of edges in images. This technique was used previously with a community detection algorithm in [5]. In this algorithm, the convolution product is applied over the line graph of the graph of interest in order to detect borders between communities. A new algorithm based on convolution was proposed by the group in [6], where the convolution product is used to detect and prune edges acting as noise in the graph.

The method presented in this work is a variation from the previous method, and it was developed in order to improve the results obtained with the previously commented community detection methods. The main differences are found in the method for computing the kernel of weights and the edges involved in the convolution. A convolution distance, n, is defined, and the convolution product is obtained taking into account all the edges of the graph at a maximum distance n from and to the edge to convolve. Figure 3 shows a diagram with the edges involved in the convolution of a given edge e, and a convolution distance $n = 2$. The nomenclature of the edges follows the scheme $w_{destination}$, distance, id.

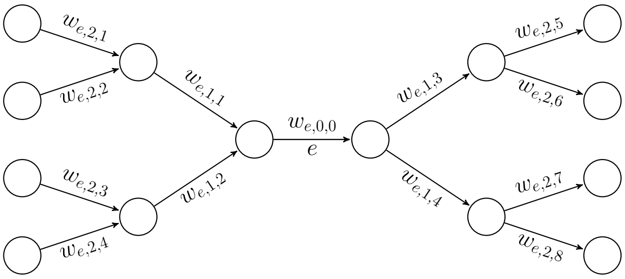


Figure 3: Edges involved in the convolution of edge e given a convolution distance $n = 2$

The convolution product of an edge e is obtained from the next equation:

$$k_a(n) = k_0 * w_{e,0,0} + \sum_{i=1}^{n} \sum_{j=1}^{m} k_i * w_{e,i,j} \tag{2}$$

where $k = [k_0, k_1, k_2, \ldots, k_n]$ is a kernel of weights computed using the Laplacian of Gaussian as described in the next equation:

$$k_i = \nabla^2 g(i, 0) \quad ; \quad i = 1, \ldots, n \tag{3}$$

$$k_0 = \sum_{i=1}^{n} q_i k_i \tag{4}$$

where q_i expresses the number of edges at a distance i from or to the edge to convolve. The kernel k_0 is computed in such a way that the sum of all kernels is equal to 0, trying to mimic the kernels used in image processing. This way, in areas where the edges are uniform, the gradient would be close to 0.

The community detection algorithms are applied to the graph after the edges that are potentially impeding the correct detection of communities are pruned. The edges to be removed are those for which the corresponding convolution product is lower than 0. In particular, those pruned edges have also a significantly lower weight than their neighbor edges.

3 Results

The efficiency of this algorithm was tested over the graph that models the access procedure to the public university system in Spain, with data from the Valencian Community in the year 2014. This graph with 215 vertices and 9615 edges is a good example of a directed badly conditioned weakly connected graph. A graphic representation can be seen in Figure 4.

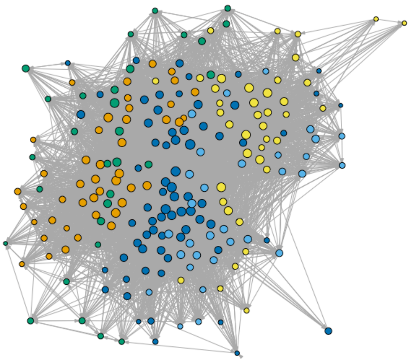


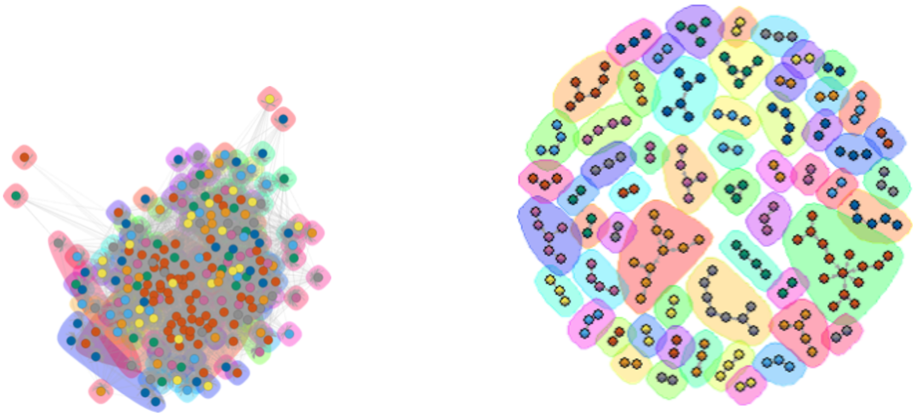
Figure 4: Spanish public university access system graph

Preliminary community detection results were obtained using classical algorithms, such as Walktrap or Girvan-Newman, but they were not good: either the modularity was too low, or the communities obtained did not provide much information. In an effort to improve the results obtained with the classical community detection algorithms, the method based on convolution presented in [6] was applied to the graph, obtaining the results shown in Figure 5. Two different convolution filters were used: the Laplacian of Gaussian and the Log kernels. As can be observed, very different results were obtained. The Log convolution filter (Figure 5.b) obtains particularly good modularity, 0.94, but at the cost of removing almost all edges, and thus losing a lot of valuable information. On the contrary, the Laplacian of Gaussian convolution (Figure 5.a) prunes a much lower number of edges, but the results obtained when detecting communities are not good (this can be seen in the low modularity obtained). These results motivate the design and implementation of improvements over the algorithms presented.

In order to improve the community detection results obtained so far, the convolution method presented in this work was applied over the graph. Using $\sigma = 0.5$ to compute the Laplacian of Gaussian, the convolution product was applied over all edges of the graph, and the edges with a negative convolution product were pruned from the graph. This pruned graph is shown in Figure 6.a, where 7924 edges were removed, the 82% of the total number of edges. The Walktrap community detection algorithm was then applied to the pruned graph, obtaining the communities shown in Figure 6.b. The modularity obtained is 0.64, which is an acceptable result, and the communities can also give useful information about the graph, and, therefore, about the entire system. For example, isolated communities can be found, that correspond to the less solicited university degrees; as well as medium size communities, that contain degrees in very specific areas, such as philology or forestry engineering. The biggest communities correspond to broader areas of knowledge: social sciences, natural and health sciences, and engineering. These results outperform greatly the results obtained with previous algorithms.

Method	**Kernel**	**Pruned edges**	**Modularity**
Previous Method [6]	Laplacian of Gaussian	332 (3.45%)	0.0207
Previous Method [6]	Log kernel	9350 (97.24%)	0.97
New Method	**Laplacian of Gaussian**	**7924 (84%)**	**0.64**

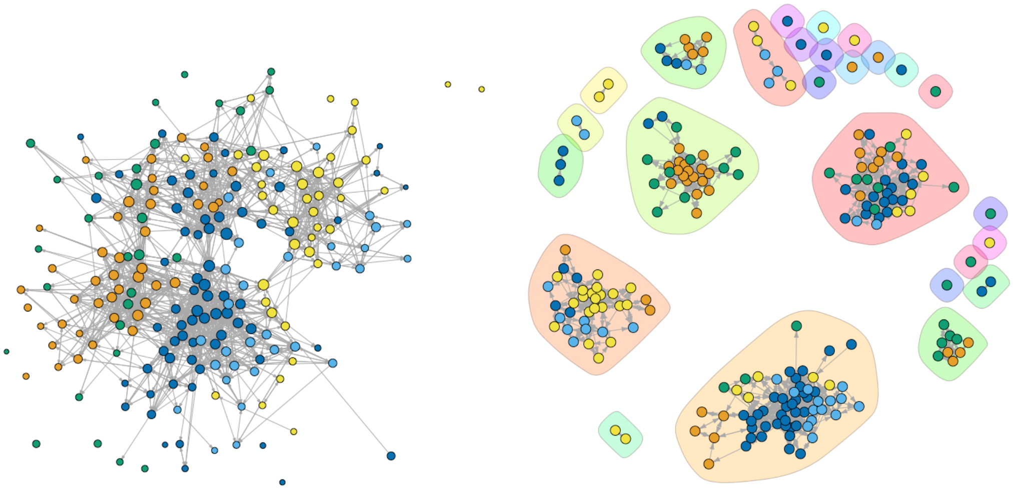
Table 1: Comparison of results



(a) Communities detected with Laplacian of Gaussian

(b) Communities detected with Log kernel Gaussian

Figure 5: Community detection results with the method described in [6]



(a) Pruned graph

(b) Communities detected with Walktrap

Figure 6: Community detection results with the new method

4 Conclusions

The results obtained with this new community detection algorithm show that the methods based on convolution techniques can be a satisfactory solution for the detection of communities in badly conditioned weakly-connected graphs. Future work can be continued in various ways. On the one hand, a method for the optimisation of the regularisation parameter of the algorithm presented in this work must be developed in order to find the optimal ratio between noise reduction and loss of information. On the other hand, other convolution filters can be used that may be more suitable for the pruning phase of the graph.

References

[1] Lancichinetti A and Fortunato S, Community detection algorithms: A comparative analysis, Phys. Rev. E 80(5), doi: 10.1103/PhysRevE.80.056117, 2009.

[2] Girvan, M. and Newman, M. E., Community structure in social and biological networks, Proc. of the National Academy of Sciences-PNAS, 99(12), pages 7821-7826, doi:10.1073/pnas.12265379, 2002.

[3] Pons, P. and Matthieu, L., Computing communities in large networks using random walks, Computer and Information Sciences - ISCIS, 284-293, doi: 10.1007/11569596_31, 2005.

[4] Clauset, A., Newman, M.E.J., and Moore, C., Finding community structure in very large networks, Physical Review E, 70(6), doi: 10.1103/PhysRevE.70.066111, 2004.

[5] Muñoz, H., Vicente, E., Gonzalez, I., Mateos A., and Jimenez-Martin, A., ConvGraph: Community Detection of Homogeneous Relationships in Weighted Graphs, Mathematics, 9(4), doi:10.3390/math9040367, 2021.

[6] Hervás, A., Montañana, J. M., Morillas, S., and Soriano, P.P., A procedure for detection of border communities using convolution techniques, Mathematical Modelling for Engineering & Human Behaviour, 2021.

24

Optimizing rehabilitation alternatives for large intermittent water distribution systems

Bruno Brentan[♭], Silvia Carpitella[♮], Ariele Zanfei[♯], Rui Gabriel Souza[♭], Andrea Menapace[♯],
Gustavo Meirelles[♭] and Joaquín Izquierdo[±]

(♭) Hydraulic and Water Resources Department, Federal University of Minas Gerais, Av.Presidente Antônio Carlos, Pampulha, Belo Horizonte, Brazil, brentan@ehr.ufmg.br, rui.g182@gmail.com, gustavo.meirelles@ehr.ufmg.br,

(♮) Department of Manufacturing Systems Engineering and Management, California State University, Northridge, CA 91330, United States, silvia.carpitella@csun.edu,

(♯) Faculty of Science and Technology, Free University of Bozen-Bolzano, Piazza Università 5, 39100, Bolzano, Italy; ariele.zanfei@natec.unibz.it, andrea.menapace@unibz.it,

(±) Institute for Multidisciplinary Mathematics, Universitat Politècnica de València, Cno. de Vera s/n, 46022 Valencia, Spain, jizquier@upv.es.

1 Introduction

Managing water distribution systems (WDS) in large metropolitan areas is a complex task. As highly connected, buried infrastructures, WDS are often exposed to failures and pose difficult control problems. To recover its capacity, various rehabilitation alternatives can be considered. Yet, specially in large, intermittent WDS, the wide spectrum of available alternatives leads to complex decision-making processes. To support WDS managers, hydraulic models can help disclose the impacts of interventions in the system. To cope with the inherent uncertainty, simulation processes built on top of those hydraulic models can shed light on each type of intervention. In-depth evaluations of the solutions under various criteria can help. Considering pipe replacement as a strategy for water network rehabilitation, we combine water distribution system analysis with multi-criteria analysis to rank alternatives for pipe replacement. Eight performance criteria are used to evaluate the rehabilitation alternatives. The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [1] is adopted to rank the solutions. Our case study is the one proposed in the Battle of Intermittent Water Supply (BIWS) [2]. A budget constraint is set for alternatives' generation. Results show as pipe replacement is an important rehabilitation strategy, since the eight evaluation criteria are improved. Sensitivity analyses show the robustness of the best solutions, with just a few variations in ranking positions. The most frequent best solution is then hydraulically evaluated, showing the real benefits of pipe replacement in terms of pressure deficit reduction.

2 Methods

2.1 Leakage simulation

This work uses hydraulic theory to generate rehabilitation solutions. Simulations use Epanet 2.2 linked to Python using the library WNTR. Simulations are assessed by using TOPSIS. Leaks are distributed along the pipes to simulate leakage in a realistic way in WDS. We create a leaking node by splitting the corresponding pipe into two pipes connected by a node without demand but provided with an emitter that characterizes the leak. To avoid computational/mathematical problems derived from possible negative pressures, we create a new node which is linked to the leaking node by an artificial pipe, a short pipe of large diameter such to avoid additional headloss. A check valve to control the flow direction is added to this pipe.

2.2 Pipe replacement simulation

Pipe replacement is used by water companies to recover the hydraulic capacities of their WDS. After a pipe replacement, leaks are fixed, since the new pipe is fully watertight. A replaced pipe has same diameter and length as the old one, but its roughness is updated. this leak repair is simulated by setting the emitter coefficient of the leaking node to zero. Each pipe has a replacement cost, which is related to the price of the pipe itself and to the civil interventions as well. This cost limits the number of pipes to be replaced according to the budget of the water Company. In this work we follow the most recent references about cost as a function of diameter.

2.3 Evaluation Criteria

In addition to the costs, we use a set of indicators to evaluate the feasibility and advantages of a solution. Solutions are evaluated using eight multifaceted criteria (formulae omitted here).

I_1: percentage of hours that a consumer is served
I_2: proportion of consumers with continuous service
I_3: volume of water leakage
I_4: percentage of volume of water supplied to the users
I_5: pressure level at consumption nodes
I_6: percentage of users supplied continuously
I_7: average length of pipes under negative pressure
I_8: total energy consumed by pump stations

2.4 Multi-criteria and sensitivity analyses

Alternatives will be evaluated and ranked by using TOPSIS, which is able to deal with a huge number of alternatives, as it is the case of the present paper and can lead the evaluation by distinctively weighting the criteria, what allows sensitivity analyses (see, for example [3]).

3 Results

The developed methodology is applied to the BIWS network [CITA]. In this work only the pipe replacement solution is explored. This network has 2.859 junctions, 3.231 pipes, 7 pump, 6 reservoirs, 4 tanks, and 15 valves. Around 3600 leaks are set at a number of pipes. The final hydraulic model is then built with more than 10.000 nodes and 6.800 pipes. This makes hydraulic simulation harder than for the original setting and application of heuristic optimization is virtually impossible. Figure 1 presents the network topology depicting node elevation. A high elevation zone corresponds to the red nodes and a low elevation zone corresponds to the dark-blue nodes. Since the operational pressure is inversely proportional to the elevation, it is expected the high elevation zone to have supply problems derived from low pressures, while the low elevation zone will have problems with leaks due to high pressures.

A budget of €650.000 is considered, following the description of the Battle, and corresponds to the total amount of money that can be invested on rehabilitation. Considering this budget, 150

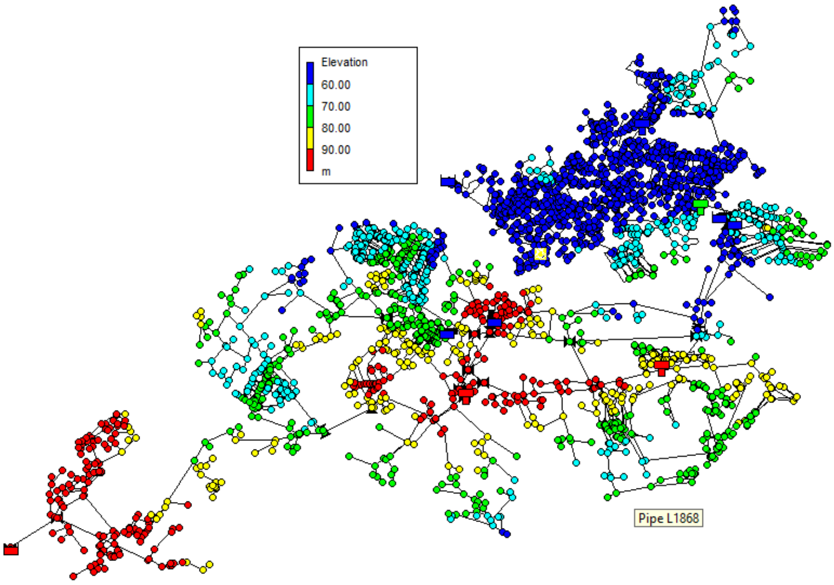


Figure 1: Water network topology of BWIS network highlighting the elevation of water system

solutions are simulated. Each solution is generated by randomly selecting pipes with leaking nodes to be replaced.

The solutions are evaluated based on the given indicators. Table 1 presents an statistical analysis of indicators. The box plot in Figure 2 analyzes the criteria value distribution. From Table 1 and the box plot in Figure 2 one verifies that some indicators do not vary significantly for pipe rehabilitation (e.g. I_1, I_4 and I_5). Comparing the statistical parameters and the criteria calculated for the original network, we note that I_1 is improved by all the solutions. However, a comparison among solutions shows slight differences between the minimum and maximum I_1 values. The same analysis can be explored for criterion I_4, which is improved by all the solutions, although, comparing minimum and maximum I_4 values, slight differences are noticed. Finally, the minimum of I_5 is virtually equal to the same indicator in the original network, while the maximum I_5 is slightly better than in the original network. The analysis on the other parameters shows that network rehabilitation by replacing pipes can improve indicators I_2, I_3, I_6, I_7 and I_8. More than that, observing the variation of these criteria, it is also possible to highlight that some solutions are better than others. It is also important to underline as some solutions can impair indicator I_6, and this is because reducing leaks in a certain region leads to higher pressures on this region but, since leaks depend on pressure, the remaining leaks can increase.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
Average	0.957	0.908	0.432	0.923	0.769	0.848	59064	6731649
Std	0.0014	0.0343	0.0061	0.0009	0.0036	0.0881	33178	79017
Min	0.954	0.729	0.380	0.920	0.763	0.547	30163	6213527
Max	0.962	0.922	0.441	0.928	0.790	0.921	155425	7037708
Original	0.882	0.708	0.489	0.874	0.760	0.698	93091	6565165

Table 1: Statistical characterization of pipe replacement solutions and evaluation criteria calculated for the original network without any intervention

A question remains: which solution is more suitable to be applied considering that all them cost virtually the same? The MCDM method TOPSIS is applied to answer this question, specifically

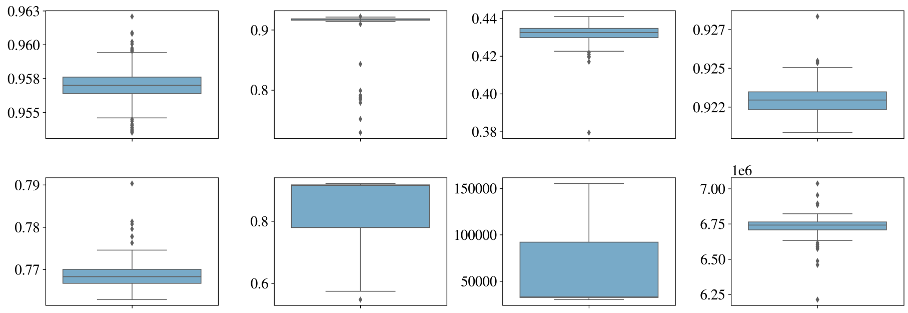


Figure 2: Box plot of each indicator used to evaluate the rehabilitation solutions

to rank the solutions based on evaluation criteria. Table 2 shows the five best solutions obtained by applying TOPSIS.

ID	TOPSIS Score	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
22	0.975254	0.961	0.921	0.423	0.926	0.781	0.922	31464.0	6459619
128	0.971187	0.958	0.918	0.432	0.925	0.770	0.918	31778.0	6769291
115	0.970083	0.957	0.917	0.431	0.923	0.770	0.917	32191.0	6714131
141	0.96904	0.958	0.918	0.433	0.922	0.768	0.911	32227.0	6736640
18	0.968505	0.955	0.917	0.433	0.923	0.769	0.916	32709.0	6680980

Table 2: Ranking of five best solutions and the corresponding evaluation criteria

First, consider that criteria may have different weights. A robust methodology for ranking solutions must handle this situation. For this reason, a sensitivity analysis of criteria weight is conducted.

Finally, to evaluate the real impact of the solutions on the hydraulic system, the best solutions are hydraulically simulated. The results of this analysis can help decision makers to understand how and where the hydraulic features (e.g. flow and pressure) are changed.

4 Conclusions

Rehabilitation is paramount for water supply managers. The diversity of alternatives to improve hydraulic and energy performance involves complex decision-making. To help in the decision-making process, this work has presented a methodology for ranking pipe replacement solutions based on the TOPSIS methodology. Based on eight multifaceted criteria, solutions are evaluated and compared. To assess the robustness of the proposal, a sensitivity assessment has to be applied. It can be done with groups of criteria. Finally, the best solutions have to be hydraulically checked and the accompanying improvements made explicit.

References

[1] Chakraborty, S, TOPSIS and Modified TOPSIS: A comparative analysis *Decision Analytics Journal*, 2:100021, 2022.

[2] https://wdsa-ccwi2022.upv.es/battle-of-water-networks/

[3] Brentan, B.M. and Carpitella, S. and Izquierdo, J. and Luvizotto Jr, E. and Meirelles, G., District metered area design through multicriteria and multiobjective optimization *Mathematical methods in the applied sciences*, 45(6):3254–3271, 2022.

25

Performance analysis of the constructive optimization of railway stiffness transition zones by means of vibration studies

M. Labrado[♭,1], J. del Pozo[♮], R. Cabezas[◇] and A. Arias[◇]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(♮) Universidad Europea,
Paseo de la Alameda 7, València, Spain.
(◇)Idvia 2020 Horizonte 2020 SL.
Av. d'Aragó 30, València, Spain.

1 Introduction

In railway tracks, transition zones are located between the civil structures and the embankments. In this sense, differential settlements take place due to the difference of stiffness between the two, resulting in direct damage to the track. Currently, railway managers have tried, without success, to solve this problem using granular wedges.

For this reason, to achieve a continuous and gradual track vertical stiffness in transition zones, a new type of transition wedge has been designed using prefabricated concrete slabs. In addition, it was studied if the new wedge was valid to be installed in urban environments, as it is thought that the transmission of railway vibrations through concrete could be very high.

With this purpose, firstly the design and validation of the new transition wedge has been carried out. To do this, the different alternatives were installed in a test field and the one with the best performance was selected. Once selected, an FEM was used to validate its performance compared to a granular wedge. Secondly, using the FEM of the validated solution, a comparison of the vibrational response of both alternatives was carried out to determine whether the new wedge is suitable for installation in urban environments.

The results showed that the new wedge using concrete slabs was an optimal solution to be installed in railway transition zones. However, these wedges did not show an adequate performance in urban environments, so the use of granular wedges was concluded to be more appropriate.

2 Design and validation of new prefabricated wedge

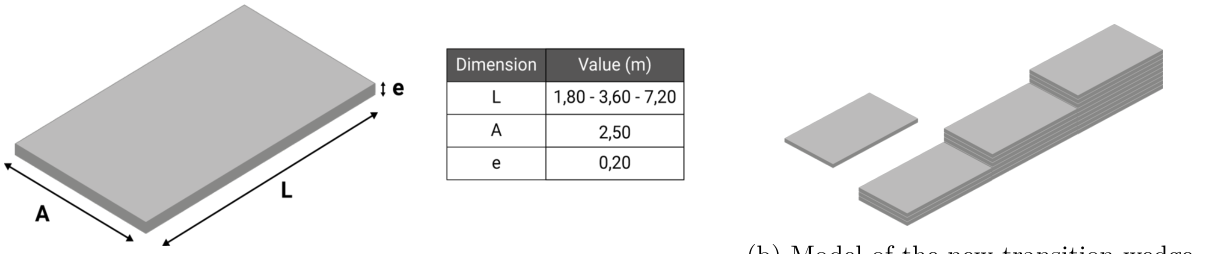
The proposed methodology for the design of a new transition wedge is then proposed with the main objective of obtaining vertical stiffness in the adjacent areas of the structures and providing optimum structural performance. Together with the methodology, the results will also be presented.

[1] milabpa@upv.es

2.1 Methodology

Approach to initial cases

The design of the transition wedges depends on the distance between the top of the structure and the runway [1]. Thus the new transition wedge has been designed to be adaptable to different scenarios and has been constructed using precast concrete slabs. For this document, four scenarios have been proposed whose specifications have been defined by ADIF (Administrador Infraestructuras Ferroviarias, Spain), depending on the distance between the top of the structure and the bottom of the sub-ballast: i) Case 1. Depth equal to 0 m; ii) Case 2. Depth less than 0.5 m; iii) Case 3. Depth between 0.5 m and 2.0 m; and, iv) Case 4. Depth greater than 2.0 m.



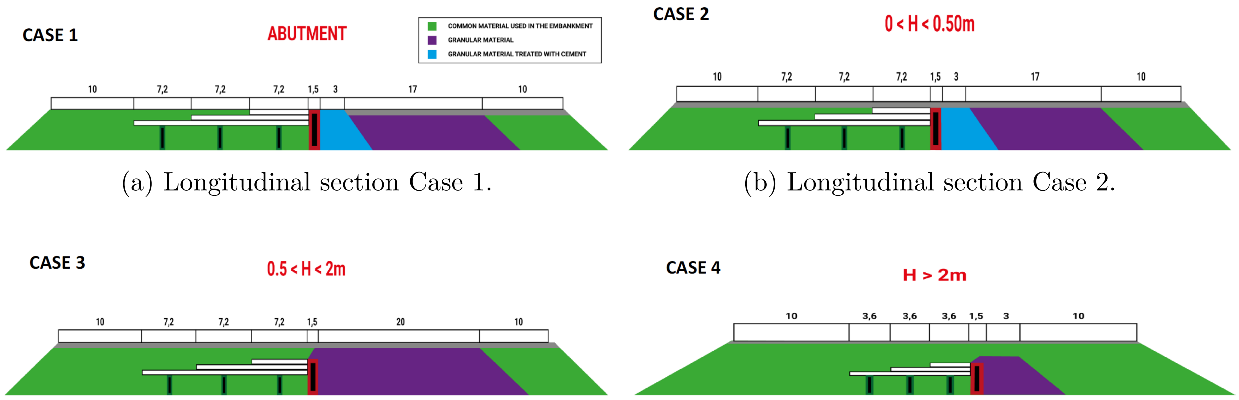
Dimension	Value (m)
L	1,80 - 3,60 - 7,20
A	2,50
e	0,20

(a) Prefabricated concrete slab dimensions.

(b) Model of the new transition wedge.

Figure 1: Design of the new transition wedge.

Two geometric models have been designed for the prefabricated wedges. The first design consists of a three-step wedge with a final length of 21.6 m and a height of 1.8 m. This design has been applied in cases 1, 2 and 3 (see Figures 2a, 2b and 2c). The second one has been designed with three steps with a final length of 10.6 m and a height of 1.8 m. This second design has been applied to case 4 (Figure 2d).



(a) Longitudinal section Case 1.

(b) Longitudinal section Case 2.

(c) Longitudinal section Case 3.

(d) Longitudinal section Case 4.

Figure 2: Longitudinal section of the case studies.

Field tests

The deformation determines the magnitude of the dynamic overloads transmitted to the track and, as a consequence, its deterioration [2]. For this reason, each case was reproduced in real size on a test track. The test field was set up in a stone quarry in Oliva (Valencia, Spain).

The tests consisted of the application of 3 load values on different sections of the wedge to study its behaviour under deformation. The results of the tests are shown in Figure 4 and, as can

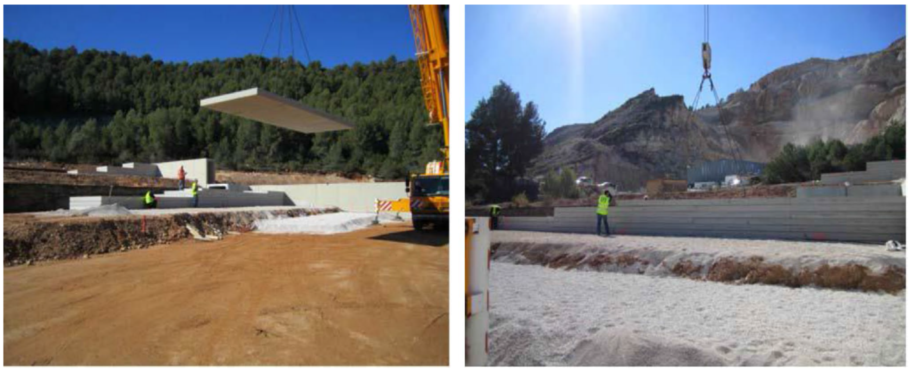
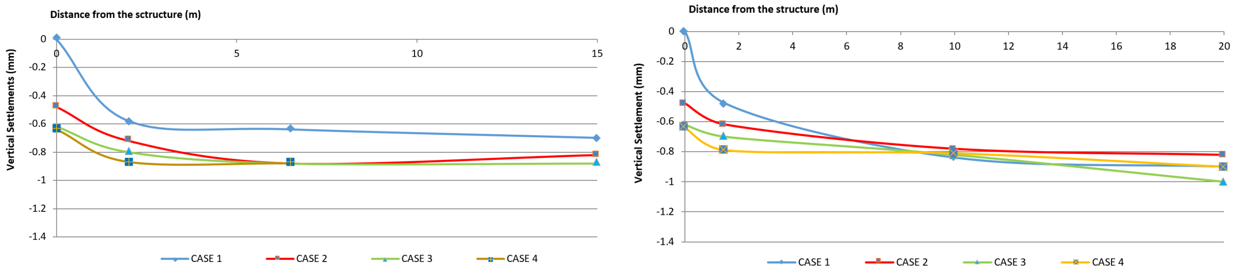


Figure 3: Installation of the prefabricated and granular wedge in a stone quarry located in Oliva (Valencia) Spain

be seen, the best performance is obtained for case 1 of depth = 0m.



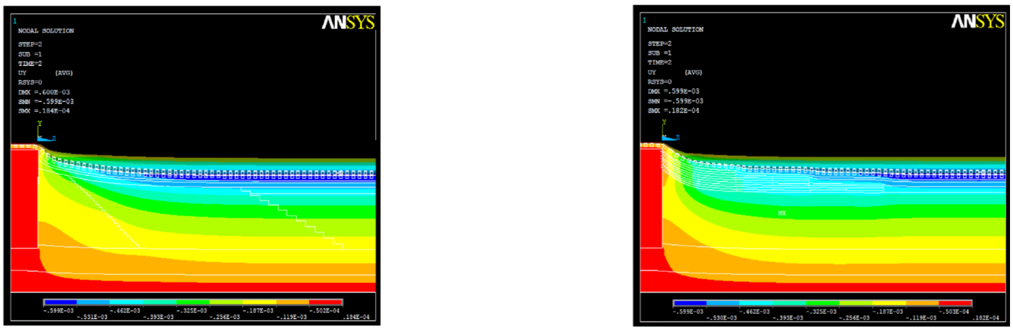
(a) Granular wedge.

(b) Prefabricated wedge.

Figure 4: Vertical settlement recorded in the field tests.

Numerical model

Following the determination of the optimal solution, the granulated and prefabricated wedges have been installed to i) calibrate and validate a Finite Element Model (FEM); and, ii) validate the prefabricated wedge as an optimal solution. The solution has been installed on the ADIF high-speed line in Antequera (Málaga, Spain). Both wedges have been calibrated using the real data acquired, comparing with the outputs obtained in the model. After this, the correct parameters have been obtained to validate the finite element model in Figure 5.



(a) Results for granular wedge model.

(b) Results for prefabricated wedge model.

Figure 5: Numerical model for the granular and prefabricated wedges.

2.2 Results

The comparison between granular and prefabricated transition wedges was made in terms of vertical displacements thanks to the calibrated and validated FEM. The deformation curve is shown in Figure 6, where the x-axis is the distance from the structure and the y-axis is the displacement of the top rail. The deformation curve is shown in Figure 7, where the x-axis is the distance from the structure and the y-axis is the displacement of the top rail. Both solutions show similar deformation, with the prefabricated solution being a better solution, as the maximum deformation is reached at a greater distance from the structure.

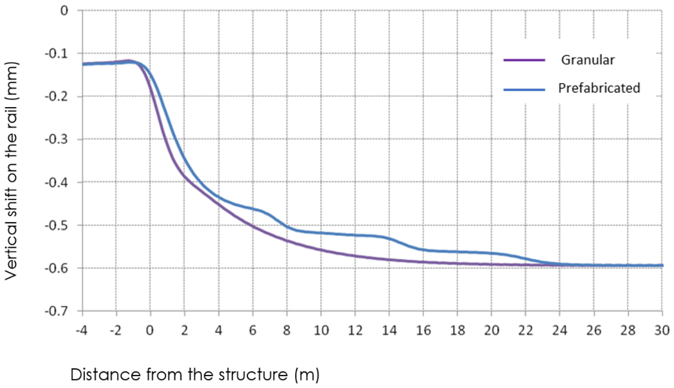


Figure 6: Deformation curve between granular and prefabricated wedges.

It is concluded that the FEM obtained has been validated and calibrated, being possible to verify that the use of a prefabricated transition wedge is an optimal solution compared to conventional granulated wedges.

3 New wedge vibration analysis in urban environments

Once the new transition wedge has been validated, a vibration study has been carried out in this section with the objective of assessing whether the new wedge is suitable for urban environments.

3.1 Methodology

Case approach

The ability of concrete structures to transmit vibrations could preclude the use of precast wedges in urban environments, as these vibrations could propagate to surrounding areas and granulated wedges would be more suitable. To determine its suitability, the validated FEM has been used to reproduce urban rail vehicle traffic.

In order to evaluate the vibration behaviour of the precast wedge and the granular wedge, the accelerations of the structure have been studied. The speed of rail vehicles has been set at a constant 80 km/h and three critical points have been evaluated: i) on a rail; ii) in the sub-ballast layer; and iii) at 0.9 m depth (half the height of the wedge). The points selected for assessment are shown in Figure 7.

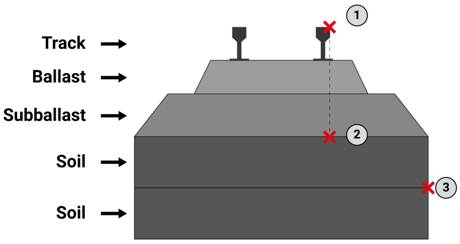


Figure 7: Data measuring points in track cross-section.

3.2 Results

Vibrations on the rail

The vibration results obtained for the railway track using the numerical model are shown in Figure 8. As can be seen, the vibration is very similar for the granular and prefabricated wedge. The vibrations induced by passing trains are between 50 and 80 Hz [3,4], so it is expected that increasing the stiffness of the materials close to the rail will increase the vibrations [5].

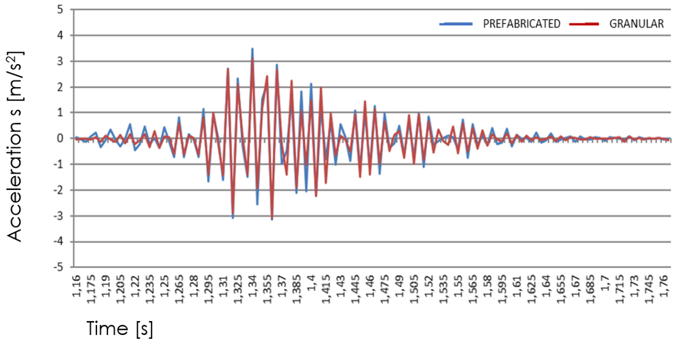


Figure 8: Accelerations on the rail at a 80 km/h train speed.

Vibration in the subballast layer

The results of the simulation are presented in Figure 10, which shows, firstly, a decrease in accelerations away from the vibrating core, as expected. Secondly, different vibration performances are observed for the two types of wedges. The prefabricated wedge has higher acceleration peaks, which leads to the conclusion that the granular wedge reduces vibration more than the prefabricated wedge. Secondly, different vibration performances are observed for the two types of wedges. The prefabricated wedge has higher acceleration peaks, which leads to the conclusion that the granular wedge reduces vibration more than the prefabricated wedge. The granular wedge is able to reduce peak vibration by 33%.

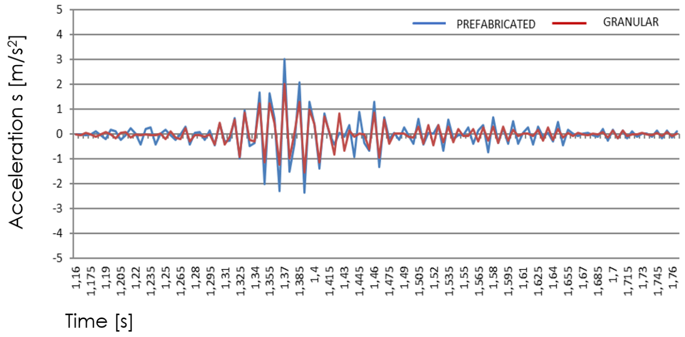


Figure 9: Accelerations in the subballast layer at a 80 km/h train speed.

Depth = 0.9 m from the subballast line

The third case study also shows the expected vibration behaviour, as shown in Figure 11. Vibrations are reduced compared to the previous case, with accelerations decreasing as they move away from the railway track. Both wedges show different vibration behaviour. Therefore, special attention was paid to the peaks at critical points, and it was determined that the granular wedge reduces the maximum vibrations by 30% compared to the prefabricated wedge.

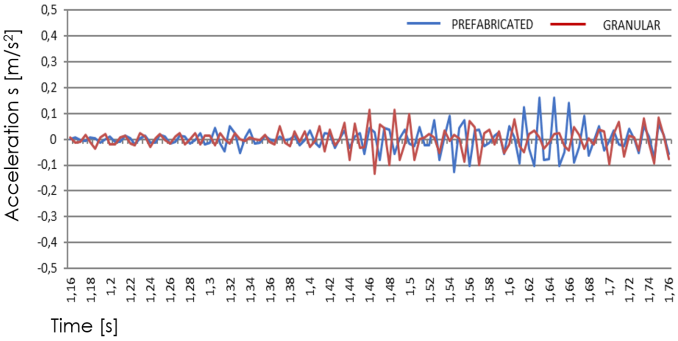


Figure 10: Accelerations on a depth of 0.9 m at a 80 km/h train speed.

4 Conclusions

In this paper, the performance of a new type of transition wedge has been evaluated, which has been made possible through the division of the study into two stages. The first stage consisted of a comparison between the new transition wedge and the traditional granular wedge. The second stage consisted of evaluating the performance of the newly designed transition wedge in an urban environment.

The evaluation of the results presented in this document has made it possible to reach the following conclusions:

- Both granular wedges (correctly executed) and precast wedges allow for a gradual transition of stiffness.
- The optimal installation scenario for the precast wedges is to place them just below the sub-ballast line.
- New prefabricated wedges improve structural performance.
- The new prefabricated wedges have poorer vibration performance and are less suitable for urban environments.

References

[1] Code UIC, 719R Earthworks and track bed for railwaylines. *Paris Union Int des Chemins Fer*, 117, 2008.

[2] Gallego I., Muãoz J., Rivas A., Sánchez-Cambronero S., Vertical track stiffness as a new parameter involved in designing high-speed railway infrastructure. *Journal of transportation engineering*, 137(12):971, 2012.

[3] Forrest J.A., Hunt H.E.M., A three-dimensional tunnel model for calculation of train-induced ground vibration. *Journal of Sound Vib*, 294(4–5):678–705, 2006

[4] Greer R, Manning C., Vibration isolation for railways. *Acoust Bull* 23:3, 1998.

[5] Gardien W., Stuit H.G., Modelling of soil vibrations from railway tunnels. *Sound Vib* 267(3):605—619, 2003

26

Can any side effects be detected as a result of the COVID-19 pandemic? A study based on social media posts from a Spanish Northwestern-region

A. Larrañaga[♭,1], G. Vilar[♮], J. Martínez[♭,♮,◇] and I. Ocarranza[♮,♯]

(♭) CINTECX, Universidade de Vigo. Campus Lagoas-Marcosende, 36310, Vigo, Spain.
(♮) Applied Mathematics I, Telecommunications Engineering School, Universidade de Vigo.
Campus Lagoas-Marcosende, 36310, Vigo, Spain.
(◇) CITMAga. Santiago de Compostela, Spain.
(♯) Possible Incorporated. Vigo.

1 Introduction

In today's world, the usage of social media platforms has popularized to the point where the volume of information they offer is unmatched by other more traditional surveying methods. For example, it is estimated that, for Instagram, over half a billion users are active everyday, and over 95 million images and videos get posted daily[2]. It is no wonder then that interest has surfaced into how to properly utilize this data in different areas of study such as tourism [1] of marketing [2].

It is important to note that, while social media posts truly contain useful data, this data does not present itself in an obvious and direct way. Instead, it has to be extracted through different tools, such as those performing face recognition or sentiment analysis tasks. Thus, the main objective of this work will be the exploration of different tools that allow for the extraction of useful information from social media, particularly from Instagram. Many previous works exist dealing with this kind of problem. This one in particular was born as an extrapolation from [3], but other relevant works such as [4] or [5] can be cited by their usage of face recognition and sentiment analysis respectively, both in the context of tourism. The base idea of our work was set as an study of the Instagram posts with location set in a particular city in 2018 and 2021 to allow for an additional study of the effects that the pandemic of COVID-19 might have had on this kind of data.

2 Data

The dataset used in this study is made up of 10140 Instagram posts, which were requested to the Instagram Basic Display API, with the requirement of having the location set in the city of Vigo (Pontevedra, Spain). This city is the most populated of the northwest region of Spain (with 293.837 inhabitants recorded as of 01-01-2021[3]), and also receives a considerable amount of visitors which ensures a good flow of social media posts coming from there.

[1] ana.larranaga.janeiro@uvigo.es

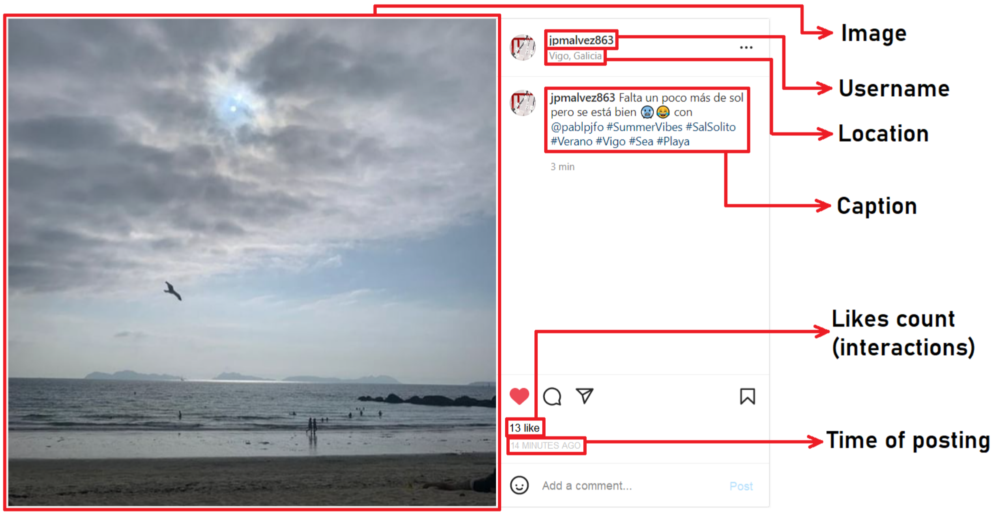
[2] https://www.wordstream.com/blog/ws/2017/04/20/instagram-statistics

[3] Spanish Statistical Office, National Statistics Institute (INE), https://www.ine.es/en/

Half of the posts are from 2018 representing the social profiles before the pandemic, while the other half are from 2021, over a year after said pandemic hit. The posts were made in both years on the months of May and June, so they belong to the same season. The good climate of the area in those months, paired with the fact that Vigo is a coastal city, guarantees that both visitors and locals will partake in many outdoor activities, and likely post about them in social media, which serves as great fuel for an study like this.

The reason why the Instagram platform was chosen for this study is due to its visual nature. Posts must be accompanied of at least one photo or video, which means an image analysis can be reliably performed as a way to obtain sufficient information. Other social media platforms such as Twitter could be used, but the methods employed should shift to a more text-based approach.

Some of the data available for each post is more straightforward, like the date and time, or the number of interactions received, that is, likes and comments. However, other more complex attributes like the image or the caption will require some methods to extract information from them, which will be introduced in the next section. An example on how all of this information would be displayed on a post can be seen on figure 1.



jpmalvez863 Falta un poco más de sol pero se está bien 🥶😂 con @pablpjfo #SummerVibes #SalSolito #Verano #Vigo #Sea #Playa

Figure 1: Example of an Instagram post, detailing the available information.

3 Methods

The posts obtained are only filtered by location and date of publication. This means that a portion of the dataset will be composed of posts made by businesses such as restaurants, which are of no relevance to this particular study. This is why a method to filter out posts that come from business accounts is required. For that purpose, a supervised machine learning model was trained. In order do so, a sample of over 400 randomly selected posts was manually labeled, based on whether each post was made by a personal account or a business account. For the purpose of this study, we considered a business account to be any account that was trying to sell a product or service to its readers, while any other account was labeled as personal. The training and test sets came from an $85-15$ split of this sample.

Four different supervised classification algorithms were considered: random forest, support vector machine, k-nearest neighbors and multilayer perceptron. The predictors used by the models were the following: number of likes and comments (interactions) received, day of week of publication, hour of publication (integer from 0 to 23), number of hashtags and mentions used in the caption, character count of the caption, amount of numbers in the username of the poster, and dummy variables indicating the year (2018 or 2021), whether the post is a video, whether it contains more than one media file (image or video) or just one, and the presence of a url or a phone number in the caption.

The hyperparameters of each model were tuned using 10-fold cross-validation, evaluating the performance through the accuracy, precision, sensitivity, and area under the receiver operating characteristic curve (AUC-ROC). Once tuned, the random forest was discarded due to poor performance, and the used model was an ensemble model of the support vector machine, the k-nearest neighbors and the multilayer perceptron. This model classified 6293 out of the 10140 posts as coming from personal accounts.

In order to obtain useful information from more complex attributes of this newly labeled data such as the image or the caption of posts, some additional methodology is required. For the images, an image classification model and a face recognition model were applied. For the captions, data about the sentiments and education level was obtained.

The education level was inferred for each post through the use of readability formulas. This kind of formulas give a score to a text that serves as a measure of how difficult it is to understand. In particular, the Flesch reading-ease score [6] was used, which follows the following formula:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59. \tag{1}$$

This formula will assign higher scores to texts with simple words and short sentences, which are easier to read, while giving lower scores to harder texts. Thus, it is possible to infer an approximate education level of the author of each caption by assigning a higher education level to the authors of posts which received a lower Flesch score, and a lower education level to authors of posts with higher scores. It is worth noting that, since the texts used in this work belong to social media, it is to be expected that they are more on the simpler side, regardless of the education of the author.

The sentiment analysis over the captions was performed through the VADER [7] model. This is a lexicon-based model that "combines these lexical features with consideration for five generalizable rules that embody grammatical and syntactical conventions that humans use when expressing or emphasizing sentiment intensity" ([7]). Some of the rules the model takes into account are, for example, punctuation signs or capital letter usage. Thus, the preprocessing step of the caption data, usually performed to clean unwanted characters from the texts, will only involve translation to English and removal of hashtags and mentions, both for this model and the Flesch score calculation.

The facial recognition software used was Face++[4]. The model returns predictions for the age and gender of each detected face, which serves as demographic information of the person behind each post. The accuracy of this model has been assessed before [8], showing that it produces acceptable results. Additional experiments were however performed while in the making of this study, utilizing the FairFace dataset [9]. The results show a face detection rate of 0.996, a gender prediction rate of 0.846, with an acceptable error on the prediction for the ages (which can not be exactly precised due to the ages of the dataset being given as an interval, instead of an integer).

Additionally, the ResNeXt model [10] was used, through its ImageNet1k [11] implementation. This allowed for general classification of the images based on objects detected in them, which was useful particularly for the many posts in which no face was present. Then, one of 67 image tags[5] can be assigned to each image based on its detected subject.

Finally, in order to perform a better analysis of the results, a clustering method was applied to the 6293 posts classified as coming from personal accounts. The model used was the k-means algorithm with $k = 3$. It was applied independently for the 2018 and the 2021 data. The variables

[4] https://www.faceplusplus.com/

[5] https://github.com/noameshed/novelty-detection/blob/master/imagenet categories.csv

used were the number of interactions received by a post (likes and comments, 2 variables), measuring the reach the post had; the number of hashtags and mentions (2 variables) used in the caption, measuring the reach the author tried to obtain for the post; and the Flesch score of the caption as given by the formula 1 after removing the hashtags and mentions.

To ensure a better performance of the algorithm, a logarithmic transformation was applied to the likes and comments variables, then all 5 variables were normalized to a $0 - 1$ range. Figure 2 shows the scatter plots of the likes received versus the hashtags used in 2018 (left) and 2021 (right), differentiating by the clusters to which each post belongs. This gives an intuition of the kind of posts that belong to each cluster, even if they are not perfectly separated. Posts from cluster 1 used few hashtags and received few likes, posts from cluster 2 used few hashtags yet received many likes, and posts from cluster 3 used a high number of hashtags. Note that this is just a generalization that is not true for every post of the dataset, as can be seen by the fact that a small subset of posts from cluster 1 seem to have received a high number of likes.

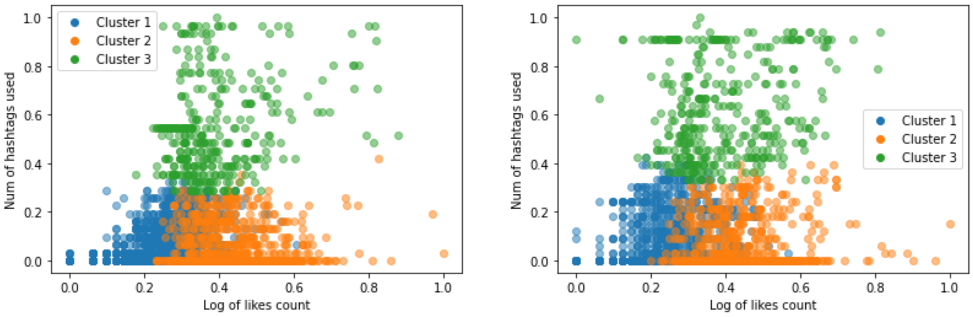


Figure 2: Scatter plots for the clusters formed for 2018 (left) and 2021 (right).

4 Results

The method chosen to present the data obtained form the tools reviewed in last section (VADER, Flesch reading-ease score, Face++) was the plotting of different histograms for each attribute, cluster and year. In particular, data about age and gender can be found in the figure 3, data about the reading ease of captions is in the figure 4, and finally figure 5 presents data about the sentiment analysis.

This graphs serve multiple purposes, since not only is it possible to study the differences of the clusters based on this attributes, but also how those attributes were affected by the global pandemic of 2020 (that is, the differences on the attributes between 2018 and 2021). For example, figure 3 shows that the amount of young people that can be classified in cluster 1 was drastically reduced, while figure 5 shows that posts belonging to clusters 2 and 3 carry more positive sentiments in their captions.

Additional context about the clusters can be obtained through the ResNeXt model, that provided information based on the posts' images without the need of faces in them. For example, the model detected a lot of swimwear in the posts of the second cluster from 2021, but not on 2018, which means more people from that cluster chose to upload images of themselves at the beach after the pandemic. It is also unlikely that this is due to a huge difference of temperatures between the periods of time considered for both years. The reason is that a lot of swimwear was detected in posts of cluster 3, while very few was detected for cluster 1, independently of the year.

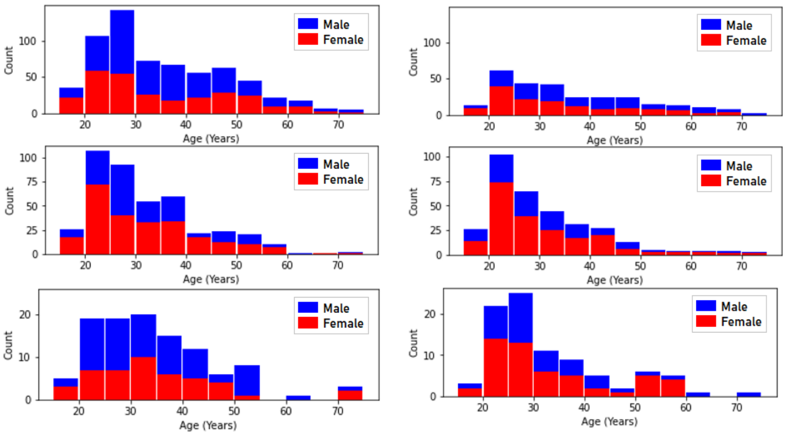


Figure 3: Ages predicted by Face++ on the faces detected on the posts of 2018 (left) and 2021 (right); for clusters 1 (top), 2 (middle) and 3 (bottom); and differentiating between men (blue) and women (red).

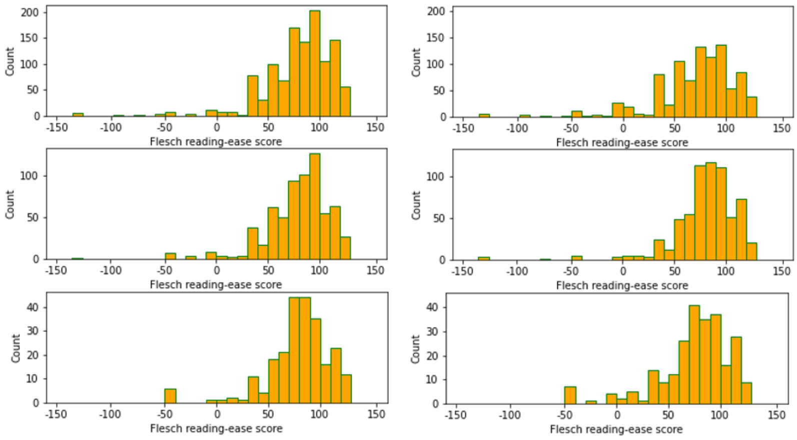


Figure 4: Flesch reading-ease scores for the captions of 2018 (left) and 2021 (right); for clusters 1 (top), 2 (middle) and 3 (bottom).

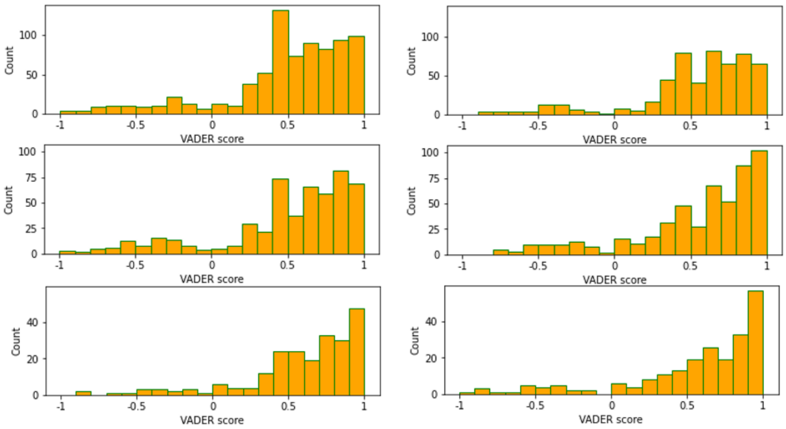


Figure 5: VADER scores for the captions of 2018 (left) and 2021 (right); for clusters 1 (top), 2 (middle) and 3 (bottom).

5 Conclusions and Future work

In this work, different tools that allow for the extraction of social and demographic information of people through their social media posts have been presented. The usage of methodology such as the one in this work could be a great addition to multiple areas including tourism and marketing. The volume of unfiltered data obtainable from social media is unmatched by any traditional survey, so learning how to make use of such data could become crucial in the near future.

This study only worked with data coming from Instagram, which is characterized by requiring posts to be accompanied by a media file, usually an image but with the possibility of a video too. Other social media sites vary from Instagram not only in the demographic of their users, but also in the structure of their posts. For example, TikTok is based around posting videos instead of images, while Twitter was built as a text-based social media page first, even if it allows media files too. All of this means that it would be interesting to explore new tools in future works that perform similarly on videos and text, to maximize the information obtained while decreasing the bias on the user base of each platform.

Acknowledgments: The work of Ana Larrañaga has been supported by the 2020 predoctoral grant of the University of Vigo.

References

[1] Tenkanen H., Di Minin E., Heikinheimo V., Hausmann A., Herbst M, Kajala L., Toivonen T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. Scientific reports, 7(1), 1-11.

[2] Park S.B., Ok CM, Chae B.K. (2016). Using Twitter data for cruise tourism marketing and research. Journal of Travel & Tourism Marketing, 33(6), 885-898.

[3] Kalra G., Yu M., Lee D., Cha M., Kim D. (2018). Ballparking the Urban Placeness: A Case Study of Analyzing Starbucks Posts on Instagram. *In International Conference on Social Informatics*, 291–307.

[4] González-Rodríguez M., Díaz-Fernández M.C., Gómez C. (2020). Facial-Expression Recognition: an emergent approach to the measurement of tourist satisfaction through emotions. Telematics and Informatics. 51. 101404.

[5] Kirilenko A.P., Stepchenkova S.O., Kim H., Li X.R. (2018). Automated Sentiment Analysis in Tourism: Comparison of Approaches. Journal of Travel Research, 57(8), 1012-1025.

[6] Flesch R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221-233.

[7] Hutto C., Gilbert E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216-225.

[8] Jung S.G., An J., Kwak H, Salminen J., Jansen B.J. (2018). Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race. *Proceedings of the 12th International AAAI Conference on Web and Social Media*, 624-627.

[9] Kärkkäinen K., Joo J. (2019). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. arXiv:1908.04913.

[10] Xie S, Girshick R., Dollar P., Tu Z., He K. (2017). Aggregated Residual Transformations for Deep Neural Networks. arXiv:1611.05431.

[11] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2014). ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575.

27

Probabilistic analysis of a cantilever beam with load modelled via Brownian motion

J.-C. Cortés[♭], E. López-Navarro[♭,1], J.I. Real Herráiz[♭], J.-V. Romero[♭] and M.-D. Roselló[♭]

(♭) I. U. de Matemática Multidisciplinar,
Universitat Politècnica de València,
Camí de Vera, s/n, 46022, València.

1 Introduction

In this contribution we deal with the stochastic analysis of the deflection of a cantilever beam with its load modelled through Brownian motion. This phenomenon can be mathematically described by means of the following fourth-order differential equation [1]

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

with boundary conditions

$$Y(0) = 0,\ Y'(0) = 0,\ Y''(l) = 0,\ Y'''(l) = 0, \tag{2}$$

where $Y(x)$ represents the deflection of the beam, E is the Young's modulus of elasticity of the material of the beam, I denotes the moment of inertia of the cross section of the beam around a horizontal line through the centroid of the cross section (hereinafter, we will denote it as i, indicating that it will be a deterministic value), l is the length of the beam and $W(x)$ represents the density of downward force acting vertically on the beam at the space point x. We will assume that E and $W(x)$ are aleatoric factors due to the heterogeneity of the material of the beam and the load. For the latter, we will assume that $W(x)$ is given by the sum of a deterministic value, w_0 and a certain random quantity given by Brownian motion, $B(x)$

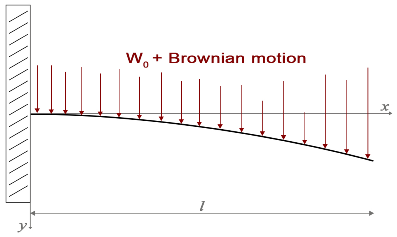


Figure 1: Cantilever beam with load modelled via Brownian motion.

$$W(x) = w_0 + B(x), \quad 0 < x \leq l. \tag{3}$$

In Figure 1 we can observe a scheme of this model.

The aim of this contribution is to obtain the first probability density (1-PDF) function of the stochastic solution, and other quantities of interest, as the maximum deflection and slope, using the Random Variable Transformation method (RVT) [2], which we will develop in Section 2. All these theoretical findings are illustrated via a numerical example in Section 2.

[1] ellona1@doctor.upv.es

2 Computing the first probability density function

This section is devoted to obtain the 1-PDF of the stochastic solution that represent the deflection of the cantilever beam, and other quantities of interest. In order to do this, we will take advantage of the Karhunen–Loève expansion of the Brownian motion [3]

$$B(x) = \mu_B(x) + \sum_{j=1}^{\infty} \sqrt{\nu_j}\phi_j(x)\xi_j(\omega), \;\; \omega \in \Omega, \;\; 0 < x \leq l, \tag{4}$$

where $\mu_B(x) = 0$, $\xi_j(\omega)$ are independent and identically distributed random variables, $\xi_j(\omega) \sim \mathrm{N}(0,1)$, $j = 1, 2, \ldots$, and

$$\nu_j = \frac{4l^2}{\pi^2(2j-1)^2}, \quad \phi_j(x) = \sqrt{\frac{2}{l}} \sin\left(\frac{(2j-1)\pi}{2l}x\right), \quad j = 1, 2, \ldots$$

are the eigenvalues and eigenfunctions obtained when solving the homogeneous Fredholm integral equation of second kind [4]. Now, we will consider the approximation of $B(x)$ obtained by truncating its Karhunen–Loève expansion at N. So, the model is approximated via the following stochastic differential equation

$$\frac{\mathrm{d}^4 Y(x)}{\mathrm{d}x^4} = \frac{1}{Ei}\left(w_0 + \sum_{j=1}^{N} \sqrt{\nu_j}\phi_j(x)\xi_j(\omega)\right), \quad 0 < x < l. \tag{5}$$

To obtain the 1-PDF of the deflection we first need to explicitly calculate the stochastic solution of model (5)-(2). We can use, for example, the Laplace transform to obtain the solution which is given by the following parametric stochastic process, $0 < x \leq l$,

$$\begin{aligned} Y(x) = \frac{1}{Ei} & \left(\frac{x^2}{2}\left(\frac{w_0}{2}l^2 + l\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2} - \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l - \sin(b_j l)}{b_j^3}\right)\right. \\ & \left. + \frac{x^3}{6}\left(-w_0 l - \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2}\right) + \frac{w_0}{24}x^4 + \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-6b_j x + b_j^3 x^3 + 6\sin(b_j x)}{6b_j^5}\right). \end{aligned} \tag{6}$$

Then, we are going to use RVT method to obtain the expression of the 1-PDF of (6). In short, RVT is a method to obtain the PDF of a random vector V that results from mapping of another random vector U whose PDF is known. We have considered that E and ξ_j, $j = 1, \ldots, N$ are independent whose PDF's are known. In order to make this abstract a light read, we suppress the calculations and give the final result of the 1-PDF, that is given by

$$\begin{aligned} f_{Y(x)}(y) = \mathbb{E}_{\xi_1,\ldots,\xi_N} & \left[f_E\left(\frac{1}{Ei}\left(\frac{x^2}{2}\left(\frac{w_0}{2}l^2 + l\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2} - \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l - \sin(b_j l)}{b_j^3}\right)\right.\right.\right. \\ & \left.\left. + \frac{x^3}{6}\left(-w_0 l - \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2}\right) + \frac{w_0}{24}x^4 + \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-6b_j x + b_j^3 x^3 + 6\sin(b_j x)}{6b_j^5}\right)\right) \\ & \cdot \left| -\frac{1}{E^2 i}\left(\frac{x^2}{2}\left(\frac{w_0}{2}l^2 + l\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2} - \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l - \sin(b_j l)}{b_j^3}\right)\right.\right. \\ & \left.\left.\left. + \frac{x^3}{6}\left(-w_0 l - \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2}\right) + \frac{w_0}{24}x^4 + \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-6b_j x + b_j^3 x^3 + 6\sin(b_j x)}{6b_j^5}\right)\right|\right], \quad 0 < x \leq l. \end{aligned} \tag{7}$$

Other important characteristics that we can obtain are the maximum deflection, D, and the maximum slope, S, these being a cantilever beam, are obtained at the free end of the beam. The maximum deflection is obtained by evaluating the solution at l and the maximum slope by evaluating the derivative at l and are given by the following expressions

$$\begin{aligned} D = Y(l) =& \frac{1}{Ei}\left(\frac{w_0}{8}l^4 + \frac{1}{3}l^3\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2} - \frac{1}{2}l^2\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l-\sin(b_j l)}{b_j^3}\right. \\ &\left. +\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-6b_j l + b_j^3 l^3 + 6\sin(b_j l)}{6b_j^5}\right), \end{aligned} \tag{8}$$

and

$$\begin{aligned} S = Y'(l) =& \frac{1}{Ei}\left(\frac{w_0}{8}l^3 + \frac{1}{2}l^2\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2} - l\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l-\sin(b_j l)}{b_j^3}\right. \\ &\left. +\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-2 + b_j^2 l^2 + 2\cos(b_j l)}{2b_j^4}\right). \end{aligned} \tag{9}$$

Then, applying again RVT method we can obtain the PDF of the maximum deflection,

$$\begin{aligned} f_D(d) =& \mathbb{E}_{\xi_1,\ldots,\xi_N}\left[f_E\left(\frac{1}{di}\left(\frac{w_0}{8}l^4 + \frac{1}{3}l^3\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2} - \frac{1}{2}l^2\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l-\sin(b_j l)}{b_j^3}\right.\right.\right. \\ &\left.\left. +\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-6b_j l + b_j^3 l^3 + 6\sin(b_j l)}{6b_j^5}\right)\right)\left|-\frac{1}{\delta^2 i}\left(\frac{w_0}{8}l^4 + \frac{1}{3}l^3\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2}\right.\right. \\ &\left.\left.\left. -\frac{1}{2}l^2\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l-\sin(b_j l)}{b_j^3} + \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-6b_j l + b_j^3 l^3 + 6\sin(b_j l)}{6b_j^5}\right)\right|\right], \end{aligned} \tag{10}$$

and the PDF of the maximum slope

$$\begin{aligned} f_S(s) =& \mathbb{E}_{\xi_1,\ldots,\xi_N}\left[f_E\left(\frac{1}{si}\left(\frac{w_0}{8}l^3 + \frac{1}{2}l^2\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2} - l\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l-\sin(b_j l)}{b_j^3}\right.\right.\right. \\ &\left.\left. +\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-2 + b_j^2 l^2 + 2\cos(b_j l)}{2b_j^4}\right)\right)\left|-\frac{1}{\theta^2 i}\left(\frac{w_0}{8}l^3 + \frac{1}{2}l^2\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{1-\cos(b_j l)}{b_j^2}\right.\right. \\ &\left.\left.\left. -l\sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{b_j l-\sin(b_j l)}{b_j^3} + \sqrt{\frac{2}{l}}\sum_{j=1}^{N}\xi_j\frac{-2 + b_j^2 l^2 + 2\cos(b_j l)}{2b_j^4}\right)\right|\right]. \end{aligned} \tag{11}$$

3 Numerical example

This section is dedicated to illustrate the above theoretical conclusions. We take the following data for the deterministic parameters of the model (5): the length of the beam, $l = 10$ m, the moment of inertia, $i = 722 \cdot 10^{-8}$ m^4, and $w_0 = 20$. For the random parameters, we will assume that the Young's modulus of elasticity, E, has a Gaussian distribution, $E \sim \mathrm{N}(210 \cdot 10^9; 420 \cdot 10^7)$. We will consider a Karhunen-Loève expansion truncated at order $N = 1$ to approximate the Brownian motion, $B(x)$.

In Figure 2, we show the graphical representation of the 1-PDF at different spatial points. We can observe, that the variance increases as the position does.

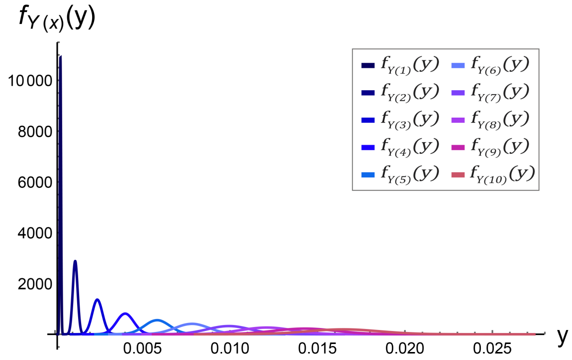


Figure 2: 1-PDF, $f_{Y(x)}(y)$, of the solution stochastic process (6), computed by (7), at different spatial position $x \in \{1, ..., 10\}$ of the cantilever beam considering an approximation of the Brownian motion, $B(x)$, via a Karhunen-Loève expansion truncated at order $N = 1$.

In Figure 3, we show on the left hand side the plot of the PDF of the maximum deflection and on the right hand side the PDF of the maximum slope.

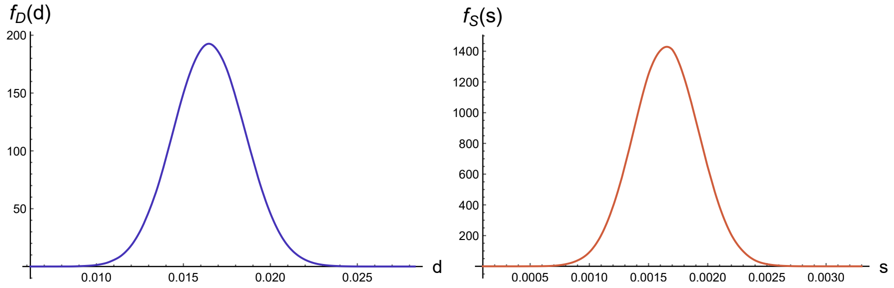


Figure 3: Left: PDF, $f_D(d)$, of the maximum deflection at the free end. Right: PDF, $f_S(s)$, of the maximum slope at free end.

Finally, in Figure 4 we show the graphical representation of the maximum deflection for different values of the truncation order, N. We can observe in this zoom, that the approximations are very similar with $N = 1$.

4 Conclusions

In this work, we have obtained a probabilistic description via the first probability density function of the stochastic solution of the fourth-order random differential equation, that describes the deflection of a cantilever beam whose load is modelled through Brownian motion. In addition, we have obtained other densities of quantities of interest such as the maximum deflection and the maximum slope at the free end of the beam. In order to obtain the probability density function we have taken advantage of the Random Variable Transformation method. Finally, we have illustrated these theoretical findings with a numerical example.

Acknowledgments

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).

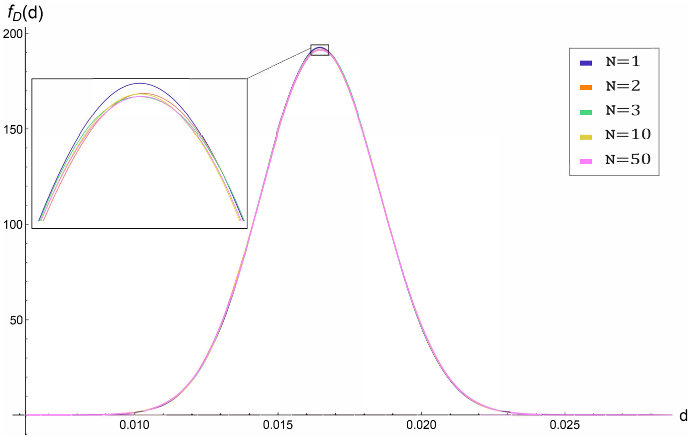


Figure 4: PDF of the maximum deflection at free end, $f_D(d)$, for different values of the truncation order, N, to approximate the Brownian motion by its Karhunen-Loève expansion.

References

[1] Öchsner, A., Classical Beam Theories of Structural Mechanics. Springer, 2021.

[2] Soong, T., Random Differential Equations in Science and Engineering. New York, Academic Press, 1973.

[3] Gabriel Lord, J., Powell, Catherine E., Shardlow, T., An Introduction to Computational Stochastic PDEs. Cambridge University Press, 2014.

[4] Ghanem, R.G., Spanos, P.D., Stochastic Finite Elements: A Spectral Approach, Dover Publications, New York, 2003.

Pivoting in ISM factorizations

J. Mas$^{\flat,1}$ and J. Marín$^{\natural}$

($\flat$) Universitat Politècnica de València,
Camino de Vera s/n, 46022 Valencia, Spain.
($\natural$) Universitat Politècnica de València,
Camino de Vera s/n, 46022 Valencia, Spain.

1 Introduction

In this work we study pivoting techniques for the balanced incomplete factorization preconditioner (BIF) [7], for solving ill-conditioned sparse nonsingular linear systems of equations of the form

$$Ax = b, \;\; A \in \mathbb{R}^{n\times n}, \;\; b \in \mathbb{R}^n \tag{1}$$

using iterative Krylov methods. There are different pivoting techniques being partial, complete and rook pivoting the more important ones [5,6]. Basically, at a given step of Gaussian elimination pivoting looks for an element sufficiently large in magnitude in the remaining submatrix, the Schur complement, to use it as the next pivot. These techniques involve row and possibly column permutations of the matrix that supposes a computational overhead. In this sense, partial pivoting is the cheapest pivoting technique, since it looks only in the first column of the Schur complement. Close behind is rook pivoting [6], it selects a pivot with maximum absolute value in his row and column, moving first to the biggest entry in magnitude in the first column, then it moves in the corresponding row, and then again in the column, and so on until the requirement is fulfilled. Finally, complete pivoting is the most expensive one, but guarantees the largest pivot at any stage however because the pivot is the entry of biggest magnitude in all the Schur complement.

BIF preconditioning is based on the incomplete Sherman-Morrison decomposition, ISM. The ISM decomposition uses recursion formulas derived from the Sherman-Morrison formula and was introduced in [7] as a method for computing approximate inverse preconditioners. In [8] and [9] the authors show that, applying the ISM algorithm to A and A^T, it is possible to compute an incomplete LDU factorization. Moreover, the inverse factors are also available and they influence the computation of the LDU factorization, and vice versa. In addition, the availability of the direct and inverse factors is exploited to implement norm based dropping rules [3]. The numerical results show that BIF is a robust algorithm comparable to other techniques as ILU(τ) [1], *ILUT* [10] and RIF [2]. Nevertheless, as mentioned above, computing stable (incomplete) factorizations for ill-conditioned problems still require the application of pivoting techniques. Here we show that with a slight modification of the ISM recursion formulas it is possible to incoporate pivoting to BIF.

[1] jmasm@imm.upv.es

2 The ISM decomposition

The ISM decomposition computes approximate inverse preconditioners since it obtains a factorization of the (shifted) inverse matrix of A, as

$$s^{-1}I - A^{-1} = s^{-2} Z D_s^{-1} V_s^T, \tag{2}$$

where $s > 0$ is a given scalar and the columns of the matrices Z and V_s are computed using the recursion formulas

$$z_k = e_k - \sum_{i=1}^{k-1} \frac{v_i^T e_k}{s r_i} z_i \quad \text{and} \quad v_k = y_k - \sum_{i=1}^{k-1} \frac{y_k^T z_i}{s r_i} v_i, \tag{3}$$

for $k = 1, 2, \ldots, n$. In (3) the vector e_k (e^k) denotes the $k-$th column (row) of the identity matrix, $y_k = (a^k - s e^k)^T$ where a^k denotes the k-th row of A, and

$$r_k = 1 + y_k^T z_k / s = 1 + v_k^T e_k / s \tag{4}$$

are the entries of the diagonal matrix D_s.

It was proved in [8] that for symmetric matrices the factorization $A = LDL^T$ and the decomposition (2) satisfy

$$D = sD_s, \quad Z = L^{-T}, \quad V_s = LD - sL^{-T}.$$

The algorithm to get the decomposition of A uses explicitly the computed factors of A^{-1}, that is, A^{-1} is implicitly factorized at the same time. Therefore, to get the LU factorization for general matrices it is necessary to compute also the ISM decomposition of A^T that gives as result

$$\tilde{Z} = L^{-T}, \quad \text{and} \quad \tilde{V_s} = LD - sU^{-1},$$

where we have denoted with tilde the factors of the ISM decomposition of A^T.

It is well known that a nonsingular matrix A has an LU factorization if there exists a lower unit triangular matrix L and an upper triangular matrix U, such that $A = LU$. The LDU factorization is obtained from the LU factorization by taking D as the diagonal matrix whose entries are the diagonal entries of U, and applying its inverse to U as $D^{-1}U$. Both factorizations are closely related with Gaussian elimination. Note that not all the nonsingular matrices have LU factorization since a zero pivot can be found during the Gaussian elimination process. However it is always possible to permute some rows, and maybe some columns of the matrix in such a way that the permuted matrix PAQ has LU factorization. Here P and Q are permutation matrices acting on rows and columns of A, respectively.

The idea is that it is possible to find permutation matrices P and Q such that at the k-th step of the Gaussian elimination process one obtains the matrix

$$(PAQ)^{(k)} = \begin{bmatrix} L_{11} & O \\ L_{21} & I \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ O & S^{(k)} \end{bmatrix} \tag{5}$$

where the Schur complement $S^{(k)} = A_{22} - A_{21} A_{22}^{-1} A_{12}$ is nonsingular and its first diagonal element is nonzero. Then, the permuted matrix PAQ is factorized as

$$PAQ = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}. \tag{6}$$

Note that in practice, the permutation matrices P and Q are not known in advance and therefore LU factorization algorithms determine which rows and columns must be interchanged during the elimination process.

3 Right looking ISM algorithm

To implement pivoting in the ISM decomposition it is necessary to know the Schur complement of the LU factorization. To accomplish that the vectors z_k and v_k must be computed in a different way. Instead of computing only one pair of vectors in the k-th step of the algorithm according to equations (3), the modification consist in updating also the remaining vectors, from $k+1$ to n. That is, the right part of the matrices Z and V are updated in each step. The following MATLAB code implements the new right looking version of ISM.

Algorithm 1 *The ISM right looking algorithm*

```
function [Z, V, D] = ismrl(A)
n = size(A,1);
Y = (A-eye(n))';
Z = eye(n);
V = A'-eye(n);
D = zeros(n,1);
for k=1:n-1
D(k) = 1+V(k,k);
for l = k+1:n
Z(:,l) = Z(:,l) - V(l,k)/D(k)*Z(:,k);
V(:,l) = V(:,l) - (Y(:,l)'*Z(:,k))/D(k)*V(:,k);
end
end
D(n)=1+V(n,n);
```

The next results show that the Schur complement $S^{(k)}$ is available from the matrix V. We denote by $V_{22}^{(k)}$ the $(n-k)\times(n-k)$ submatrix of V in Algorithm 1 after step k, with rows and columns with indexes in $\{k+1,\ldots,n\}$.

Theorem 12. *If A is a nonsingular matrix, then at the k-th step of the right looking ISM algorithm 1*

$$V_{22}^{(k)} = S^{(k)^T} - I \tag{7}$$

Corollary 2. *If the right looking algorithm is applied to A^T then*

$$\tilde{V}_{22}^{(k)} = S^{(k)} - I$$

To introduce pivoting strategies the relation

$$V_{22}^{(k)} = S^{(k)^T} - I$$

should be taken into account. The new pivot is looked for into the submatrix $V_{22}^{(k)} + I$ that corresponds to the transpose of the same submatrix in $A^{(k)}$ in Gaussian elimination. Thus, in partial pivoting if two columns k and $p > k$ are permuted at step k in matrix $V^{(k)}$, the rows k and p should be permuted in A.

Also note that the pivoting strategy should be decided looking into the Schur complement contained in V_s, or that in $\tilde{V}_s$, but not both. In contrast, for complete pivoting it is clear that V_s or $\tilde{V}_s$ produce the same pivot in exact arithmetic so any of them or both may be used.

Table 1: Test problems

Matrix	n	nz	$cond(A)$	Application
adder_dcop_06	1,813	11,224	$1.1 \cdot 10^{12}$	circuit simulation matrix
adder_dcop_19	1,813	11,224	$5.9 \cdot 10^{11}$	circuit simulation matrix
oscil_dcop_01	1,813	11,224	$5.9 \cdot 10^{12}$	circuit simulation problem
oscil_dcop_57	1,813	11,224	$1.4 \cdot 10^{21}$	circuit simulation problem
radfr1	1,048	13,299	$5.9 \cdot 10^{10}$	chemical process separation

4 Numerical experiments

In this section we report the results of some numerical experiments with a set of matrices from The Univesity of Florida Sparse Matrix Collection [4]. The matrices are listed in Table 1 where their size, number of nonzeros, condition number and application are indicated. They correspond to very ill-conditioned and highly indefinite problems for which Gaussian elimination without pivoting fails to compute good quality L and U factors, so the same is expected to be the case for incomplete LU factorizations. Partial, rook and complete pivoting techniques have been tested. The experiments have been implemented and run in MATLAB R2022. As iterative solvers the MATLAB implementation of full GMRES [11] and BiCGStab [12] were used. The iterations were stopped when the initial residual was reduced by 8 orders of magnitude with a maximum number of $1,000$ iterations. The right hand side vector was computed such that the solution was the vector of all ones. To compare the results obtained with BIF the problems were also solved with the MATLAB's incomplete LU preconditioner with partial pivoting, ILUTP.

The implementation of the BIF preconditioner is based on the algorithm described in [9] but with the right looking modification described in Section 3. For simplicity, all the experiments have been done with the the parameter s of the ISM decomposition equal to one. The algorithm is implemented such that the ISM decompositions of A and A^T are computed at the same time. Therefore, accessing to A and A^T simultaneously is needed. The pivot is choosen from the Schur complement contained in V_s rather than $\tilde{V}_s$. We note that for complete pivoting the same pivot could be obtained working either with V_s or $\tilde{V}_s$ but we choose working with V_s for simplicity.

In Table 2 the pivoting strategy is indicated with C, P and R for the complete, partial and rook pivoting strategies, respectively. *Density* is the ratio between the number of nonzeros of the preconditioner and the matrix. Column *iter* shows the number of iterations of the solver and *droptol* is the tolerance used to drop elements in BIFP and ILUTP. The other columns are self explanatory. To reduce the numbers in the tables, a blank space means that the value is the same appearing in previous rows. For instance, in Table 2 the *droptol* value for BIFP was always 10^{-6} and therefore it appears only in the first row. The same holds for the preconditioner densities which are the same for GMRES and BiCSTAB and therefore only indicated once.

Next, we will comment on the results. We note that the matrices tested can not be solved without pivoting with both BIFP and ILUTP preconditioners. Thus, pivoting is an essential tool to gain robustness for these factorizations. From the University of Florida test matrices, Table 2, we observe for the adder group that there are not big differences between the different pivoting strategies for BIFP. Density is small, except for adder_dcop_06. The same can be said for the number of iterations spent by both iterative solvers. For the rest of matrices one can see that BIFP with complete pivoting computes sparser preconditioners than partial and rook pivoting. The iteration count does not present remarkable differences except for the oscil_dcop_01 matrix for which GMRES with partial pivoting, although with larger nonzero density, doubles the number of iterations.

Finally, comparing the performance of BIFP with ILUTP we did not observed significative differences, specially with the preconditioned BiCGStab method. We recall that ILUTP uses partial pivoting and we observe that BIFP with this pivoting strategy performed closely in most cases.

5 Conclusions

In this paper we have presented an improved version of the BIF preconditioner that incorporates pivoting. The algorithm relies on a modification of the recursion formulas such that the Schur complement of standard Gaussian elimination is available at each step of the factorization. Thus, the application of different pivoting techniques, as for instance partial, rook and complete pivoting, can be done in a straightforward maner. Incorporating pivoting turns out to be an important step in order to achieve our initial goal of obtaining a more robust preconditioner since it is able to solve very ill-conditioned and indefinite problems that it may not be possible to solve in other way. The results of the numerical experiments with several matrices arising in different applications confirm that BIF with pivoting is a robust algorithm. Partial, rook and complete pivoting has been tested. Although complete pivoting very often produces sparser preconditioners with a competitive iteration count, rook and partial pivoting perform also quite well. Taking into account that partial and rook are less expensive from a computational point of view since they need less comparisons in order to determine the pivot, these two techniques may be prefereable as default.

References

[1] M. Benzi and M. Tůma. A comparative study of sparse approximate inverse preconditioners. *Appl. Numer. Math.*, 30(2-3):305–340, 1999.

[2] M. Benzi and M. Tůma. A robust incomplete factorization preconditioner for positive definite matrices. *Numer. Linear Algebra Appl.*, 10(5-6):385–400, 2003.

[3] M. Bollhöfer. A robust and efficient *ILU* that incorporates the growth of the inverse triangular factors. *SIAM J. Sci. Comput.*, 25(1):86–103, 2003.

[4] Timothy A. Davis and Yifan Hu. The university of florida sparse matrix collection. *ACM Trans. Math. Softw.*, 38(1):1:1–1:25, December 2011.

[5] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* SIAM, Philadelphia, 2002.

[6] George Poole and Larry Neal. The rook's pivoting strategy. *Journal of Computational and Applied Mathematics*, 123:353–369, 11 2000.

[7] J. Marín R. Bru, J. Cerdán and J. Mas. Preconditioning sparse nonsymmetric linearsystems with the sherman-morrison formula. *SIAM J. Sci. Comput.*, 25(2):701–715, 2003.

[8] J. Mas R. Bru, J. Marín and M. Tuma. Balanced incomplete factorization. *SIAM J. Sci. Comput.*, 30(5):2302–2318, 2008.

[9] J. Mas R. Bru, J. Marín and M. Tuma. Improved balanced incomplete factorization. *SIAM J. Matrix Anal. Appl.*, 31(5):2431–2452, 2010.

[10] Y. Saad. ILUT: A dual threshold incomplete lu factorization. *Numer. Linear Algebra Appl.*, 1(4):387–402, 1994.

[11] Y. Saad and M. Schulz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[12] H.A. van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems. *SIAM Journal on Scientific and Statistical Computing.*, 12:631–644, 1992.

Table 2: Test results for the University of Florida matrices

Matrix	*precond*	*solver*	*droptol*	*piv*	*density*	*iter*
adder_dcop_06	BIFP	GMRES	10^{-6}	C	1.76	4
				P	1.81	3
				R	2.92	4
		BiCGStab		C		1
				P		1
				R		1
	ILUTP	GMRES	10^{-7}	P	1.67	3
		BiCGStab		P		1
adder_dcop_19	BIFP	GMRES	10^{-6}	C	0.69	3
				P	0.94	3
				R	0.72	3
		BiCGStab		C		2
				P		1
				R		2
	ILUTP	GMRES	10^{-2}	P	0.70	6
		BiCGStab		P		2
oscil_dcop_01	BIFP	GMRES	10^{-7}	C	2.06	10
				P	2.64	19
				R	2.08	11
		BiCGStab		C		1
				P		4
				R		1
	ILUTP	GMRES	10^{-9}	P	2.70	3
		BiCGStab		P		1
oscil_dcop_57	BIFP	GMRES	10^{-16}	C	2.31	11
			10^{-11}	P	2.28	28
			10^{-16}	R	2.57	11
		BiCGStab		C		1
				P		2
				R		1
	ILUTP	GMRES	10^{-16}	P	2.74	11
		BiCGStab		P		1
radfr1	BIFP	GMRES	10^{-2}	C	2.25	3
			10^{-5}	P	3.86	1
			10^{-2}	R	2.70	2
		BiCGStab		C		9
				P		3
				R		8
	ILUTP	GMRES	10^{-3}	P	2.69	2
		BiCGStab		P		10

29

Relative research contributions towards the characterization of scour in bridge piers based on operational modal analysis techniques

S. Mateo[♭,1], J. H. Alcañiz[♭], J. I. Real[♮], E. A. Colomer[♮]

(♭) Universidad Católica San Antonio de Murcia,
Av. de los Jerónimos 135, Guadalupe de Maciascoque, Murcia, Spain.
(♮) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Transportation infrastructure is one of the most important and valuable assets globally v, specially bridges since they are designed for a useful life of 100 years [2]. It is precisely its long service life the reason that causes them to suffer different pathologies throughout their life cycle, which must be identified to guarantee their structural safety.

However, despite the need for bridge auscultation, bridges continue to experience collapses around the world, such as the subway bridge in Mexico City, the Morandi viaduct (Italy), the bridge in Jaracoba (Dominican Republic), or the Ventana 2 bridge on the Coca River (Ecuador), all of which are events that have occurred in recent years [3–6].This is why infrastructure managers need to implement a predictive maintenance philosophy aimed at the pathologies affecting bridges, which would require quasi-continuous monitoring campaigns.

For bridges with piers over river channels, the main pathology responsible for bridge collapse is scour [7]. This pathology is the removal of the soil under pier and abutment foundations following a hydraulic phenomenon, or the replacement of the existing soil in the foundations by another of inferior quality (filling phenomenon), thereby causing a decrease in the stiffness of the pier-soil assembly and, therefore, decreasing the structural stability of the bridge. Its nature makes it difficult to detect since the presence of soil does not guarantee the stiffness of the pile-soil assembly. The importance of bridge scour was confirmed by the study carried out by the European Commission, which determined that around 20 % of bridges in Europe will be at high risk of scour in the next 20 years, affecting 48,000 road bridges in Europe [7, 8].

In this paper we present a bridge scour detection method based on the measurement of the accelerations at the pier heads and at the centre of the bridge span by means of an operational modal analysis. The analysis is based on the frequency domain decomposition method and the dynamic behaviour based on a finite-element model of the bridge over Leza's river near the city of Logroño. It should be noted that during the data acquisition phase, one of the bridge piers was repaired due to suffering from scour, so records are available before and after said action and will

[1] smateo0@alu.ucam.edu

be used to verify the performance of the system. We present some results and end the paper with conclusions

2 Method

This section details the methodology to diagnose bridge scour from an operational modal analysis. Figure 1 shows all the steps to be followed to obtain it from a combined random load applied to the bridge.

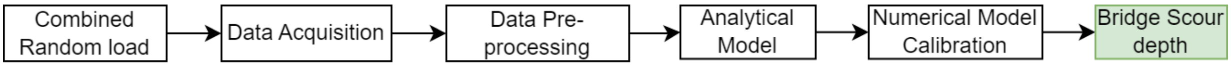


Figure 1: Methodology process scheme.

2.1 Data Acquisition System

The data acquisition system consists of wireless nodes composed of an ultra-wide bandwidth, low-noise, 3-axis digital accelerometer sensor (IIS3DWB) and other components such as the transmission module, internal storage, battery and solar panel. Additionally, to be able to transmit the data, a commercial gateway system was used (Xbee 4G). As mentioned above, these devices have been installed at the pier heads and at the centre of the bridge span.

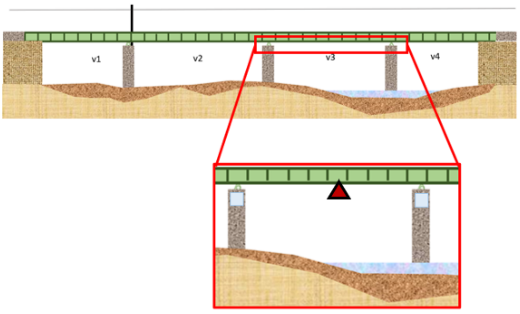


Figure 2: Data acquisition position scheme of the bridge over the Leza river.

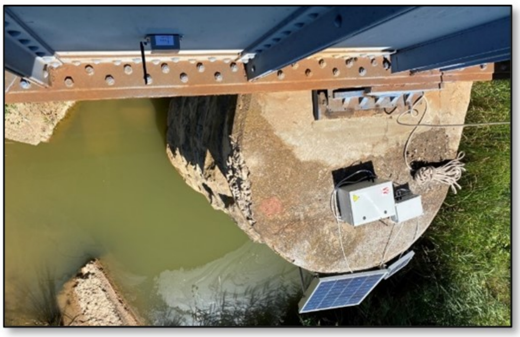


Figure 3: Hardware installed on the bridge.

2.2 Data Pre-processing

Data processing is a time-consuming and computationally expensive process. Even more so given that the information acquired can correspond to signals with low energy that can affect the accuracy of modal identification methods.

For that reason, a methodology is developed and implemented to categorise signals using a cumulative intensity function and the Rainflow-counting algorithm. The cumulative intensity of a record can be determined by Eq (1), where ü is the acceleration measured between t0 and t1. This measure considers the amplitude of the signal and its duration, achieving a more complete representation compared to, for example, taking the maximum value of the record. In addition, by analysing the acceleration acquired in different windows, it is also possible to identify the periods that contain the most interesting information for modal analysis.

$$IA_i = \alpha \int_{t_0}^{t_1} (\ddot{u}(t))^2 \, dt \tag{1}$$

The previous results are complemented with the Rainflow-counting algorithm, which determines the number of cycles and their corresponding amplitude. These values can be operated according to EQ 2, where m is the number of cycles found in window i, rn is the amplitude of cycle n and cn is the number of times the cycle is repeated.

$$RF_i = \sum_{n=1}^{m} c_n r_n \tag{2}$$

2.3 Analytical Model

The analytical model in charge of obtaining the structure's natural frequencies is based on the frequency domain decomposition method (FDD) (shown in Figure 4) which performs an approximate decomposition of the system response into a set of independent single degree of freedom (SDOF) systems, one for each mode.

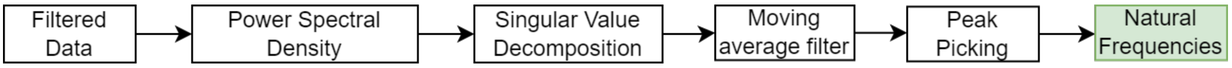


Figure 4: Methodology process scheme.

As described by Rainieri and Fabbrocino [9] the theoretical proof of the method is based on the modal expansion of the structural response:

$$y(t) = [\phi]p(t) \tag{3}$$

Where ϕ is the modal matrix and p(t) the vector of modal coordinates. From eq 3 the correlation matrix of the responses can be computed as follows:

$$[R_{yy}(\tau)] = E[\{y(t+\tau)\}\{y(t)\}^T] = [\tau][R_{pp}(\tau)][\phi]^T \tag{4}$$

The PSD matrix can be obtained from eq 4 by taking the Fourier transform:

$$[G_{YY}(\omega)] = [\phi][G_{PP}(\omega)][\phi]^H \tag{5}$$

The PSD matrix of the modal coordinates is diagonal if they are uncorrelated. A similar decomposition is obtained in the case of uncorrelated excitation forces characterised by a flat spectral density function.

Taking into account that the SVD of the PSD matrix at a certain frequency ω leads to the following factorisation:

$$[G_{YY}(\omega)] = [U][\Sigma][V]^H \tag{6}$$

Where [U] and [V] are the unitary matrices holding the left and right singular vectors and [Σ] is the matrix of singular values (arranged in descending order), for a Hermitian and positive definite matrix, such as the PSD matrix, it follows that [U]=[V] and the decomposition of eq 6 can be rewritten as:

$$[G_{YY}(\omega)] = [U][\Sigma][U]^H \tag{7}$$

The comparison between eq 5 and eq 7 suggests that it is possible to identify a one-to-one relationship between singular vectors and mode shapes; moreover, the singular values are related to the modal responses and they can be used to define the spectra of equivalent Single Degree of Freedom (SDOF) systems characterized by the same modal parameters as the modes contributing to the response of the MDOF system under investigation. Since the SVD provides the singular values arranged in descending order, near a resonance the first singular value contains the information about the dominant mode at that frequency. Moreover, since the number of nonzero elements in [Σ] equals the rank of the PSD matrix at the considered frequency, this property can be used to identify closely spaced or even coincident modes. In fact, the number of dominant singular values (defining the rank of the output PSD matrix) at a certain frequency equals the number of modes that give a significant contribution to the structural response at that particular frequency. Assuming that only one mode is dominant at the frequency ω, and that the selected frequency is associated to the peak of resonance of the k-th mode, the PSD matrix approximates to a rank one matrix with only one term on the right side of eq 7:

$$[G_{YY}(\omega)] = \sigma_1\{u_1\}\{u_1\}^H, \quad \omega \to \omega_k \tag{8}$$

In such a case, the first singular vector u_1 represents an estimate of the mode shape of the k-th mode:

$$\{\hat{\phi}_k\} = \{u_1(\omega_k)\} \tag{9}$$

and the corresponding singular value σ_1 belongs to the auto PSD function of the equivalent SDOF system corresponding to the mode of interest. The equivalent SDOF PSD function is identified as the set of singular values around a peak of the singular value plots that are characterized by similar singular vectors.

2.4 Numerical Model

Operational Modal Analysis is a process that allows the dynamic properties of the structure to be determined from its natural frequencies, damping ratios and vibration modes. In this way, it makes it possible to obtain a mathematical model that is a faithful representation of the real structure, on which different types of verifications can be carried out to find out the current state of the bridge, or even simulations of highly demanding scenarios for the structure and predict its behaviour.

The differential equation governing the linear, time-invariant, multi-degree-of-freedom system, in which both the excitation and response of the structure are stationary, random processes, is shown in equation 10:

$$[M]\{\ddot{X}\} + [C][\{\dot{X}\}][K][\{X\}] = \{F(t)\} \tag{10}$$

Where [M], [C] and [K] are the mass, damping and stiffness matrices, respectively; X is the vector of displacements in each degree of freedom of the system and F(t) is the external excitation.

Once the numerical model has been set up, it can be calibrated by matching the natural frequencies obtained previously with the analytical model. This is done by modifying the properties of the materials and the stiffness of the soil iteratively. The model validation is subsequently performed on the basis of the deformation under self-weight and the damping ratio measured at the centre of the bridge span.

3 Results

Once all the equipment has been fully installed on the bridge and the equipment has started to operate, the data obtained from each of the sensors and their respective frequency registers can be observed by applying the mathematical model. A series of records obtained by the system at each of the sensor nodes of the bridge are shown below.

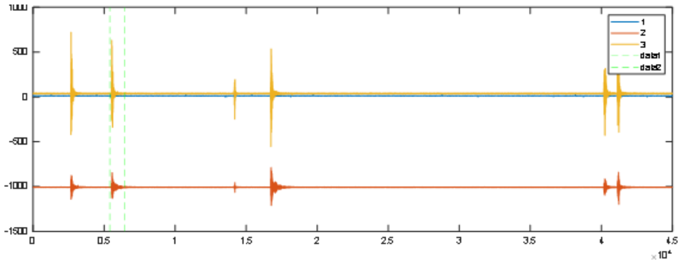


Figure 5: Accelerometer registry at the centre of the bridge span. Each colour represents an axis.

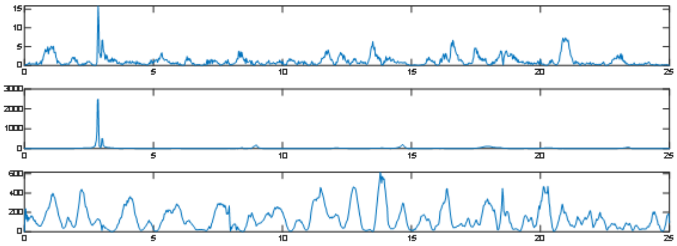


Figure 6: Spectral frequency in each axis at the centre of the bridge span.

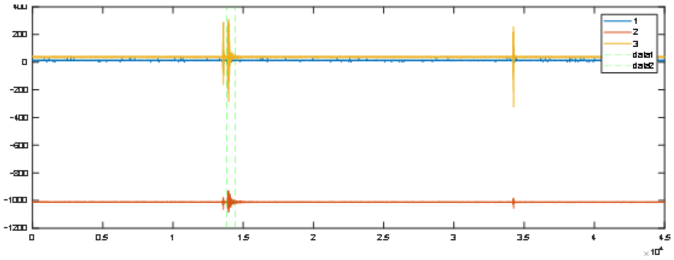


Figure 7: Accelerometer registry at the pier head. Each colour represents an axis.

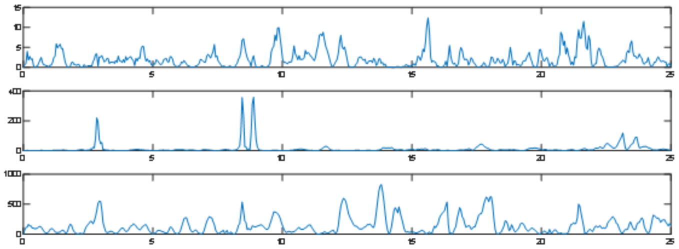


Figure 8: Spectral frequency in each axis at the pier head.

Figure 9 shows the frequency associated with the first singular value obtained after performing singular value decomposition of the spectral density matrices. While Figure 10 shows the maximum frequency after applying the peak-picking algorithm.

4 Conclusions

The data acquisition system has been developed and installed on the bridge over the river Leza is capable of recording and transmitting data continuously and in real time. It has also been

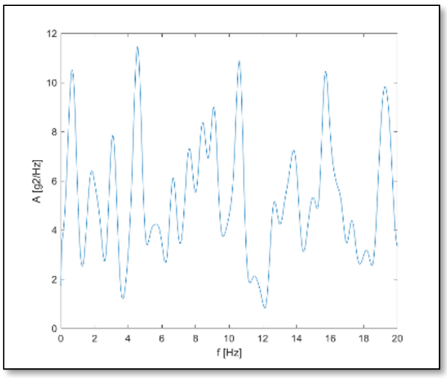


Figure 9: First singular value of all spectral density matrices.

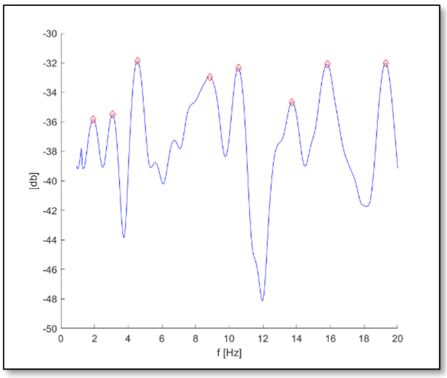


Figure 10: Maximum frequencies after applying the peak-picking algorithm.

verified that both the pre-processing algorithms and the analytical model to obtain the vibration frequencies of the pier are working successfully. Although future work will focus on specifying the best combination of spectral density matrices (i.e., how many registries to employ and from which sensor) to be used for the singular value decomposition method. Additionally, the authors will work on the calibration of the numerical model and the comparison between the results before and after the scour reparation action.

References

[1] AEC. Sector viario y crisis. Análisis de situación. Catálogo de propuestas. 2010.

[2] Ministerio de Fomento. IAP-11. Instrucción sobre las acciones a considerar en el proyecto de puentes de carreteras. Segunda. Madrid; 2012.

[3] Rodríguez D. Accidente Línea 12: Al menos 25 muertos por el derrumbe de un tramo del metro de Ciudad de México [Internet]. El País México. 2021 [cited 2022 Oct 27]. Available from: https://elpais.com/mexico/2021-05-04/un-accidente-en-la-linea-12-de-metro-de-ciudad-de-mexico-deja-al-menos-50-heridos.html

[4] Delle L. Génova: El viaducto Morandi, una “obra maestra” que resultó letal — Internacional [Internet]. El País. 2018 [cited 2022 Oct 27]. Available from: https://elpais.com/internacional/2018/08/14/actualidad/1534271879_092079.html

[5] Fulcar R. Lluvias provocan colapso de un puente en Jarabacoa [Internet]. El Nuevo Diario. 2021 [cited 2022 Oct 27]. Available from: https://elnuevodiario.com.do/lluvias-provocan-colapso-de-un-puente-en-jarabacoa/

[6] Coba G. Tramo del puente Ventana 2 colapsa por erosión regresiva del río Coca [Internet]. Primicias. 2021 [cited 2022 Oct 27]. Available from: https://www.primicias.ec/noticias/economia/puente-acceso-ventana-colapsa-erosion-coca/

[7] Brady K, O’Reilly M, Bevc L, Znidaric A, O’Brien E, Jordan R. Cost 345. Procedures required for assessing highway structures. Final report [Internet]. 2003 [cited 2022 Oct 27]. Available from: http://www.cordis.lu/cost-transport/home.html

[8] Nemry F., Hande D., Impacts of climate change on transport: a focus on road and rail transport infrastructures. *Publications Office*, 2012.

[9] Rainieri C., Fabbrocino G., Operational Modal Analysis of Civil Engineering Structures. An Introduction and Guide for Applications. 127-–133, 2014. .

Short-term happiness dynamics as a consequence of an alcohol or caffeine intake

Salvador Amigó[*], Antonio Caselles[♭], Joan C. Micó[◇] and David Soler[◇,1]

(*) Departament de Personalitat, Avaluació i Tractaments Psicològics.
Universitat de València,
Av. Blasco Ibáñez 21, 46010. València, Spain,
Salvador.Amigo@uv.es.
(♭) IASCYS member, Departament de Matemàtica Aplicada,
Universitat de València (retired),
Dr. Moliner 50, 46100 Burjassot, Spain,
Antonio.Caselles@uv.es.
(◇) I. U. de Matemàtica Multidisciplinar,
Universitat Politècnica de València,
Camí de Vera s/n, 46022, València, Spain,
jmico@mat.upv.es, dsoler@mat.upv.es

1 Introduction

There exists a scientific debate about whether happiness is a trait or a state (see [1] for a review). However, some authors go beyond by an inclusive proposal about the happiness nature: it is both a trait and a state [2]. We adopt that approach in this study.

If happiness has a state nature as well as a trait one, we will be able to study its dynamics as both short and long term, even as a result of a unique eliciting stimulus in a single session, as it can be a drug intake. For instance, it has been demonstrated that, after a single dose intake, both alcohol and caffeine can increase happiness and feelings such as euphoria in the short term [3-5]. On the other hand, the existence of individual differences inside the acute effects of both drugs has been proved [6-8]. But its short-term dynamics has not been well described yet.

There exists a mathematical dynamical model to predict and describe how the whole personality changes (The General Factor of Personality or GFP) during a single session in response a single dose of caffeine or alcohol and how the responses vary between individuals [9, 10], but it has not been applied to the study of happiness yet. In this study, we present a dynamical model to predict the evolution of a subject's happiness in response a single dose of alcohol or caffeine.

The model is provided by the following integrodifferential equation:

$$\begin{aligned} &\dot{q}\left(t\right) = a\left(b - q\left(t\right)\right) + \frac{\delta}{M}s(t)q(t) - \frac{\gamma}{M}\int_{t_0}^{t}\exp\left(\frac{r-t}{\tau}\right)s(r)\cdot q(r)\,dr \\ &q\left(t_0\right) = q_0 \end{aligned} \tag{1}$$

[1] dsoler@mat.upv.es

In Eq. (1), $q(t)$ represents the GFP dynamics; and b and q_0 are respectively its trait level and its initial value. Its dynamics is a balance of three terms, which provide the time derivative of the GFP: the homeostatic control $(a(b - q(t)))$, i.e., the cause of the fast recovering of the tonic level b, the excitation effect $((\delta/M)s(t)q(t))$, which tends to increase the GFP per drug unit, and the inhibitor effect $((\gamma/M)\mathring{u}\int_0^t \exp\frac{r-t}{\tau}s(r)q(r)dr)$, which tends to decrease the GFP per drug unit and is the cause of a continuous delayed recovering, being M the amount of drug intake. Parameters α, δ, γ and τ are named respectively the homeostatic control power, the excitation effect power, the inhibitor effect power and the inhibitor effect delay. In addition, $s(t)$ provides the dynamics of the stimulus by the drug kinetics:

$$s\left(t\right) = s_0 \exp(-\beta t) = \begin{cases} \frac{\alpha M}{\beta - \alpha}\left(\exp(-\alpha t) - \exp(-\beta t)\right) & \alpha \neq \beta \\ \alpha M t \exp(-\alpha t) & \alpha = \beta \end{cases} \tag{2}$$

In Eq. (2) α is the drug assimilation rate and β is the drug elimination rate, being again M the amount of drug intake.

The model given by eqs, 1 and 2 has been applied in a study with two participants with a different level of the happiness trait, by using a Trait-State Scale of Happiness previously validated in [11]. This scale was based on the Euphoria Scale [12], and it has been proved that this scale is closely related with the Oxford Happiness Inventory (short version) [13] and how this scale is sensitive to the changes produced by eliciting stimulus [14]. We also use a Smiling Face Scale. Both scales will be described below.

2 Methodology

Two voluntary men participated in this study. A single-case experimental ABC design was used. In phase A the participants received no treatment. At the start of phase B, both participants received 26.51 ml of alcohol and a slight food. In phase C, both participants received 330 mg of caffeine. Two instruments to evaluate happiness were used: 1) The trait-State Scale of Happiness [11] in its trait-format ("*Are you like this in general?*") and its state-format ("*Are you like this at this moment?*" or "do you feel so at this moment?"). It is a 4-item Likert-type response scale with the following self-descriptive adjectives: cheerful, elated, exhilarated, and lively. The scale score goes from 0 (no effect) to 5 (maximum effect); 2) The Smiling Face Scale, that is a 7-item Likert-type response scale, that shows images with very sad to very happy faces, so ranging from negative to neutral to positive values [15, 16]. Both participants filled in the Euphoria Scale in its trait format at the very beginning of this study, and every 5 minutes over a 1,5-hour period, all the three phases long. For the mathematical analysis, the modified response model was applied, whose usefulness has been shown to model the dynamic effect of both drugs. Figures 1 to 4 show the data (Expression) jointly with the predicted by the model response curves (Happiness).

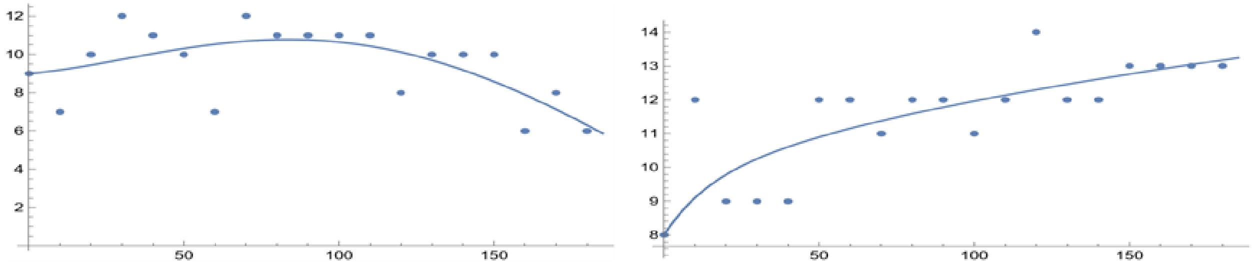


Figure 1: Participant 1: (Left) Happiness response to alcohol dose (R2=.43); (Right) Happiness response to caffeine dose (R2=.45).

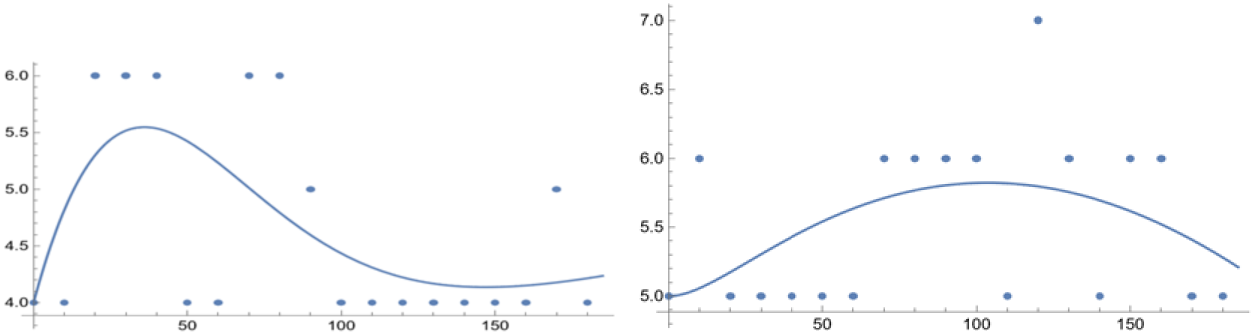


Figure 2: Participant 1: (Left) Expression response to alcohol dose (R2=.36); (Right) Expression response to caffeine dose (R2=.16).

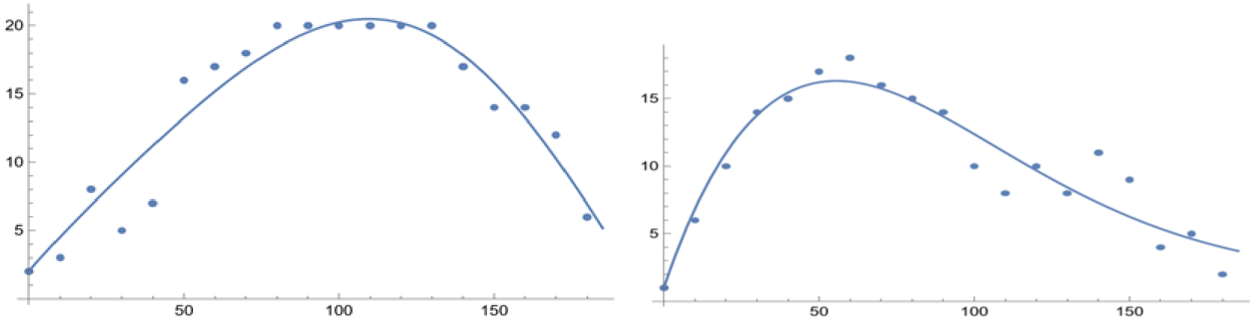


Figure 3: Participant 2 (Left) Happiness response to alcohol dose (R2=.90); (Right) Happiness response to caffeine dose (R2=.87).

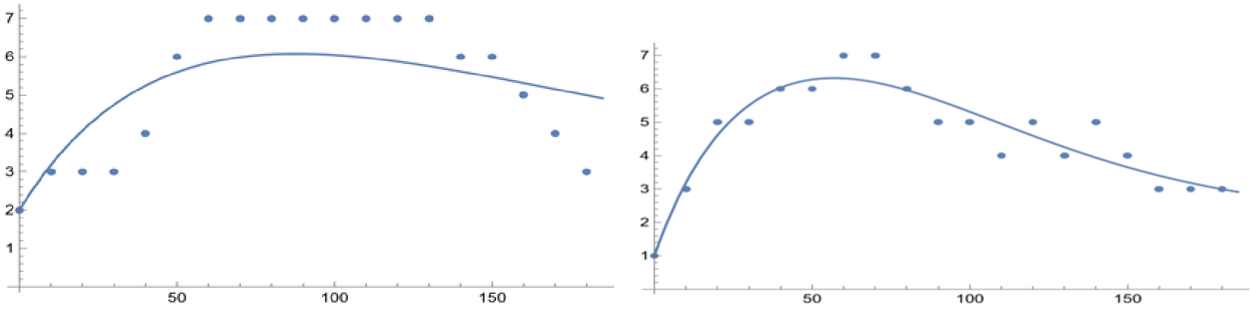


Figure 4: Participant 2: (Left) Expression response to alcohol dose (R2=.74); (Right) Expression response to caffeine dose (R2=.83).

3 Conclusions

The difference between both participants in the experiment is evident: Participant 1 presents more dispersion than Participant 2. In addition, the dynamic model presented here adapts better to Participant 2. However, the results provide random residuals in both cases, and so the model represents the deterministic predictive part of the dynamical responses.

As other studies pointed out [9, 10], a model describe and predict how the GFP changes in response to a single dose of caffeine or alcohol. This mathematical model predicts that the lower the GFP-trait score is the higher the response to both caffeine and alcohol intake will be, and better the corresponding model evolution curve fits the data. That is precisely what happens in the present study, so the participant 2 scores lower in the GFP-trait (17 points) than participant 1 (23 points), and that is because his response is higher, and the model fits the experimental data better. This fact is coherent with the mathematical model prediction [17], and with the fact that GFP and Happiness scores are closely related [11, 14].

Regarding the Happiness scores, the relationship between the trait and the state evolution after a single dose intake of caffeine or alcohol is not the same as regarding the GFP. So, the difference between both participants is not as high as regarding the FGP (10 and 14 points respectively for the Trait Scale of Happiness, and the same score (5) for the Face Scale). So that, this discrepancy should be studied in the future.

Besides, these results are consistent with the ones obtained for the FGP in the sense that the lower the initial score is, the higher level will be achieved during the response to the drug intake. Likewise, the same happens regarding Happiness as this study reveals. So, the initial scores for participant 1 are 9 and 4 for the State Scale of Happiness and for the Face Scale, respectively, while for participant 2 they are 3 for both scales This result is coherent with the model prediction for the personality responses to drug intakes [17].

Note however, that this study is a first approach to the relationship between happiness and drug consumption, taking into account that this subject can provide an upset social discussion due to somebody can understand that this paper suggests consuming drugs to reach short periods of happiness, while its objective is the opposite: preventing consumers that drug consumption must be done rationally. From this result, in a future research, the objective could be to relate happiness with personality dynamics, such as it has been already with the General Factor of Personality dynamics [18].

References

[1] Diener, E.; Larsen, R.J. y Emmons, R.A. (1984), Person x situation interactions: Choice of situations and congruence response models. Journal of Personality and Social Psychology, 47, 580-592.

[2] Stones, M.J.; Hadjistavopoulos, T.; Tuuko, H. y Kozma, A. (1995), Happiness has traitlike and statelike properties: a reply to Veenhoven. Social Indicators Research, 36, 129-144.

[3] Ben Baumberg Geiger, George MacKerron. Can alcohol make you happy? A subjective wellbeing approach. Social Science & Medicine, 2016; 156: 184 DOI: 10.1016/j.socscimed.2016.03.034.

[4] Warburton DM. Effects of caffeine on cognition and mood without caffeine abstinence. Psychopharmacology (Berl). 1995 May;119(1): 66-70.

[5] Childs E, de Wit H. Subjective, behavioral, and physiological effects of acute caffeine in light, nondependent caffeine users. Psychopharmacology (Berl). 2006;185: 514–523.

[6] Hammersley, R., Finnigan, F. and Millar, K. (1994). Individual differences in the acute response to alcohol. Personality and Individual Differences, 17(4), 497–510.

[7] Yang, A., Palmer, A. A., and de Wit, H. (2010). Genetics of caffeine consumption and responses to caffeine. Psychopharmacology, 211(3), 245–257.

[8] Wit, H. d. (2005). Relationships Between Personality and Acute Subjective Responses to Stimulant Drugs. In M. Earleywine (Ed.), Mind-altering drugs: The science of subjective experience (pp. 258–274). Oxford University Press.

[9] Caselles, A., Micó, J.C. y Amigó, S. (2011). Dynamics of the General Factor of Personality in response to a single dose of caffeine. Spanish Journal of Psychology, 14, 675-692.

[10] Amigó S, Caselles A, Micó JC, Sanz MT, Soler D. Dynamics of the general factor of personality: A predictor mathematical tool of alcohol misuse. Mathematical Methods in the Applied Sciences 2020, 1–20.

[11] Amigó, S. y Hernández, N.E. (2012). Factor general de personalidad y felicidad: un estudio desde la perspectiva rasgo-estado en una muestra colombiana. Pensando Psicología, 8, 39-49.

[12] Kjellberg, A. y Bohlin, G. (1974), “Self-reported arousal: further development of a multifactorial inventory”, en Scandinavian Journal of Psychology, vol. 15, pp. 285-292.

[13] Hills, P. y Argyle, M. (1998), (2002), “The Oxford Happiness Questionnaire: a compact scale for the measurement of psychological well-being”, en Personality and Individual Differences, vol. 33, pp. 1073-1082.

[14] Amigó, S. (2014). Drugs, self-control and happiness. ACTAS del 9th Congress of the EUS-UES Globalization and Crisis. Complexity and governance of systems. Celebrado en Valencia del 15 al 17 de octubre de 2014. (pp. 273-280).

[15] Reynolds-Keefer, L., Johnson, R., Dickenson, T. and McFadden, L. Validity issues in the use of pictorial Likert scales. Studies in Learning, Evaluation Innovation and Development 6 (2009) 15–24.

[16] Hall, Lynne & Hume, Colette & Tazzyman, Sarah. (2016). Five Degrees of Happiness: Effective Smiley Face Likert Scales for Evaluating with Children. 311-321.

[17] Amigó, S., Caselles, A. and Micó, J.C. (2008). A dynamical model of extraversion. British Journal of Mathematical and Statistical Pychology, 61, 211-231.

[18] Caselles, A., Micó, J.C. and Amigó, S. (2021). Energy and Personality: A Bridge between Physics and Psychology. Mathematics, 9, 1331.

First Order Hamiltonian Systems

Joan C. Micó$^{\diamond,1}$
($\diamond$) I. U. de Matemàtica Multidisciplinar,
Universitat Politècnica de València,
Camí de Vera s/n, 46022, València, Spain.

1 Introduction

This paper presents a way to get a Hamiltonian function for a coupled first order differential equation system whose independent variable is time or, as it is called shortly in the literature about the subject, a dynamical system. In addition, the Hamiltonian is also first order in momenta, oppositely to physics dynamics, where the Hamiltonian is second order respect the momenta (note that the corresponding dynamical systems in physics are, by the Newton laws of Nature, coupled second order differential equation systems). Besides, while the temporal variables involved in physics are of spatial nature, in the dynamical systems here considered the temporal variables involved are abstract, i.e., they are of arbitrary nature, such as, for instance, biological populations, chemical components, or any other variables related with social or behavioural nature. The way used to get the Hamiltonian is that provided by Dirac by his Generalized Hamiltonian Dynamics method [1]. This method is applied for those systems for which the Hamiltonian cannot be provided from the Lagrangian function for two cases: either for those where the generalized velocities cannot be isolated as a function of the momenta, or for those with singular Hamiltonians, for which the momenta vanish. The second case is the corresponding to this paper. This method was developed by Dirac for fields in order to get the Hamiltonian of the electromagnetic field, which was also singular. Note that here, at difference of fields, which are infinite-dimensional systems, the system dimension is finite and equal to the number of the first order differential equations involved in the model studied. There are other approaches different to Dirac's method, such as Hava's approach [2] or Pontryagin's approach [3]. Hava's approach [2] focuses on the Lagrangian to discuss possible conservation laws, and Pontryagin's approach [3] uses a different Hamiltonian way steered to optimize dynamical functions, although Dirac's and Pontryagin's results are mathematically similar but with different objectives. In fact, the real Dirac's objective is to get the Schrödinger equation from the Hamiltonian. Different authors have followed Pontryagin's method [3] to get the same quantum approach objective [4, 5] in other contexts. However, this paper objective does not consider the quantum approach, which is so considered in [6, 7]. In addition, the equivalence between Hava's and Dirac's approach is addressed in [8] and also more recently in [6, 7]. Section 2 is devoted to get the Hamiltonian following Dirac's approach [1]. Section 3 is devoted to use the formalism presented to get a General System Thermodynamics (GST), while Section 4 points out how this formalism could reproduce the classical reversible Thermodynamics as a particular case of the GST. Section 5 is devoted to the paper discussion and the possible future applications of the formalism.

[1] jmico@mat.upv.es

2 Getting the Hamiltonian

Let $q_k\left(t\right)$, $k = 1, 2, \ldots, n$, the abstract variables of a dynamical system, with $\mathbf{q} = \left(q_1, q_2, \ldots,\ q_n\right)$:

$$\dot{q}_k\left(t\right) = f_k\left(t, \mathbf{q}\right); \qquad k = 1, 2, \ldots, n. \tag{1}$$

The problem of minimum action principle consists in finding an integral action $\Lambda\left(t\right)$:

$$\Lambda\left(t\right) = \int_{t_1}^{t_2} L\left(t, \mathbf{q}, \dot{\mathbf{q}}\right)\ dt \tag{2}$$

with arbitrary t_1 and t_2 times, being $L\left(t, \mathbf{q}, \dot{\mathbf{q}}\right)$ the Lagrangian, such that $\Lambda\left(t\right)$ be a minimum, i.e., $\delta\Lambda\left(t\right) = 0$. This process of minimization provides the Euler-Lagrange equations:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = 0; \qquad k = 1, 2, \ldots, n. \tag{3}$$

Eq. (3) must reproduce the system of Eq. 1. This problem is known in the scientific literature as the Lagrange inverse problem [2].

As Dirac [1] and Havas [2] argue, the only possibility to solve the inverse Lagrange problem for Eq. 1 is as a linear combination for the Lagrangian such as the following:

$$L\left(t, \mathbf{q}, \dot{\mathbf{q}}\right) = \sum_{j=1}^{n} g_j\left(t, \mathbf{q}\right)\ \dot{q}_j - h\left(t, \mathbf{q}\right). \tag{4}$$

In Eq. (4) the $g_j\left(t, \mathbf{q}\right)$ and $h\left(t, \mathbf{q}\right)$ functions are unknown functions by the moment. Note that the momenta p_k can be defined from Eq. (4) as:

$$p_k = \frac{\partial L}{\partial \dot{q}_k} = g_k\left(t, \mathbf{q}\right) \tag{5}$$

As it is well known, the Hamiltonian $\left(H_0\left(t, \mathbf{q}, \mathbf{p}\right)\right)$ is defined from the Lagrangian as:

$$H_0\left(t, \mathbf{q}, \mathbf{p}\right) = \sum_{j=1}^{n} p_j \dot{q}_j - L\left(t, \mathbf{q}, \dot{\mathbf{q}}\right) = h\left(t, \mathbf{q}\right) \tag{6}$$

Note that the Hamiltonian $H_0\left(t, \mathbf{q}, \mathbf{p}\right)$ of Eq. 6 is singular due to it does not depend on the momenta. Dirac's method [1] solves the problem by the steps described in the beginning. The first step is defining the primary constraints $\phi_j\left(t, \mathbf{q}, \mathbf{p}\right)$ from Eq. 5:

$$\phi_j\left(t, \mathbf{q}, \mathbf{p}\right) = p_j - g_j\left(t, \mathbf{q}\right) = 0; \qquad\qquad j = 1, 2, \ldots, n \tag{7}$$

that are added to the Hamiltonian Eq. 6 by the $\lambda_j\left(t, \mathbf{q}, \mathbf{p}\right)$ unknown multiplying functions as:

$$H\left(t, \mathbf{q}, \mathbf{p}\right) = H_0\left(t, \mathbf{q}, \mathbf{p}\right) + \sum_{j=1}^{n} \lambda_j\left(t, \mathbf{q}, \mathbf{p}\right)\ \phi_j\left(t, \mathbf{q}, \mathbf{p}\right) = h\left(t, \mathbf{q}\right) + \sum_{j=1}^{n} \lambda_j\left(t, \mathbf{q}, \mathbf{p}\right)\ \phi_j\left(t, \mathbf{q}, \mathbf{p}\right) \tag{8}$$

Note that the primary constraints are non-zero valued inside the Hamiltonian. By considering Eq. 7, the Hamilton equations for the new Hamiltonian $H\left(t, \mathbf{q}, \mathbf{p}\right)$ are:

$$\left.\begin{array}{l} \dot{q}_k = \frac{\partial H(t,\mathbf{q},\mathbf{p})}{\partial p_k} = \lambda_k(t, \mathbf{q}, p) \\ \dot{p}_k = -\frac{\partial H(t,\mathbf{q},\mathbf{p})}{\partial q_k} = -\frac{\partial h(t,\mathbf{q})}{\partial q_k} + j = \sum_{j=1}^{n} \lambda_j(t, \mathbf{q}, \mathbf{p}) \frac{\partial g_j(t,q)}{\partial q_k} \end{array}\right\}; \qquad k = 1, 2, \ldots, n \tag{9}$$

The first result obtained by comparing Eqs. 1 and 9 is that:

$$\lambda_k(t,\mathbf{q},\mathbf{p}) = f_k(t,\mathbf{q}); \qquad k = 1,2,\ldots,n \tag{10}$$

Note that in Eq. 9 the zero value of the primary constraints outside the Hamiltonian has been used. In order to get the $g_j(t,\mathbf{q})$ and $h(t,\mathbf{q})$ functions of the formalism, the functions primary constants $\phi_j(t,\mathbf{q},\mathbf{p})$ must hold the consistency conditions, i.e., $\dot{\phi}_l = 0(l = 1,2,\ldots,n)$, that is:

$$\dot{\phi}_l(t,\mathbf{q},\mathbf{p}) = \frac{\partial\phi_l(t,\mathbf{q},\mathbf{p})}{\partial t} + \sum_{k=1}^{n}\frac{\partial\phi_l(t,\mathbf{q},\mathbf{p})}{\partial q_k}\dot{q}_k + \sum_{k=1}^{n}\frac{\partial\phi_l(t,\mathbf{q},\mathbf{p})}{\partial p_k}\dot{p}_k = 0; \qquad l = 1,2,\ldots,n \tag{11}$$

The substitution of Eqs. 9 and 10 in Eq. 11 provides, taking into account the zero value of the primary constraints, and after some calculations:

$$-\frac{\partial g_l(t,\mathbf{q})}{\partial t} - \frac{\partial h(t,\mathbf{q})}{\partial q_l} + \sum_{k=1}^{n}\left(-\frac{\partial g_l(t,\mathbf{q})}{\partial q_k} + \frac{\partial g_k(t,\mathbf{q})}{\partial q_l}\right) f_k(t,\mathbf{q}) = 0; \qquad l = 1,2,\ldots,n \tag{12}$$

In Eq. 12 the following $F_{kl}(t,\mathbf{q})$ functions can be defined:

$$F_{kl}(t,\mathbf{q}) = -F_{lk}(t,\mathbf{q}) = -\frac{\partial g_l(t,\mathbf{q})}{\partial q_k} + \frac{\partial g_k(t,\mathbf{q})}{\partial q_l}; \qquad k,l = 1,2,\ldots,n \tag{13}$$

Eq. 12 can be rewritten by using Eq. 13:

$$\sum_{k=1}^{n} F_{lk}(t,\mathbf{q})\ f_k(t,\mathbf{q}) = -\frac{\partial g_l(t,\mathbf{q})}{\partial t} - \frac{\partial h(t,\mathbf{q})}{\partial q_l}; \qquad l = 1,2,\ldots,n \tag{14}$$

In order to get an equation for the $F_{lk}(t,\mathbf{q})$ functions without the $h(t,\mathbf{q})$ function, such as in Eq. 14, the following steps are followed: 1. Take the derivative respect an arbitrary q_j in Eq. 14; 2. Rewrite Eq. 14 by replacing l by j; 3. Take the derivative respect q_j in the rewritten equation; 4. Subtract both equations. Taking into account the equality of both $h(t,\mathbf{q})$ crossed derivatives, the result is:

$$\frac{\partial F_{jl}(t,\mathbf{q})}{\partial t} = \frac{\partial}{\partial q_j}\left(\sum_{k=1}^{n} F_{lk}(t,\mathbf{q})\ f_k(t,\mathbf{q})\right) - \frac{\partial}{\partial q_l}\left(F_{jk}(t,\mathbf{q})\ f_k(t,\mathbf{q})\right); \qquad j,l = 1,2,\ldots,n \tag{15}$$

Therefore, the process to get the $g_j(t,\mathbf{q})$ and $h(t,\mathbf{q})$ functions of the Hamiltonian given by Eq. 8 is: 1. Get the $F_{lk}(t,\mathbf{q})$ functions by Eq. 15; 2. Substitute these results in Eq. 13 to get the $g_j(t,\mathbf{q})$ functions; 3. Substitute these results in Eq. 14 to get the $h(t,\mathbf{q})$ function. Take into account in this process that $F_{ll}(t,\mathbf{q}) = 0$ and that $F_{kl}(t,\mathbf{q}) = -F_{lk}(t,\mathbf{q})$.

However, two different classes of solutions must be considered depending on the system dimension n. This is due to the antisymmetric definition of the $F_{lk}(t,\mathbf{q})$ functions. On the one hand, if n is even, then, in general $\det(F_{lk}(t,\mathbf{q})) \neq 0$. In this case Eqs. 14 or 15 are independent. On the other hand, if n is odd, then always $\det(F_{lk}(t,\mathbf{q})) = 0$, and being $\det(F_{lk}(t,\mathbf{q})) \neq 0$ for the $n-1$ even dimension, one of the Eq. 14 is dependent on the others, which makes that some $F_{lk}(t,\mathbf{q})$ functions become undetermined parameters from which the rest ones depend on.

Actually, this last case always happens due to Eq. 14 is a coupled set of n equations and $n+1$ unknown variables: $g_l(t,\mathbf{q})$ $(l = 1,2,\ldots,n)$ and $h(t,\mathbf{q})$. For instance, let the special one-dimensional $(n = 1)$ odd case be. The consistence conditions that provide Eq. 14 become:

$$\eta(t,q) := \frac{\partial g(t,q)}{\partial t} + \frac{\partial h(t,q)}{\partial q} = 0 \tag{16}$$

Note that Eq. 16 does not provide the $\lambda(t,q,p) = f(t,q)$ multiplying function. When this case happens, the Dirac's method [1] prescription is to consider the equations such as Eq. 16 as secondary constraints. In order to get the multiplying function in an equation, the time derivative is taken in Eq. 16:

$$\dot{\eta}(t,q) = \frac{\partial \eta(t,q,p)}{\partial t} + \frac{\partial \eta(t,q,p)}{\partial q}\dot{q} + \frac{\partial \eta(t,q,p)}{\partial p}\dot{p} = 0 \tag{17}$$

Taking into account the Hamilton equations Eq. 9, as for the primary constraints, Eq. 17 becomes, after some calculations:

$$\frac{\partial^2 g(t,q)}{\partial t^2} + \frac{\partial^2 h(t,q)}{\partial t\, \partial q} + f(t,q)\left(\frac{\partial^2 g(t,q)}{\partial q\, \partial t} + \frac{\partial^2 h(t,q)}{\partial q^2}\right) = 0 \tag{18}$$

Note in Eq. 18 that just an equation is provided for two unknown variables, $g(t,q)$ and $h(t,q)$, then one of the two variables become undetermined.

Now, with all the formalism background developed, the $g_j(t,\mathbf{q})$ $(j = 1,2,\ldots,n)$ and $h(t,\mathbf{q})$ functions can be found, and the Hamiltonian can be written from Eqs. 7 and 8, as:

$$\begin{aligned} H(t,\mathbf{q},\mathbf{p}) &= \sum_{j=1}^{n} f_j(t,\mathbf{q})\,(p_j - g_j(t,\mathbf{q})) + h(t,\mathbf{q}) = \\ &= \sum_{j=1}^{n} f_j(t,\mathbf{q}) \cdot p_j - \sum_{j=1}^{n} f_j(t,\mathbf{q}) \cdot g_j(t,\mathbf{q}) + h(t,\mathbf{q}) \end{aligned} \tag{19}$$

Observe in addition that, if the dynamical model Eq. 1 is autonomous, thus $f_k(t,\mathbf{q}) = f_k(\mathbf{q})$ and both $g_j(t,\mathbf{q}) = g_j(\mathbf{q})$ $(j = 1,2,\ldots,n)$ and $h(t,\mathbf{q}) = h(\mathbf{q})$ can be found as time independent functions, thus the Hamiltonian $H(t,\mathbf{q},\mathbf{p}) = H(\mathbf{q},\mathbf{p})$ and it is a constant of the dynamics, then it can be identified with the system energy E, that is:

$$E = \sum_{j=1}^{n} f_j(\mathbf{q})\,(p_j - g_j(\mathbf{q})) + h(\mathbf{q}) = \sum_{j=1}^{n} f_j(\mathbf{q}) \cdot p_j - \sum_{j=1}^{n} f_j(\mathbf{q}) \cdot g_j(\mathbf{q}) + h(\mathbf{q}) \tag{20}$$

3 A proposal of a General System Thermodynamics (GST)

A first theoretical application of the Hamiltonian formalism developed is to state a general or abstract theory of systems, called here as a System General Thermodynamics (GST), by introducing some new postulates. The **First Postulate** is: **identify the Hamiltonian of Eq. 19 as the system Internal Energy**. The Hamilton equations will correspond to the state equations. Note that time is indispensable in this formalism, unlikely to the classical quasi-static evolution of Thermodynamics [9]. The **Second Postulate** is: **assume the two following equations**:

$$g_k(t,\mathbf{q}) = \frac{\partial \chi(t,\mathbf{q})}{\partial q_k}; \qquad k = 1,2,\ldots,n \tag{21}$$

$$h(t,\mathbf{q}) = -\frac{\partial \chi(t,\mathbf{q})}{\partial t} \tag{22}$$

From Eqs. 21 and 22, Eq. 14 hold identically (even for the case $n = 1$ given by Eq. 16), being $\chi(t,\mathbf{q})$ an arbitrary function, and Eq. 15 and 18 are unnecessary to compute $g_k(t,\mathbf{q})$ and $h(t,\mathbf{q})$. Therefore, the primary constraints of Eq. 7 become:

$$\bar{\phi}_j(t,\mathbf{q},\mathbf{p}) = p_j - \frac{\partial \chi(t,\mathbf{q})}{\partial q_j} = 0; \qquad j = 1,2,\ldots,n \tag{23}$$

and the Hamiltonian (Eq. 19) becomes:

$$H\left(t,\mathbf{q},\mathbf{p}\right)\sum_{j=1}^{n} f_j(t,\mathbf{q})\cdot\bar{\phi}_j(t,\mathbf{q},\mathbf{p})-\frac{\partial\chi(t,q)}{\partial t}=\sum_{j=1}^{n} f_j(t,\mathbf{q})\cdot p_j-\sum_{j=1}^{n}\frac{\partial\chi(t,\mathbf{q})}{\partial q_j}f_j(t,\mathbf{q})-\frac{\partial\chi(t,q)}{\partial t} \tag{24}$$

The corresponding Hamilton equations to Eq. 24 are:

$$\left.\begin{array}{l} \dot{q}_k=\frac{\partial H(t,\mathbf{q},\mathbf{p})}{\partial p_k}=f_k(t,q) \\ p_k=-\frac{\partial H(t,\mathbf{q},\mathbf{p})}{\partial q_k}\sum_{j=1}^{n} f_j(t,\mathbf{q})\frac{\partial^2\chi(t,\mathbf{q})}{\partial q_k\partial q_j}+\frac{\partial^2\chi(t,\mathbf{q})}{\partial q_k\partial t}; \end{array}\right\}; \qquad k=1,2,\ldots,n \tag{25}$$

and that the time derivative of the Hamiltonian is:

$$\frac{dH\left(t,\mathbf{q},\mathbf{p}\right)}{dt}=\frac{\partial H\left(t,\mathbf{q},\mathbf{p}\right)}{\partial t}=\frac{\partial^2\chi\left(t,\mathbf{q}\right)}{\partial t^2}-\sum_{k=1}^{n} f_k\left(t,\mathbf{q}\right)\frac{\partial^2\chi\left(t,\mathbf{q}\right)}{\partial q_k\,\partial t} \tag{26}$$

Note in Eq. 26, as it is well-known, that the fact that the Hamiltonian total time derivative be equal to its partial time derivative is a general property of the Hamiltonian systems. On the other hand, $\chi\left(t,\mathbf{q}\right)$ can be fixed by the Third Postulate, which introduces in the formalism the generalized temperature $T\left(t,q\right)$ and the generalized Entropy and $S\left(t,q\right)$ in the Hamiltonian (Eq. 24) as:

$$\sum_{j=1}^{n}\frac{\partial\chi\left(t,\mathbf{q}\right)}{\partial q_j}f_j\left(t,\mathbf{q}\right)+\frac{\partial\chi\left(t,\mathbf{q}\right)}{\partial t}=-T\left(t,\mathbf{q}\right)\cdot St,q \tag{27}$$

Then, this Hamiltonian of Eq. 24 can also be written as:

$$H\left(t,\mathbf{q},\mathbf{p}\right)\sum_{j=1}^{n} f_j(t,\mathbf{q})\cdot p_j+T(t,\mathbf{q})\cdot S(t,\mathbf{q}) \tag{28}$$

that mathematically looks like much more to the Internal Energy of the classical Thermodynamics. The **Fourth Postulate** is a **generalized Gibbs-Duhem equation** [9], written as:

$$\sum_{k=1}^{n} f_k(t,\mathbf{q})\cdot dp_k+S(t,\mathbf{q})\cdot dT(t,\mathbf{q})=0 \tag{29}$$

Dividing Eq. 29 by dt, developing the total time derivative of $T\left(t,\mathbf{q}\right)$ and making use subsequently of Eq. 25, the generalized Gibbs-Duhem equation becomes:

$$S\left(t,\mathbf{q}\right)\frac{\partial T\left(t,\mathbf{q}\right)}{\partial t}+S\left(t,\mathbf{q}\right)\sum_{k=1}^{n} f_k\left(t,\mathbf{q}\right)\frac{\partial T\left(t,\mathbf{q}\right)}{\partial q_k}=-\sum_{k=1}^{n} f_k\left(t,\mathbf{q}\right)\left(\sum_{j=1}^{n} f_j\left(t,\mathbf{q}\right)\frac{\partial^2\chi\left(t,\mathbf{q}\right)}{\partial q_k\,\partial q_j}+\frac{\partial^2\chi\left(t,\mathbf{q}\right)}{\partial q_k\,\partial t}\right) \tag{30}$$

Taking the differential of $H\left(t,\mathbf{q},\mathbf{p}\right)$ in Eq. 28 and considering the generalized Gibbs-Duhem equation Eq. 29:

$$dH\left(t,\mathbf{q},\mathbf{p}\right)\sum_{k=1}^{n} dp_k\cdot df_k(t,\mathbf{q})+T(t,\mathbf{q})\cdot dS(t,\mathbf{q}) \tag{31}$$

Dividing Eq. 31 by dt, developing the total time derivative of $S\left(t,\mathbf{q}\right)$ and making use subsequently of Eqs. 23, 25 and 26, Eq. 31 becomes:

$$T\left(t,\mathbf{q}\right)\frac{\partial S\left(t,\mathbf{q}\right)}{\partial t}+T\left(t,\mathbf{q}\right)\sum_{k=1}^{n} f_k\left(t,\mathbf{q}\right)\frac{\partial S\left(t,\mathbf{q}\right)}{\partial q_k}=$$

$$= -\frac{\partial^2 \chi(t,\mathbf{q})}{\partial t^2} - \sum_{k=1}^{n} f_k(t,\mathbf{q}) \frac{\partial^2 \chi(t,\mathbf{q})}{\partial q_k \, \partial t} - \sum_{k=1}^{n} \frac{\partial \chi(t,\mathbf{q})}{\partial q_k} \left(\frac{\partial f_k(t,\mathbf{q})}{\partial t} + \sum_{j=1}^{n} \frac{\partial f_k(t,\mathbf{q})}{\partial q_j} f_j(t,\mathbf{q}) \right) \quad (32)$$

Eqs. (27), (30) and (32) define a system of three partial differential equations for $\chi(t,\mathbf{q})$, $T(t,\mathbf{q})$ and $S(t,\mathbf{q})$. However, the $\chi(t,\mathbf{q})$ function can be uncoupled by taking the total time derivative of $T(t,\mathbf{q}) \cdot St, q$ in Eq. 27, developing it by its partial derivatives, and comparing the result with Eqs. 30 and 32. This process provides the following equation for the $\chi(t,\mathbf{q})$ function:

$$\sum_{k=1}^{n} f_k(t,\mathbf{q}) \frac{\partial^2 \chi(t,\mathbf{q})}{\partial q_k \, \partial t} = 0 \quad (33)$$

Eq. 33 allows computing the $\chi(t,\mathbf{q})$ function, and at once simplifying Eqs. 30 and 32:

$$S(t,\mathbf{q}) \frac{\partial T(t,\mathbf{q})}{\partial t} + S(t,\mathbf{q}) \sum_{k=1}^{n} f_k(t,\mathbf{q}) \frac{\partial T(t,\mathbf{q})}{\partial q_k} = -\sum_{k=1}^{n} f_k(t,\mathbf{q}) \left(\sum_{j=1}^{n} f_j(t,\mathbf{q}) \frac{\partial^2 \chi(t,\mathbf{q})}{\partial q_k \, \partial q_j} \right) \quad (34)$$

$$T(t,\mathbf{q}) \frac{\partial S(t,\mathbf{q})}{\partial t} + T(t,\mathbf{q}) \sum_{k=1}^{n} f_k(t,\mathbf{q}) \frac{\partial S(t,\mathbf{q})}{\partial q_k} =$$

$$= -\frac{\partial^2 \chi(t,\mathbf{q})}{\partial t^2} - \sum_{k=1}^{n} \frac{\partial \chi(t,\mathbf{q})}{\partial q_k} \left(\frac{\partial f_k(t,\mathbf{q})}{\partial t} + \sum_{j=1}^{n} \frac{\partial f_k(t,\mathbf{q})}{\partial q_j} f_j(t,\mathbf{q}) \right) \quad (35)$$

In conclusion, Eqs. (33), (34) and (35) are the base to compute the $\chi(t,\mathbf{q})$, $T(t,\mathbf{q})$ and $S(t,\mathbf{q})$ functions.

4 Reversible Thermodynamics is a particular case of the GST?

In order to answer the question of this section, the non-explicit time dependence of Hamiltonian Eq. 28 is assumed. Therefore: (a) the system Eq. 1 is autonomous ($f_k(t,\mathbf{q}) = f_k(\mathbf{q})\,; k = 1, 2, \ldots, n$); (b) Temperature holds $T = T(\mathbf{q})$ and Entropy holds $S = S(\mathbf{q})$. In addition, it is assumed that there exists at least one steady state $\mathbf{qe} = (qe_1, qe_2, \ldots, qe_n)$, such that $f_k(\mathbf{qe}) = 0$; $k = 1, 2, \ldots, n$. Expanding at first order the $f_k(\mathbf{q})$ functions and substituting them in the Hamiltonian Eq. 28:

$$H(\mathbf{q},\mathbf{p}) \sum_{j=1}^{n} \left(\sum_{l=1}^{n} \nu_j l \, (q_l - qe_l) \right) \cdot p_j + T(\mathbf{q}) \cdot S(\mathbf{q}) =$$

$$= \sum_{l=1}^{n} \left(\sum_{j=1}^{n} \nu_j l \cdot p_j \right) (q_l - qe_l) + T(\mathbf{q}) \cdot S(\mathbf{q}) \quad (36)$$

In Eq. 36: $\nu_{jl} = \frac{\partial f_j}{\partial q_l}(\mathbf{qe})$. If the following canonical transformation is done:

$$\left. \begin{array}{l} Q_l = q_l - qe_l \\ P_l = \sum_{j=1}^{n} \nu_{jl} \cdot p_j \end{array} \right\}; \qquad l = 1, 2, \ldots, n \quad (37)$$

Then, the Hamiltonian Eq. 36 becomes:

$$H(\mathbf{Q},\mathbf{P}) = \sum_{l=1}^{n} P_l \cdot Q_l + T(\mathbf{Q}) \cdot S(\mathbf{Q}) \quad (38)$$

Eq. 38 is then similar to that corresponding to a reversible thermodynamics.

5 Conclusions

Note that all the formalism presented is an attempt to develop a complete classical *analytical dynamics* or *mechanics* for dynamical systems, here represented as coupled first order differential equation systems.

On the one hand, the formalism is called classical versus the quantum possible development from the Hamiltonian. In fact, a first attempt to bring the formalism to the quantum context is done in [7]. On the other hand, it is a first attempt because more background can be developed, such as the Hamilton-Jacobi equation, or the corresponding canonical transformations.

It is important to emphasize that the formalism is completely open to solve many theoretical problems as well as its applications. For instance, about the relationship with the physical formalism, already faced by Havas in [2] from a Lagrangian perspective, in which the energy conservation should be also faced from a Hamiltonian perspective. This is important because a second order differential equation formalism can be reduced to a first order formalism by introducing the velocities as new variables. The same happens with the relationship of both formalisms in the quantum context [7]: the probability is conserved but the Hamilton-Jacobi equation does not present any stochastic further term.

Another problem faced from the presented formalism is, as pointed out already in [6], that the Hamiltonian can be reinterpreted as a nonlinear version of the Thermodynamics Internal Energy, expressed by the second Thermodynamics postulate, identifying the Internal Energy with the Hamiltonian. In fact this formalism goes beyond, because from the Hamiltonian function, an attempt to develop a General System Thermodynamics (GST) is presented. Observe that the GST presented is nonlinear and a comparing with the classical Thermodynamics referred to quasi-static or reversible systems. Thus, the irreversible systems [9] should be described by the GST. Note that, at this point, Section 4 shows how a Hamiltonian that does not depend explicitly on time in the linear context about a steady state reproduces the classical reversible Internal Energy. However, both linear and nonlinear general approaches should reproduce for Eqs. 34 and 35 the third Thermodynamics postulate, i.e., $T\left(t,\mathbf{q}\right) > 0$, while any sign of the Entropy time derivative $\left(\dot{S}\left(t,\mathbf{q}\right) > 0, \dot{S}\left(t,\mathbf{q}\right) = 0 \text{ or } \dot{S}\left(t,\mathbf{q}\right) < 0\right)$ should inform us whether the system is tending to order or to disorder. The most interesting application in this context is that related with the dynamics of the chemical reactions, due to it is modelled with coupled first order differential equation systems [9]. It could be a way to state the whished objective of physics to unify dynamics and thermodynamics.

References

[1] Dirac, P. A. M., Lectures on Quantum Dynamics. New York, Belfer Graduate School of Science, Yeshiva University, 1964.

[2] Havas, P., The connection between Conservation laws and Invariance Groups: Folklore, Fiction, and Fact. Acta Physica Austriaca, 38: 145-167, 1973.

[3] Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V. and Mishchenko, E. F., The Mathematical Theory of Optimal Processes. New York, Gordon and Breach Science Publishers, 1986.

[4] Hooft, G.'t, The mathematical basis for deterministic quantum mechanics. arXiv: quant-ph/0604008v2, 26 Jun 2006.

[5] Blasone, M., Jizba, P. and Scardigli, F., Can quantum mechanics be an emergent phenomenon?. arXiv:0901.3907v2, 11 Feb 2009.

[6] Micó, J. C., An analytical formalism of dynamical systems. 9th Congress of the EUS-UES, 501-510, 2014.

[7] Micó, J. C., A quantum formalism of dynamical systems. 9th Congress of the EUS-UES, 511-518, 2014.

[8] Govaerts, J., Hamiltonian Reduction of First Order Actions. International Journal of Modern Physics A, 05: 3625-3640, 1990.

[9] Callen, H. B., Thermodynamics and an Introduction to Thermostatistics. New York, John Wiley and Sons, 1985.

32

Dynamical analysis of a new sixth-order parametric family for solving nonlinear systems of equations

Marlon Moscoso-Martínez[♭,♮,1], Alicia Cordero[♮], Juan R. Torregrosa[♮] and F. I. Chicharro [♮]

(♭) Faculty of Sciences,
Escuela Superior Politécnica de Chimborazo (ESPOCH).
Panamericana Sur km 1 1/2, 060106 Riobamba, Ecuador.
(♮) Institute for Multidisciplinary Mathematics,
Universitat Politècnica de València.
Camino de Vera s/n, 46022 València, Spain.

1 Introduction

A large number of problems in Computational Sciences and other disciplines can be stated in the form of a nonlinear equation or nonlinear system of equations using mathematical modelling. Finding the solution ξ of a nonlinear system of equations $F(x) = 0$ is a classical and difficult problem in Numerical Analysis, Applied Mathematics and Engineering, wherein $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is a sufficiently Frechet differentiable function in an open convex set D. We can find in [1, 2], and in the references therein, several overviews of the iterative methods for solving nonlinear systems published in the last years. The best known method for finding a solution $\xi \in D$ is Newton's scheme.

The dynamical behavior of the rational operator associated to iterative schemes for solving nonlinear systems, applied to low-degree polynomial systems, has shown to be an efficient tool for analyzing the stability and reliability of the methods, see for example [1, 3] and the references therein.

In this manuscript, we introduce a new sixth-order parametric family of multistep iterative schemes for solving nonlinear systems of equations as an extension of the family presented in [4] for solving nonlinear equations. This family is built from the Ostrowski's scheme, adding a Newton step with a "frozen" derivative and using a divided difference operator. We study its convergence, its real dynamics for stability and its numerical behavior. The dynamical planes are presented showing the complexity of the family. From the parameter spaces, presented in [4] for scalar functions, we have been able to determine different members of the family for vector functions that have bad convergence properties, since attracting periodic orbits and attracting strange fixed points appear in their dynamical planes. Moreover, this same study has allowed us to detect family members with specially stable behavior and suitable for solving practical problems. Several numerical tests are performed to illustrate the efficiency and stability of the presented family.

[1] marmosma@doctor.upv.es

2 New parametric family

The new triparametric family called MCTC(α, β, γ), object of study in this manuscript, has the following iterative expression:

$$\begin{cases} y^{(k)} = x^{(k)} - [F'(x^{(k)})]^{-1}F(x^{(k)}) \\ z^{(k)} = y^{(k)} - [2[x^{(k)}, y^{(k)}; F] - F'(x^{(k)})]^{-1}F(y^{(k)}) \\ x^{(k+1)} = z^{(k)} - (\alpha I + \beta u^{(k)} + \gamma v^{(k)})[F'(x^{(k)})]^{-1}F(z^{(k)}) \end{cases}, \tag{1}$$

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, β and γ are arbitrary parameters. The divided difference operator $[x, y; F]$, defined in [5], 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 13 (triparametric family)**.** *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 nonlinear 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* (1) *converges to ξ with order four, being its error equation*

$$e^{(k+1)} = (1 - \alpha - \gamma)\left(C_2^3 - C_3C_2\right)e^{(k)^4} + \mathcal{O}(e^{(k)^5}),$$

where $e^{(k)} = x^{(k)} - \xi$, $C_q = \frac{1}{q!}[F'(\xi)]^{-1}F^{(q)}(\xi)$ and $q = 2, 3, ...$

From Theorem 13, it follows that the new triparametric family has an order of convergence of four for any value of α, β and γ. However, convergence can be speed-up if only one parameter is held and the family is reduced to an uniparametric iterative scheme.

Theorem 14 (uniparametric family)**.** *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 nonlinear 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* (1) *converges to ξ with order six, provided that $\beta = 1 + \alpha$ and $\gamma = 1 - \alpha$, 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 $q = 2, 3, ...$

From Theorem 14, it follows that if we only hold α in (1), the triparametric family is reduced to an uniparametric family with an order of convergence of six, for any value of α, as long as $\beta = 1+\alpha$ and $\gamma = 1 - \alpha$. So, the iterative expression of the new three-step uniparametric family, dependent only of α and which we will call MCTC(α) family, is defined as

$$\begin{cases} y^{(k)} = x^{(k)} - [F'(x^{(k)})]^{-1}F(x^{(k)}) \\ z^{(k)} = y^{(k)} - [2[x^{(k)}, y^{(k)}; F] - F'(x^{(k)})]^{-1}F(y^{(k)}) \\ x^{(k+1)} = z^{(k)} - (\alpha I + (1+\alpha)u^{(k)} + (1-\alpha)v^{(k)})[F'(x^{(k)})]^{-1}F(z^{(k)}) \end{cases}, \tag{2}$$

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.

Because of the results obtained with the convergence analysis carried out, from now on we only work with MCTC(α) family of iterative methods and, to select the best members of this family, we use the real dynamics tools discussed in Section 3.

3 Real dynamics for stability

This section refers to the study of the dynamical behavior of the rational operator associated with iterative schemes of MCTC(α) family. This study give us important information about the stability and reliability of the family. We will construct dynamical planes in order to show the behavior of particular methods in terms of the basins of attraction of their fixed points, periodic points, etc.

3.1 Rational operator

The rational operator can be built on any nonlinear system; however, we construct this operator on the following low-degree nonlinear polynomial system:

$$F(x_1,x_2)=\left(x_1^2-1,x_2^2-1\right)=(0,0)\,, \tag{3}$$

since the criterion of stability or instability of a method applied to this system can be generalized for other multidimensional cases.

Proposition 1 (rational operator R_F). Let the polynomial system $F(x_1,x_2)$ given in (3), with roots $(-1,-1)$, $(-1,1)$, $(1,-1)$, $(1,1)\in\mathbb{R}^2$. The rational operator associated with MCTC(α) family and applied on $F(x_1,x_2)$, with $\alpha\in\mathbb{R}$ an arbitrary parameter, is

$$R_F(x_1,x_2,\alpha)=\left(R_{F_{11}},R_{F_{12}}\right), \tag{4}$$

where

$$R_{F_{11}}=\frac{1}{32}\left(\frac{\left(x_1^2-1\right)^4\left(\alpha+(\alpha-19)x_1^4-2(\alpha-1)x_1^2+1\right)}{4x_1^5\left(x_1^2+1\right)^2\left(3x_1^2+1\right)}+\frac{8\left(x_1^4+6x_1^2+1\right)}{x_1^3+x_1}-\frac{\alpha\left(x_2^2-1\right)^4}{x_2^3\left(x_2^2+1\right)^2}\right),$$

$$R_{F_{12}}=\frac{1}{32}\left(\frac{\left(x_2^2-1\right)^4\left(\alpha+(\alpha-19)x_2^4-2(\alpha-1)x_2^2+1\right)}{4x_2^5\left(x_2^2+1\right)^2\left(3x_2^2+1\right)}+\frac{8\left(x_2^4+6x_2^2+1\right)}{x_2^3+x_2}-\frac{\alpha\left(x_1^2-1\right)^4}{x_1^3\left(x_1^2+1\right)^2}\right).$$

To simplify the rational operator R_F defined in Proposition 1, we can select a value of α that cancels terms of the expression and reduces it. It is easy to show that for $\alpha=0$, the rational operator is simpler and there will be fewer fixed and critical points that can improve the stability of the associated method. Also, the components of this $R_F(x_1,x_2,0)$ will be of separate variables.

3.2 Fixed points and their stability

We calculate the fixed points of the rational operator $R_F(x_1,x_2,\alpha)$ given in (4), and analyze their stability.

Proposition 2 (fixed points). The real fixed points of $R_F(x_1,x_2,\alpha)$ are the roots of the equation $R_F(x_1,x_2,\alpha)=(x_1,x_2)$. That is

$$fp_1=(-1,-1), fp_2=(-1,1), fp_3=(1,-1), fp_4=(1,1),$$

that correspond to the roots of the polynomial system $F(x_1,x_2)$ given in (3), and they are also superattracting. Other strange fixed points may appear but their components are roots of polynomials of very high degrees.

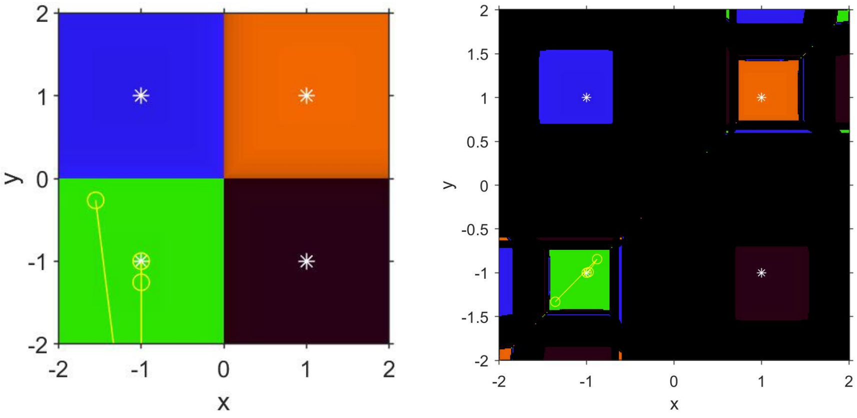
From Proposition 2, we establish there is a minimum of 4 fixed points. Of these, from fp_1 to fp_4 correspond to the roots of the original polynomial system $F(x_1,x_2)$ and are attractive and critical points.

3.3 Dynamical planes

Here, we study the stability of two MCTC(α) family methods as representatives. The first method is for $\alpha = 0$, whose value is inside the stability region of the parameter spaces shown in [4], that is, it is in the red area. The second method is for $\alpha = 200$, whose value is outside the stability region of the same parameter spaces, located in the black area.

Dynamical planes are built with a mesh from -2 to 2, with a step equal to 0.01. Every initial estimation is iterated 100 times (maximum) with a tolerance of 10^{-3}. The points in the mesh are represented based on the roots to which they converge: the 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. Fixed points are illustrated with a white circle '○', critical points with a white square '□' and attractors with a white asterisk '$*$'. Also, the basins of attraction are depicted in different colors. The resulting graphic is made in Matlab R2020b with a resolution of 400x400 pixels.

Thus, the dynamical planes for $R_F(x_1, x_2, 0)$ and $R_F(x_1, x_2, 200)$, with some convergence orbits in yellow, are shown in Figure 1. On the one hand, the method for $\alpha = 0$ presents four basins of attraction associated with the roots. Also, there are no black areas of non-convergence to the solution. Consequently, this method shows good dynamical behavior: it is very stable. On the other hand, the method for $\alpha = 200$ presents the same four basins of attraction associated with the roots, but of reduced size, which minimizes the chances of convergence to the solution. Likewise, there are black areas of slow convergence of the method. Consequently, this method has poor dynamical behavior: it is unstable.



(a) $\alpha = 0$ and $pf = (-1, -1)$

(b) $\alpha = 200$ and $pf = (-1, -1)$

Figure 1: Dynamical planes for R_F.

4 Numerical results

In this section, we perform several numerical tests to illustrate the efficiency and stability of the presented family. We consider the same two members of MCTC(α) proposed in Section 3.3, for $\alpha = 0$ and $\alpha = 200$. These methods are applied on two nonlinear test systems, whose expressions and corresponding roots are shown in Table 1.

Table 1: Nonlinear test systems and corresponding roots.

Nonlinear test system	Roots
$F_1(x_1,x_2) = (x_1^2-1, x_2^2-1) = (0,0)$	$\xi \approx (1,1)^T$
$F_2(x_1,x_2) = \left(x_1^2+x_2^2-1, x_1^2-x_2^2-\frac{1}{2}\right) = (0,0)$	$\xi \approx \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)^T$

The calculations have been developed in Matlab R2020b programming package using variable precision arithmetics with 200 digits of mantissa. For each method, we analyze the number of iterations (iter) required to converge to the solution, so that the stopping criteria $||x^{(k+1)} - x^{(k)}|| < 10^{-100}$ or $||F(x^{(k+1)})|| < 10^{-100}$ are satisfied.

To check the theoretical order of convergence of the methods, we calculate the approximate computational order of convergence (ACOC) given in [12]. In the numerical results, if the ACOC vector inputs do not stabilize their values throughout the iterative process, it is marked as '-'; and, if any of the methods used does not reach convergence in 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 near and far from the solution, that is, for $x^{(0)} \approx 2\xi$ and $x^{(0)} \approx 10\xi$. The results are encouraging because we can notice that MCTC(0) always converges to the solution in the two nonlinear test systems, regardless of the initial estimates used. Therefore, we verify this method is robust, according to the stability results shown in Section 3.

Table 2: Numerical performance of MCTC(0) on test problems.

System	$x^{(0)}$		$\|\|x^{(k+1)} - x^{(k)}\|\|$	$\|\|F(x^{(k+1)})\|\|$	iter	ACOC
F_1	$\approx 2\xi$	$(2,2)^T$	6.4031e-71	8.0537e-173	5	3.1906
F_2	$\approx 2\xi$	$(1.7,1)^T$	6.6576e-68	1.4261e-166	6	3.6518
F_1	$\approx 10\xi$	$(10,10)^T$	6.7666e-41	3.4488e-113	10	6.4467
F_2	$\approx 10\xi$	$(9,5)^T$	3.9e-70	5.5314e-172	12	3.4528

Now, in Table 3 we show the numerical performance of MCTC(200) for initial estimations very close to ($x^{(0)} \approx \xi$) and near to ($x^{(0)} \approx 2\xi$) the solution.

Table 3: Numerical performance of MCTC(200) on test problems.

System	$x^{(0)}$		$\|\|x^{(k+1)} - x^{(k)}\|\|$	$\|\|F(x^{(k+1)})\|\|$	iter	ACOC
F_1	$\approx \xi$	$(1.5,1.5)^T$	8.1881e-94	1.5574e-207	4	2.5252
F_2	$\approx \xi$	$(1.3,0.8)^T$	nc	nc	nc	nc
F_1	$\approx 2\xi$	$(2,2)^T$	nc	nc	nc	nc
F_2	$\approx 2\xi$	$(1.7,1)^T$	nc	nc	nc	nc

Note that the results shown in Table 3 also corroborate the stability analysis performed in Section 3. The MCTC(200) presents convergence problems even for estimates very close to the root ($x^{(0)} \approx \xi$), this method does not converge to the solution in one of two cases. Furthermore, for estimations near to the root ($x^{(0)} \approx 2\xi$), it does not converge to the solution in all cases, establishing a dependency on the initial estimates used. Therefore, the instability of this method is verified.

5 Conclusions

A highly efficient family of iterative methods MCTC(α) has been designed to solve nonlinear systems. This family proved to have an excellent numerical performance considering stable members as representatives. The method for $\alpha = 0$ proved to be robust (stable), according to the real dynamics analysis performed. The method for $\alpha = 200$ proved to be unstable, chaotic and cannot converge to the solution according to the initial estimate and the nonlinear system used. Also, the order of convergence is verified by the ACOC, which is close to 6. Numerical experiments confirm the theoretical results.

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] Ortega, J. M., Rheinboldt, W. C., Iterative Solution of Nonlinear Equations in Several Variables. New York, Academic Press, 1970.

[6] Neta, B., Numerical methods for the solution of equations. California, Net-A-Sof, 1983.

[7] Petković, M., Neta, B., Petković, L., Džunić, J., Multipoint Methods for Solving Nonlinear Equations. Boston, Academic Press, 2013.

[8] Amat, S., Busquier, S., Advances in Iterative Methods for Nonlinear Equations. Switzerland, Springer, 2017.

[9] Ortega, J. M., Rheinboldt, W. C., Iterative Solution of Nonlinear Equations in Several Variables. New York, Academic Press, 1970.

[10] Ostrowski, A. M., Solutions of Equations and Systems of Equations. New York, Academic Press, 1966.

[11] Hueso, J. L., Martínez, E., Teruel, C., Convergence, efficiency and dynamics of new fourth and sixth order families of iterative methods for nonlinear systems *Journal of Computational and Applied Mathematics*, 275:412–420, 2015.

[12] Cordero, A., Torregrosa, J.R., Variants of Newton's Method using fifth-order quadrature formulas. *Applied Mathematics and Computation*, 190(1): 686—698, 2007.

Higher order numerical methods for addressing an embedded steel constitutive model

J.J. Padilla[♮,1], A. Cordero[♭], A.M. Hernández-Díaz[◇] and J.R. Torregrosa[♭]

(♮) Departamento de Ingeniería Civil,
UCAM, Universidad Católica de Murcia.
Avenida de los Jerónimos, 135, Murcia, Spain.
(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(◇) Área de Mecánica de Medios Continuos y Teoría de Estructuras
Universidad de La Laguna
Avda. Ángel Guimerá s/n, 38200, Santa Cruz de Tenerife

1 Introduction

Structural models usually present several types of nonlinearities [1,2]. Particularly, some materials, such as the reinforced concrete (RC), introduce nonlinear stress-strain relationships in the mechanical model. The implementation of these models usually requires the application of numerical methods for solving the corresponding nonlinear equations, such as, for example, the Newton-type methods [3]. In fact, the correct solver for solving a nonlinear problem is often a choice between computational cost and accuracy [4]. However, despite of the high nonlinear nature of the problem, sometimes the previous determination of a solvability region for the equation to solve is possible using algebraic procedures, what allows improving the efficiency of the numerical solver.

2 Problem statement

2.1 Solvability of the steel constitutive model

Several authors [5–7] consider a steel reinforcement stiffened by the concrete bonded to it (or also called as "embedded bar model"). One of these approaches, the so-called Refined Compression Field Theory (RCFT) [8], includes in the steel constitutive model an equilibrium condition that takes into account the concrete tension stiffening effect between cracks. As result, a nonlinear equation is introduced in the steel constitutive model in terms of the apparent yield strain. This last theory predicts the average stress of an embedded bar as a function of the average strain (i.e., measured on certain length including several cracks) such as follows:

$$\sigma_{s,av} = \begin{cases} f_y - \dfrac{A_c}{A_s}\dfrac{f_{ct}}{1+\sqrt{3.6M\varepsilon_{s,av}}} & \text{if} \quad \varepsilon_{s,av} \geq \varepsilon_{max}, \\ E_s\varepsilon_{s,av} & \text{if} \quad \varepsilon_{s,av} < \varepsilon_{max}, \end{cases} \tag{1}$$

[1] jjpadilla@ucam.edu

with,

$$\varepsilon_{max} = \frac{f_y}{E_s} - \frac{\overline{\dfrac{f_{ct}}{1+\sqrt{3.6M\varepsilon_{max}}}}}{E_sA_s}, \qquad M = \frac{A_c}{\sum \pi\phi},$$

where E_s is the elastic modulus of the steel, f_y is the steel yield stress, f_{ct} is the tensile concrete strength, $\sigma_{s,av}$ and $\sigma_{ct,av}$ are the average tensile stresses in the reinforcing steel and in the concrete, respectively, A_s is the cross section of the steel bars, $\varepsilon_{s,av}$ is the average strain in the reinforcing bar and A_c is the area of concrete bonded to the bar participating in the tension stiffening effect. The previous formulation is based on the concept of equilibrium of forces between a cracked section (where only the reinforcement contributes) and a generic section (where both steel plus the surrounding concrete contribute; see Figure 1). Technical codes [9] propose a value for A_c equal to the rectangular area (surrounding the bar) over a distance not exceeding 7.5Ø from the center of the bar (Figure 1), and Ø is the diameter of the bar. It can be proven that when this value is adopted, for certain specimens (specifically those with high values of the ratio f_{ct}/ρ, being ρ the reinforcement ratio) it is not possible to obtain a positive real solution for the apparent yield strain (ϵ_{max}) defined in Equation (1). Moreover, if the solvability analysis is posed in terms of the variables f_{ct} and ϵ_{max} and the value of the area A_c is increased monotonically, from a certain value of this area the bijection between f_{ct} and ϵ_{max} is broken (*problem of uniqueness*), and subsequently, for higher values of the area A_c until up the value prescribed by technical codes, the absence of a real solution is achieved (*problem of existence*) [10]. This last fact indicates that the equilibrium of internal forces along the cracked member is not verified.

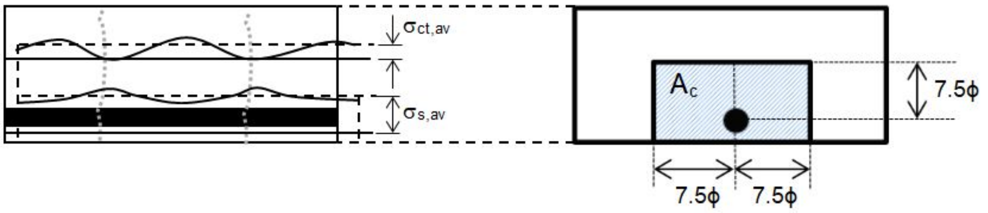


Figure 1: Average stresses profiles ($\sigma_{ct,av}$ and $\sigma_{st,av}$) and prescribed tension stiffening area (A_c) effectively bonded to the steel.

In [10] the greatest value of the area A_c for which the embedded steel constitutive model has, at least, a positive real solution, is determined using algebraic procedures. In this sense, the greatest portion of the tension stiffening area which may be taken in order to preserve the solvability of the constitutive model (i.e., in order to preserve the internal equilibrium of forces, in such a way that as concrete participation increases, the steel stress diminishes) is represented by the following coefficient:

$$\lambda = \frac{A_s \cdot f_y}{A_c \cdot f_{ct}} \cdot \left(\frac{2}{3} + \frac{\sqrt{(1+10.8 \cdot M \cdot \epsilon_y)^3}}{48.6 \cdot M \cdot \epsilon_y} \right) \tag{2}$$

being ϵ_y the strain corresponding to the steel yield stress (i.e.,$\epsilon_y = f_y/E_s$). Therefore, the coefficient λ represents the boundary of the solvability region for the embedded steel constitutive model proposed by the RCFT; since this region is only obtained from algebraic considerations, it may be applied to every experimental approach of the concrete tension stiffening model.

The previous boundary, for certain design cases, may be lower than the value prescribed by the technical codes for the area A_c (i.e., the solvability region lies within the design range prescribed by technical codes). In fact, several works [8,11,12] point out the convenience of correcting the tension stiffening area in order to adjust the shear response of reinforced concrete members, particularly for high shear strains, where the technical codes underestimate the concrete tension stiffening.

Input	**Range**
$E_s\,(MPa)$	$[195000, 205000]$
$f_y\,(MPa)$	$[400, 500]$
$f_{ct}\,(MPa)$	$[25, 50]$
$\phi\,(mm)$	$[6, 40]$
$\lambda\,(dimensionless)$	$[0, \lambda_{lim}]$

Table 1: Ranges for input parameters of the Equation (2).

3 Higher order numerical methods

The Newton-Raphson method belongs to the family of the so-called linearization methods for solving non-linear equations and systems of equations and its iterative scheme is based on a local convergence process [13], what might be critical in the context of solvability described in the Section 2.1 (particularly, in the boundary of such region).

Due to this local convergence, several strategies have been developed to allow the convergence to be achieved from far starting points into the neighbourhood of a local minimizer. Likewise, one of the main drawbacks of the Newton-type methods is the evaluation of the Hessian matrix, since, in certain contexts (i.e., physics equations), the second-order derivative may be highly expensive from a computational point of view. One of these strategies are the higher-order modifications of Newton's method for solving nonlinear equations in order to increase the order of convergence [14], or even, the combination of the high-order Newton method and the Newton method in order to reduce the calculation time and to improve the efficiency [15].

3.1 A parametric family of root-finding iterative methods

We will use a two-step uniparametric family, of fourth-order iterative methods, to find roots of nonlinear equation (1). This family uses a damped Newton in the first step (predictor), and the second step (corrector) is defined as a Newton-type scheme, in which three functional evaluations are used.

Theorem 1. *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a sufficiently differentiable function in an open interval I and $x^* \in I$ a simple root of equation $f(x) = 0$. Let $G(\eta)$, be a real function, satisfying $G(1) = 1$, $G'(1) = \dfrac{-3}{4}$, $G''(1) = \dfrac{9}{4}$, and $|G'''(1)| < +\infty$. If $\gamma = \dfrac{2}{3}$ and we choose an initial approximation x_0 close enough to x^*, then iterative family defined by*

$$\begin{aligned} y_k &= x_k - \gamma \frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} &= x_k - G(\eta) \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, 2, ..., \end{aligned} \tag{3}$$

satisfying the error equation below:

$$e_{k+1} = \left[\left(5 + \frac{32G'''(1)}{81} \right) c_2^3 - c_2 c_3 + \frac{c_4}{9} \right] e_k^4 + O(e_k^5),$$

where $c_k = \dfrac{1}{k!}\dfrac{f^{(k)}(\alpha)}{f'(\alpha)}$, $\eta = \frac{f'(y_k)}{f'(x_k)}$, $k = 2, 3, \ldots$ *and* $e_k = x_k - x^*$*. Therefore, all the members of class 3 converges to* x^* *with order of convergence four.*

We select a particular subclass of iterative methods, depending on a parameter α, whose iterative expression is

$$\begin{aligned} y_k &= x_k - \frac{2}{3}\frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} &= x_k - \Big(1 - \frac{3}{4}(\eta_k - 1) + \frac{9}{8}(\eta_k - 1)^2 + \alpha(\eta_k - 1)^3\Big)\frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, 2, ... \end{aligned} \tag{4}$$

where the variable of the weight function is $\eta_k = \dfrac{f'(y_k)}{f'(x_k)}$.

In [16] the dynamical behavior of this class of iterative methods was studied, finding the most stable members (in particular $\alpha = 1$) and also those showing chaotic performance (as $\alpha = -20 + 45i$). These values are used in the next section.

4 Numerical results

In order to evaluate the efficiency of different numerical iterative methods (among them, the parametric family previosly described) regarding to the physical problem presented in Section 2, a sample of 100 values for the inputs of the Equation (2) has been adopted. This sample has been generated randomly, within the ranges presented in Table 1 for each input parameter.

Numerical computations have been carried out by using $MATLAB^{ő}$ R2019a, on a PC equipped with a $Intel^{ő}$ Core™ i5-5200U CPU 2.20GHz. In Table 1, we show the residuals $|f(x_{k+1})|$ and $|x_{k+1} - x_k|$ at the last iteration, the estimation of the solution found, the number of iterations needed and the execution time in seconds, calculated with the command *cputime*. The stopping criterion used is $|x_{k+1} - x_k| + |f(x_{k+1})| < 10^{-6}$.

Solution	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1} - x_k\rvert$	Iteration
0.000635867598422514	6,56612e-11	5,33986e-07	7
0.000517406121549559	6,75826e-11	4,96554e-07	7
0.000446396692152001	6.81218e-11	4.86223e-07	7
0.000599111127570943	6,64413e-11	5,22456e-07	7
0.000600487981552110	6,64141e-11	5,26344e-07	7
0.000413110652300844	6,81422e-11	4,79587e-07	7
0.000605207957076228	6,62967e-11	5,33798e-07	7
0.000525637445813513	6,74148e-11	4,92513e-07	7
0.000537940095960127	6,74358e-11	5,12146e-07	7
0.000451978308427051	6,81995e-11	5,01625e-07	7

Table 2: Method (3), $\alpha = 1$, $x_0 = \dfrac{450}{2e5}$, t=4.5 s

Solution	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1}-x_k\rvert$	Iteration
0.000635026251270566 - 1.82794944086520e-07i	3,46e-10	9,75837e-07	37
0.000517111210020908 + 5.82759991250184e-07i	3,84e-10	9,43327e-07	36
0.000446140213854697 + 5.57181484272692e-07i	3,66636e-10	8,99109e-07	36
0.000598305827351175 - 1.46005246945238e-07i	3,19626e-10	9,12315e-07	37
0.000599671246190458 - 1.54467715109439e-07i	3,27595e-10	9,30678e-07	37
0.000412872936808692 + 5.43899516322722e-07i	3,5893e-10	8,77331e-07	36
0.000604369040412069 - 1.71703918414669e-07i	3,43246e-10	9,6699e-07	37
0.000525350335557838 + 5.70774200440979e-07i	3,72896e-10	9,23301e-07	36
0.000537166425339973 - 1.08432551780129e-07i	2,97758e-10	8,56745e-07	37
0.000451683580721046 + 6.05570673260231e-07i	4,09452e-10	9,79693e-07	36

Table 3: Method (3), $\alpha = -20 + 45i$, $x_0 = \dfrac{450}{2e5}$, t=8.0 s

In order to compare the results, in terms of robustress and effectiveness, we check the performance of high-order schemes (3) with $\alpha = 1$ and $\alpha = -20 + 45i$. However, we perform also the numerical test on known classical procedures as Newton and Jarrat. By using $x_0 = \dfrac{450}{2e5}$ as initial guess of the apparent yield strain ϵ_{max}, we notice in Table 1 and 3 that performance of the stable element of family (3) $\alpha = 1$ is much better than the unstable one $\alpha = -20 + 45i$, as in terms of number of iterations needed to converge (7 in case of $\alpha = 1$, and around 37 for $\alpha = -20 + 45i$) as in precision and, of course, in terms of mean execution time.

Solution	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1}-x_k\rvert$	Iteration
0.000636206293661355	3,42126e-10	6,03471e-07	10
0.000517716984601877	3,47046e-10	5,57127e-07	10
0.000446700051134824	3,49e-10	5,44532e-07	10
0.000599441292717526	3,44772e-10	5,8924e-07	10
0.000600820952928857	3,45036e-10	5,93973e-07	10
0.000413409408484667	3,48016e-10	5,36655e-07	10
0.000605546303030168	3,45179e-10	6,0304e-07	10
0.000525945349712532	3,45645e-10	5,52161e-07	10
0.000538262467333356	3,48397e-10	5,76359e-07	10
0.000452293149904906	3,5123e-10	5,63653e-07	10

Table 4: Newton, $x_0 = \dfrac{450}{2e5}$, t=4.4 s

Regarding known methods, we observe in Table 4 that Newton's scheme is able to find the solution in a similar time as the best fourth order method, but using almost a 50% more iterations. Also Jarratt's procedure whose numerical results can be seen in Table 5, show the same mean time as the best proposed scheme (3) with $\alpha = 1$ with a slightly lower number of iterations.

Solution	$\lvert f(x_{k+1})\rvert$	$\lvert x_{k+1} - x_k\rvert$	Iteration
0.000635801360966030	3,6846e-11	5,93549e-07	6
0.000517349955326238	4,01937e-11	5,68203e-07	6
0.000446343620616066	4,13593e-11	5,62143e-07	6
0.000599048149568916	3,79913e-11	5,86213e-07	6
0.000600424161803673	3,7832e-11	5,89457e-07	6
0.000413059263537066	4,18013e-11	5,57337e-07	6
0.000605142462375379	3,74734e-11	5,95492e-07	6
0.000525581928713347	4,0176e-11	5,64157e-07	6
0.000537880456824904	3,94078e-11	5,80921e-07	6
0.000451922171196378	4,08155e-11	5,75796e-07	6

Table 5: Jarratt, $x_0 = \dfrac{450}{2e5}$, t=4.5 s

5 Conclusions

In this manuscript, a new strategy for solving the yield strain equation is proposed. The main approach is to include values of parameters in the boundary of the solvability region and use a high-order family of iterative methods. In particular, stable members of this class of iterative methods show very good numerical performance in terms of precision and number of iterations/execution time.

References

[1] Hernández-Díaz, A. M., Muñoz, A., Jiménez-Alonso, J. F., Sáez, A. (2019). Buckling design of submerged arches via shape parameterization. Computational and Mathematical Methods, 1(5), e1057.

[2] Jiménez-Alonso, J. F., Pérez-Aracil, J., Hernández Díaz, A. M., Sáez, A. (2019). Effect of Vinyl flooring on the modal properties of a steel footbridge. Applied Sciences, 9(7), 1374.

[3] Galántai, A.: The theory of newton's method. Journal of Computational and Applied Mathematics, 124(1-2):25–44, 2000.

[4] Pérez-Aracil, J., Camacho-Gómez, C., Hernández-Díaz, A. M., Pereira, E., Camacho, D., Salcedo-Sanz, S. (2021). Memetic coral reefs optimization algorithms for optimal geometrical design of submerged arches. Swarm and Evolutionary Computation, 67, 100958.

[5] Belarbi, Abdeldjelil and Thomas TC Hsu: Constitutive laws of concrete in tension and reinforcing bars stiffened by concrete. Structural Journal, 91(4):465–474, 1994

[6] Pang, Xiao Bo David and Thomas TC Hsu: Behavior of reinforced concrete membrane elements in shear. Structural Journal, 92(6):665–679, 1995.

[7] Gil-Martín, Luisa María, Enrique Hernández-Montes, Mark Aschheim, and Stavroula Pantazopoulou: Refinements to compression field theory, with application to wall-type structures. American Concrete Institute Special Publication, 265:123–142, 2009.

[8] Palermo, M, LM Gil-Martin, E Hernandez-Montes, and M Aschheim: Refined compres- sion field theory for plastered straw bale walls. Construction and Building Materials, 58:101–110, 2014

[9] Code, Price: Eurocode 2: design of concrete structures-part 1–1: general rules and rules for buildings. British Standard Institution, London, 2005.

[10] Hernández-Díaz, A.M. and M.D. García-Román: Computing the refined compression field theory. International Journal of Concrete Structures and Materials, 10(2):143–147, 2016

[11] Palermo, M., L.M. Gil-Martín, T. Trombetti, and E. Hernández-Montes: In-plane shear behaviour of thin low reinforced concrete panels for earthquake re-construction. Materials and structures, 46(5):841–856, 2013.

[12] España, R. M., Hernández-Díaz, A. M., Cecilia, J. M., García-Román, M. D. (2017). Evolutionary strategies as applied to shear strain effects in reinforced concrete beams. Applied Soft Computing, 57, 164-176.

[13] Rheinboldt, W.C.: Methods for solving systems of nonlinear equations. SIAM, 1998.

[14] Guo, C., Y. Gao, and C. Xia: Improved newton iteration method and convergence order analysis. In Proceedings of the 2020 4th International Conference on Digital Signal Processing, pages 29–32, 2020

[15] Wei, Y., Q. Li, P. Wang, M. Yang, Z.Zeng, and X. Wang: A new combination algorithm based on higher-order newton and simplified newton method. In The 16th IET International Conference on AC and DC Power Transmission (ACDC 2020), volume 2020, pages 1804–1809. IET, 2020.

[16] Padilla, J.J.; Chicharro, F.I.; Cordero, A.; Torregrosa, J.R. Parametric Family of Root-Finding Iterative Methods: Fractals of the Basins of Attraction. Fractal Fract. 2022, 6, 572.

A generalization of subdirect sums of matrices

F. Pedroche♮,1

♮ Institut Universitari de Matemàtica
Multidisciplinària, Universitat Politècnica de València
Camí de Vera s/n, 46022, València, España

1 Introduction

In this paper we present the concept of *weighted subdirect sum of matrices* that has been recently published [5]. The definition of this new concept arises as an extension of the term *subdirect sum of matrices* coined in [3]. A review of the literature concerning this topic can be found in [4], where it can be checked that the subdirect sum of matrices has been analyzed for a wide variety of matrix classes (for example, in [1] is studied the case of inverses of M-matrices, in [2] the case of S-Strictly Diagonally Dominant matrices, etc.). The subdirect sum of matrices appears naturally when studying systems that present some overlapping regions, such as in graphs, in numerical applications of finite elements, when studying overlapping areas in a brain [6], etc.. Therefore, it seems useful to define an extension of the subdirect sum as we comment in this paper.

2 Subdirect sum of matrices

The subdirect sum of matrices was introduced in [3], as follows. Given two matrices

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \tag{1}$$

with A_{22} and B_{11} of size $k \times k$, the sum

$$C = \begin{bmatrix} A_{11} & A_{12} & O \\ A_{21} & A_{22} + B_{11} & B_{12} \\ O & B_{21} & B_{22} \end{bmatrix} \tag{2}$$

is called the *k-subdirect sum* (or simply the *subdirect sum*) of A and B, and it is denoted as $C = A \oplus_k B$. Notice that the particular case $k = 0$ can be considered as the usual direct sum of matrices.

Example 1. Given

$$A = \begin{bmatrix} 2 & 3 \\ 5 & 4 \end{bmatrix} \quad B = \begin{bmatrix} 7 & -2 \\ 1 & 9 \end{bmatrix}$$

[1] pedroche@imm.upv.es

The 1-subdirect sum of A and B is

$$A \oplus_1 B = \begin{bmatrix} 2 & 3 & 0 \\ 5 & 11 & -2 \\ 0 & 1 & 9 \end{bmatrix}$$

Given a matrix class $\mathcal{S}$, the questions that were addressed in [3], for some positivity matrix classes, were the following.

Q1 If A and B belong to $\mathcal{S}$, does it hold that $A \oplus_k B$ belongs to the same class $\mathcal{S}$, for $k = 1$?

Q2 Given a matrix of the form

$$C = \begin{bmatrix} C_{11} & C_{12} & O \\ C_{21} & C_{22} & C_{23} \\ O & C_{32} & C_{33} \end{bmatrix}$$

belonging to a class $\mathcal{S}$, can it be written as $A \oplus_1 B$ with A and B in the same class as C, for $k = 1$?

Q3 Question 1 but taking $k > 1$.

Q4 Question 2 but taking $k > 1$.

3 Weighted subdirect sum of matrices

In [5] the weighted subdirect sum of matrices is defined as follows. Given two matrices A and B partitioned as in (1), and for nonnegative α and β

$$A \oplus_k^{\alpha,\beta} B := \begin{bmatrix} A_{11} & A_{12} & O \\ A_{21} & \alpha A_{22} + \beta B_{11} & B_{12} \\ O & B_{21} & B_{22} \end{bmatrix} \tag{3}$$

Note that the weighted subdirect sum can be written in terms of the subdirect sum in the form

$$A \oplus_k^{\alpha,\beta} B = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & \alpha A_{22} \end{bmatrix} \oplus_k \begin{bmatrix} \beta B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

Example 2. Given

$$A = \begin{bmatrix} 1 & 5 & 4 \\ 1 & 2 & 5 \\ 3 & 1 & 9 \end{bmatrix}, B = \begin{bmatrix} 6 & 12 & -5 \\ 1 & -5 & 1 \\ 2 & 4 & 7 \end{bmatrix}$$

we have

$$A \oplus_2^{2,3} B = \begin{bmatrix} 1 & 5 & 4 & 0 \\ 1 & 22 & 46 & -5 \\ 3 & 5 & 3 & 1 \\ 0 & 2 & 4 & 7 \end{bmatrix}$$

Class	Question			
	$\mathcal{Q}1$	$\mathcal{Q}2$	$\mathcal{Q}3$	$\mathcal{Q}4$
PSD	A	A	A	A
PD	A	A	A	A
SM	A	A	A	A
CP	A	A	A	A
DN	A	A	A	A
M	A	A		A
P	A	A		A
P_0	A	A		A
TN	A	A		A

Table 1: Summary of affirmative answers to questions $\mathcal{Q}1$ to $\mathcal{Q}4$.

4 Results

In [5] we have studied the questions $Q1$ to $Q4$ when using the weighted subdirect sum of matrices. In more detail, the questions that have been addressed were the following:

$\mathcal{Q}1$ If A and B belong to $\mathcal{S}$, do exist some $\alpha > 0$ and $\beta > 0$ such that $A \oplus_k^{\alpha,\beta} B$ belongs to the same class $\mathcal{S}$, for $k = 1$?

$\mathcal{Q}2$ Given a matrix of the form

$$C = \begin{bmatrix} C_{11} & C_{12} & O \\ C_{21} & C_{22} & C_{23} \\ O & C_{32} & C_{33} \end{bmatrix}$$

belonging to a class $\mathcal{S}$, can it be written as $A \oplus_1^{\alpha,\beta} B$ with A and B in the same class as C, for $k = 1$?

$\mathcal{Q}3$ Question $\mathcal{Q}1$ but taking $k > 1$.

$\mathcal{Q}4$ Question $\mathcal{Q}2$ but taking $k > 1$.

The positivity matrix classes that were analysed were the following: Positive definite (PD) and positive semidefinite (PSD) matrices, symmetric M matrices (SM), completely positive matrices (CP), double nonnegative matrices(DN), M matrices, P matrices, P_0 matrices, and totally nonnegative matrices (TN).

Regarding the questions $\mathcal{Q}1$ to $\mathcal{Q}4$, we summarized in Table 1 the questions with affirmative answer (see [5] for details) for each class.

5 Conclusions

We have presented the concept of *weighted subdirect sum of matrices* that has been recently published in [5], jointly with some properties for the following matrix classes: Positive definite (PD) and positive semidefinite (PSD) matrices, symmetric M matrices (SM), completely positive matrices (CP), double nonnegative matrices(DN), M matrices, P matrices, P_0 matrices, and totally nonnegative matrices (TN).

References

[1] Bru, R., Pedroche, F., Szyld, D. B., Subdirect sums of nonsingular M-matrices and of their inverses, *Electron. J. Linear Algebra*, 13:162–174, 2005.

[2] Bru, R., Pedroche, F., Szyld, D. B., Subdirect sums of S-Strictly Diagonally Dominant matrices, *Electron. J. Linear Algebra,* 15: 201–209, 2006.

[3] Fallat, S. M., Johnson, C. R., Sub-direct sums and positivity classes of matrices. *Linear Algebra and its Applications*, 288:149–173, 1999.

[4] Pedroche, F. Subdirect sums of matrices. Definitions, methodology and known results. In: Torregrosa JR. *et al* eds. *Modelling for Engineering & Human Behaviour 2021.* I.U. de Matemàtica Multidisciplinar, Universitat Politècnica de València. Spain. 2021: 180–185.

[5] Pedroche, F. Weighted subdirect sum of matrices: Definition and properties for positivity classes, *Mathematical Methods in the Applied Sciences.* 2022;1-22. doi: 10.1002/mma.8787

[6] Wu, K., Taki, Y., Sato, K., Sassa, Y., Inoue, K., Goto, R., Okada, K., Kawashima, R., He, Y., Evans, A. C., Fukuda, H., The Overlapping Community Structure of Structural Brain Network in Young Healthy Individuals *PLoS ONE* 6(5): e19608, 2011.

35

A new iterative inverse display model

M.J. Pérez-Peñalver[◇], S.-W.Lee[△], C. Jordán[♮], E. Sanabria-Codesal[◇] and S. Morillas[♭,1]

(◇) Departamento de Matemática Aplicada, Universitat Politècnica de València
(△) Department of Information Display and Advanced Display Research Center,
Kyung Hee University.
(♮) Intituto de Matemática Multidisciplinar, Universitat Politècnica de València.
(♭) Instituto de Matemática Pura y Aplicada, Universitat Politècnica de València.

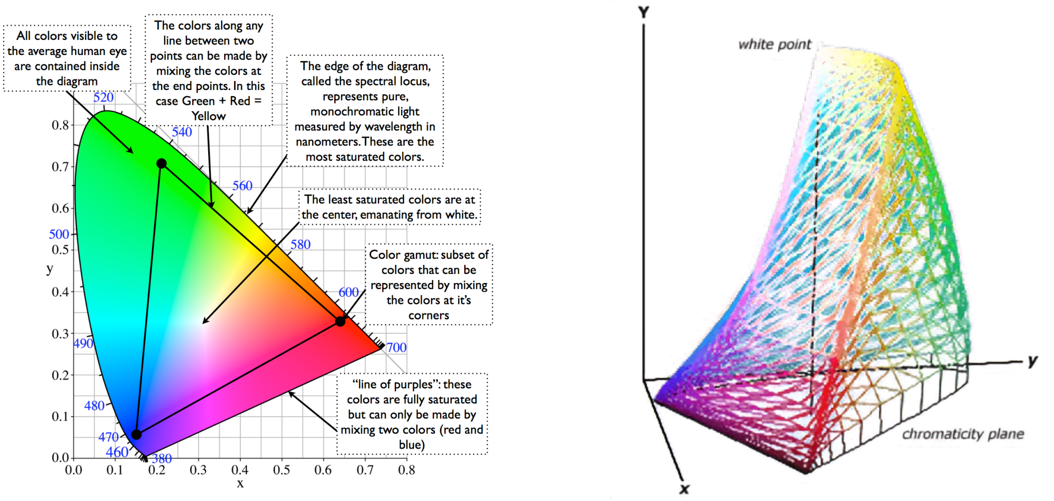
1 Abstract

In this manuscript we propose a new inverse model for a display based on the direct model developed in [6]. We use an iterative method to compute what inputs are able to produce a desired color. Whereas this iterative approach has been used in the past, we take advantage here of the model [6] to design a method able to converge in few iterations yielding a very accurate result.

2 Introduction

Display characterization has been an important topic in the field of color imaging for years, gaining interest with the recent availability increase of a variety of display technologies. Having a precise display characterization allows accurate image reproduction. This is important not only from the consumer point of view in order to optimize viewing experience, but, also, it is critical for any color imaging application, where accurate image reproduction is paramount, and for vision science research to precisely control the stimuli shown to observers when carrying out psychophysics. In general, display characterization means to create a model able to relate device dependent DACRGB inputs with display outputs expressed in an appropriate device independent color space (usually tristimulus values $XY\ Z$ or xyY color coordinates). A colorimeter, or even better a spectrophotometer or spectroradiometer, is used to measure outputs related to a set of $DACRGB$ inputs. Some years ago, usually, the measurements needed to be taken manually. This led to the development of display models based on their physical behavior and that could be built using just a few measurements, which was convenient. Examples of these models are those in [1, 2] that are based on representing the nonlinear behaviour of inputs with some power function or look-up-table and then relate the linearized inputs with the outputs through a simple linear transformation (and so assuming constant chromaticity for them). An alternative model presented [3, 4] is based on using measurements in xyY coordinates and processing separately the Y component with a power function of the inputs and the xy chromaticity using linear interpolation between measurements. On the other hand, mathematical models have been also developed. One of the classical ones is based on trying to find the best linear application able to relate inputs (or some function of them

[1] smorillas@mat.upv.es


All colors visible to the average human eye are contained inside the diagram
The colors along any line between two points can be made by mixing the colors at the end points. In this case Green + Red = Yellow
The edge of the diagram, called the spectral locus, represents pure, monochromatic light measured by wavelength in nanometers. These are the most saturated colors.
The least saturated colors are at the center, emanating from white.
Color gamut: subset of colors that can be represented by mixing the colors at it's corners
"line of purples": these colors are fully saturated but can only be made by mixing two colors (red and blue)
520
540
560
580
600
700
500
490
480
470
460
380
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
y
x
Y
white point
y
chromaticity plane
x

a) $xyCIE1931$ b) xyY color space.

Figure 1: Color spaces $xyCIE1931$ (https://dot-color.com/) and xyY space adapted from a wire-frame animation © 2007 Bruce Lindbloom.

like powers or square roots) and outputs (or some function of them) [5]. To find this linear application an overdetermined equation system is formulated using the measurements taken and the best solution is found by squared error minimization using Penrose pseudo-inverse [7], which is the reason why this method is named the Pseudo-Inverse method.

3 Introductory concepts

RGB is a color representation space based on additive synthesis with which it is possible to represent a color by the addition mixing of the three primary light colors. It is a device-dependent color space, that is, the same *RGB* values may display markedly different colors on different displays using this color space. The *RGB* model is the one used by most devices. It is a three-dimensional space where each color is represented by a vector (r,g,b) with each coordinate belonging to the naturals from 0 to 255 (in the common case of 8-bit representation). Each color is a combination of the three pure colors: red $(255, 0, 0)$, green $(0, 255, 0)$, and blue $(0, 0, 255)$.

In 1931, the Commission International de l'Eclairage (CIE) established a frame of reference to define colors univocally, based on their wavelengths. The (X, Y, Z) CIE tristimulus values can describe any color visible by the human eye. It is a device independent color space. The XYZ values are transformed into two x and y values known as *chromatic coordinates* through the expressions,

$$x = \frac{X}{X+Y+Z} \qquad y = \frac{Y}{X+Y+Z}$$

These chromatic coordinates correspond to the colors that human eye sees and are represented on a plane $xyCIE1931$ (see Figure 1). This is complemented by the Y value which gives the color what is known as luminance, represented in the xyY color space (see Figure 1). It is important to note that although all the colors seen by the human eye are represented in the color tongue on the xy space, the different devices can only represent a part of these colors, namely those contained in the triangle in the Figure 1.

4 Methods

As we said in the introduction we need a display characterization, i.e. models that, given a specific monitor, allow us, on the one side, to predict the (X, Y, Z) values corresponding to some given RGB values (direct model), and on the other side, approximate the RGB inputs me should use to show a certain color expressed in device-independent coordinates (XYZ, xyY or other)(inverse model). In this paper we focus our interest on the last question.

4.1 Direct model our proposal is based on

The direct method developed by Kim and Lee [3] estimates a variable of light leakage from the screen through the color filters to calculate what they call pure XYZ colors from the XYZ measurements of the colorimeter. This model has the advantage of predicting the colorimetric information for all types of screens with high accuracy, which is not the case with other direct models. We use these set of pure colors in our inverse model.

4.2 Inverse model

Since given a physical color with coordinates (X, Y, Z), our objective is to determine which RGB coordinates give an approximation of XYZ from the ramps of RGB pure colors, we need to solve a system that has no analytical solution, so we use an iterative method. As the space of solutions to explore in the method is very large and for image reproduction a high number of coordinates need to be computed, we need to be very efficient (a few number of iterations) in order to be useful in practice. We have used certain heuristics when performing the iterations that has turned out to be very fast.

Since we will need it during the application of the algorithm, the first step is to determine what are the values of X, Y, Z corresponding to the RGB values of pure red, green and blue, obtained using a colorimeter and the ramps of RGB given in [3], i.e., of the respective RGB values: $(i, 0, 0), (0, i, 0)$, $(0, 0, i)$ and (i, i, i) where $i \in \{0, 1, \cdots, 255\}$.

We begin with the study of the chromaticity. We draw on the plane $xyCIE1931$ of chromatic coordinates a triangle T_0 whose vertices, R, G and B, are the corresponding x and y values of the RGB coordinates of each of the most saturated three pure colors, that is, those whose RGB coordinates are $(255, 0, 0)$ (for red), $(0, 255, 0)$ (for green) and $(0, 0, 255)$ (for blue).

Let (X, Y, Z) be the device-independent coordinates of certain color. Let (x, y) be its chromatic coordinates that we represent as the point P on the space $xyYCIE1931$. We now obtain its orthogonal projections, P_R, P_G and P_B, onto the sides of the triangle T_0. (see Figure 2). That is

$$\begin{array}{l} P_R = t_B^{R,1}B + t_G^{R,1}G \\ P_G = t_B^{G,1}B + t_R^{G,1}R \\ P_B = t_R^{B,1}R + t_G^{B,1}G \end{array} \quad \text{with} \quad t_B^{R,1} + t_G^{R,1} = 1, t_B^{G,1} + t_R^{G,1} = 1 \text{ and } t_R^{B,1} + t_G^{B,1} = 1.$$

Now, as an initial iteration of the RGB searched for, we consider (R_1, G_1, B_1) where:

$$R_1 = rpi(\tfrac{t_R^{G,1}+t_R^{B,1}}{2} \cdot 255) \qquad G_1 = rpi(\tfrac{t_G^{R,1}+t_G^{B,1}}{2} \cdot 255) \qquad B_1 = rpi(\tfrac{t_B^{G,1}+t_B^{R,1}}{2} \cdot 255)$$

where rpi means the nearest integer.

We obtain an approximation $P(1)$ of P looking for, among the values previously obtained, the (X, Y, Z) values corresponding to the pure RGB: $(R_1, 0, 0), (0, G_1, 0)$ and $(0, 0, B_1)$ and suming the corresponding coordinates:

$$\begin{array}{l} X(P(1)) = X(R_1, 0, 0) + X(0, G_1, 0) + X(0, 0, B_1) \\ Y(P(1)) = Y(R_1, 0, 0) + Y(0, G_1, 0) + Y(0, 0, B_1) \\ Z(P(1)) = Z(R_1, 0, 0) + Z(0, G_1, 0) + Z(0, 0, B_1) \end{array}$$

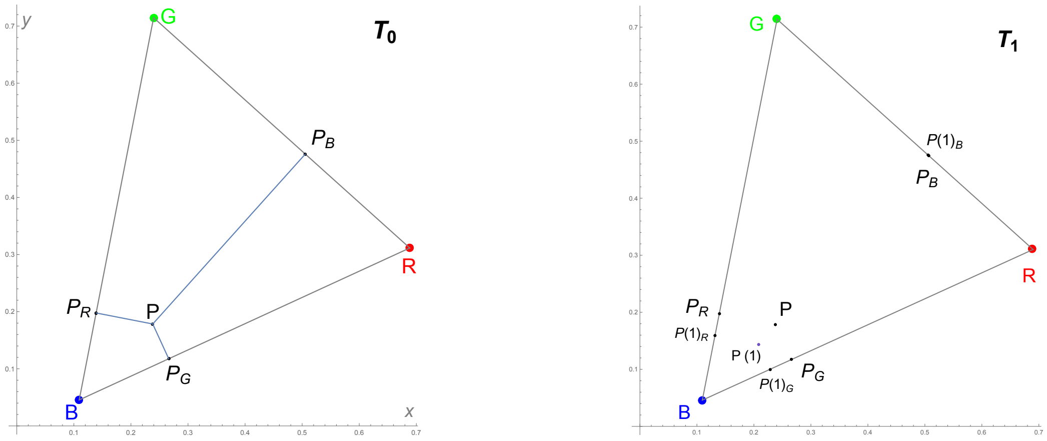


Figure 2: Plane $xyYCIE1931$, points P and $P(1)$ and its orthogonal projections.

From these we obtain the $xyYCIE1931$ coordinates of $P(1)$, a point that approximates P with the above RGB values.

For the next iteration we consider, on the space $xyCIE1931$, a new triangle T_1 whose vertices, called again R,G, and B for simplicity, are the chromatic coordinates of the new $(R_1, 0, 0)$ $(0, G_1, 0)$ and $(0, 0, B_1)$. Now we project orthogonally P and $P(1)$ on the sides of the triangle T_1 obtaining respectively P_R, P_G and P_B, and $P(1)_R$, $P(1)_G$ and $P(1)_B$.

$$\begin{array}{ll} P_R = r_B^R B + r_G^R G \qquad\qquad & P(1)_R = s_B^R B + s_G^R G \\ P_G = r_B^G B + r_R^G R & P(1)_G = s_B^G B + s_R^G R \\ P_B = r_R^B R + r_G^B G & P(1)_B = s_R^B R + s_G^B G \end{array}$$

We can see points P, $P(1)$ and its projections in Figure 2.

Iterating, we reach a triangle T_n whose vertices, called again R,G, and B for simplicity, are the chromatic coordinates of $(R_n, 0, 0)$ $(0, G_n, 0)$ and $(0, 0, B_n)$. We project P and $P(n)$ on the sides of the triangle T_n obtaining respectively P_R, P_G and P_B, and $P(n)_R$, $P(n)_G$ and $P(n)_B$. Their values are:

$$\begin{array}{ll} P_R = r_B^R B + r_G^R G \qquad\qquad & P(n)_R = s_B^R B + s_G^R G \\ P_G = r_B^G B + r_R^G R & P(n)_G = s_B^G B + s_R^G R \\ P_B = r_R^B R + r_G^B G & P(n)_B = s_R^B R + s_G^B G \end{array}$$

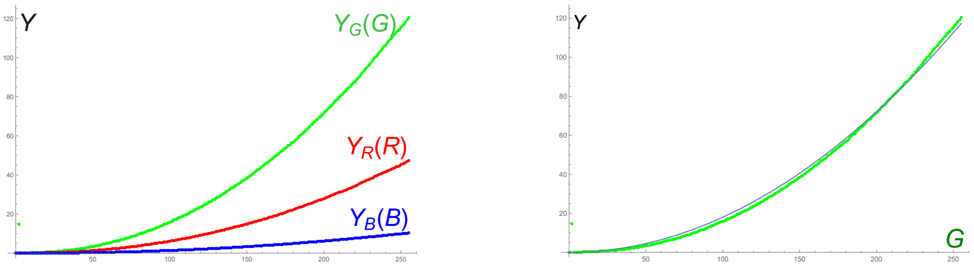
In order to find a better approximation we can modified, for example, the values t by adding to the its last value the half differences between the proportions of the projections of P and $P(n)$:

$$\begin{array}{lll} t_B^{R,n+1} = t_B^{R,n} + \frac{1}{2}\left(r_B^R - s_B^R\right) \qquad & t_B^{G,n+1} = t_B^{G,n} + \frac{1}{2}\left(r_B^G - s_B^G\right) \qquad & t_R^{B,n+1} = t_R^{B,n} + \frac{1}{2}\left(r_R^B - s_R^B\right) \\ t_G^{R,n+1} = t_G^{R,n} + \frac{1}{2}\left(r_G^R - s_G^R\right) & t_R^{G,n+1} = t_R^{G,n} + \frac{1}{2}\left(r_R^G - s_R^G\right) & t_G^{B,n+1} = t_G^{B,n} + \frac{1}{2}\left(r_G^B - s_G^B\right) \end{array}$$

Similarly to what we did in the first iteration we obtain new values RGB $(R_{n+1}, 0, 0)$, $(0, G_{n+1}, 0)$ and $(0, 0, B_{n+1})$:

$$R_{n+1} = rpi\left(\frac{t_R^{G,n+1} + t_R^{B,n+1}}{2} \cdot 255\right) \qquad G_{n+1} = rpi\left(\frac{t_G^{R,n+1} + t_G^{B,n+1}}{2} \cdot 255\right) \qquad B_{n+1} = rpi\left(\frac{t_B^{G,n+1} + t_B^{R,n+1}}{2} \cdot 255\right)$$

By using the ramps obtained at the beginning of the subsection we obtain the pure values (X, Y, Z) associated to them. By summing them we obtain the (X, Y, Z) coordinates of a new approximation



Luminance of red, green and blue

Aproximation of green luminance

Figure 3: Luminance

of the point P. We compute its corresponding chromatic coordinates and obtain a new point $P(n+1)$ on the $xyCEI1931$ space.

Up to now, we have estudied the chromaticity. The next step is the study of luminance.

From the initial coordinates of P we get its luminance Y. We consider now that we have reached the iteration n, and got the RGB values: (R_n, G_n, R_n). Due to the proportionality between the R, G, and B values and since the luminance of a point with coordinates (R_n, G_n, R_n) are obtained by summing the luminance of $(R_n, 0, 0), (0, G_n, 0), (0, 0, R_n)$ we can say that,

$$(R_{n+1}, G_{n+1}, R_{n+1}) = K\,(R_n, G_n, R_n) \quad \text{and} \quad Y_n = Y(R_n) + Y(G_n) + Y(B_n)$$

for some constant K. Then we must choose a suitable constant K.

The relationship of the luminance of red, green and blue dates of our display is represented in Figure 3. We see that the color that provides the most luminance is green and the graphic of the green has a approximate parabola shape that an approximation by least squares gives $Y(G) = 0.001832G^2$ (see Figure 3). Then $G(Y) = \sqrt{\frac{Y}{0.001832}}$ and if $G_{n+1} = KG_n$, we then take

$$K = \frac{G_{n+1}}{G_n} \sim \frac{G_n + \alpha\,(G(Y) - G(Y_n))}{G_n} = 1 + \frac{\alpha}{G_n}\,(G(Y) - G(Y_n))\,.$$

We consider $\alpha = \frac{1}{2}$ and then

$$K = 1 + \frac{1}{2G_n}\left(\sqrt{\frac{Y}{0.001832}} - \sqrt{\frac{Y_n}{0.001832}}\right) \qquad \text{and}$$

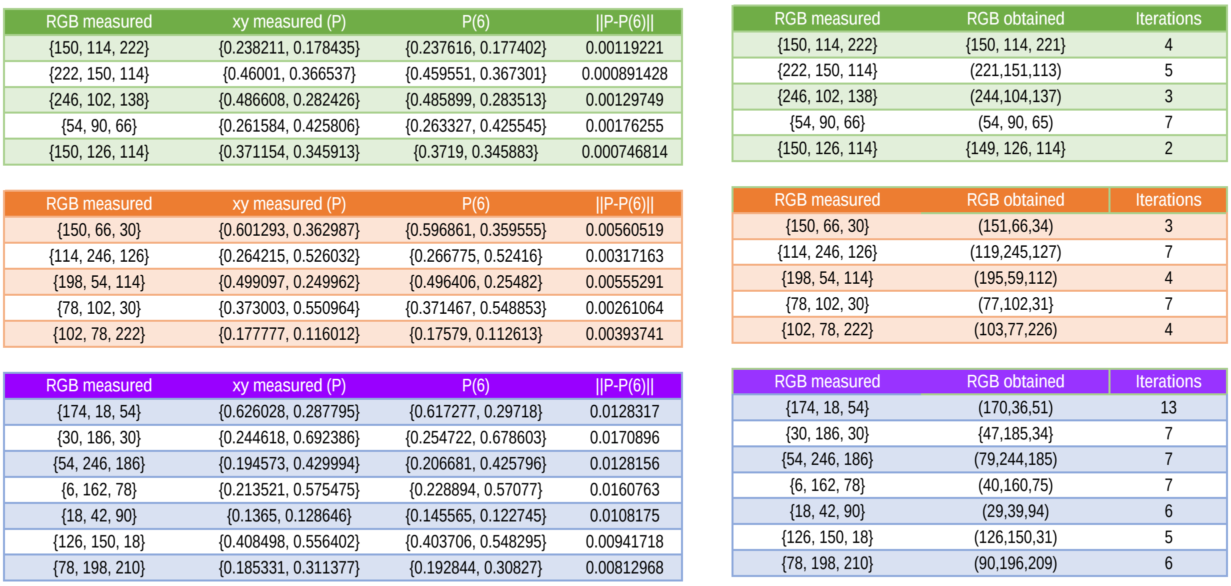
$$R_{n+1} = rpi(K \cdot R_n) \qquad G_{n+1} = rpi(K \cdot G_n) \qquad B_{n+1} = rpi(K \cdot B_n)$$

We go iterating until they stabilize.

5 Results

To discriminate the advantages or disadvantages of the method we will measure with the colorimeter 9260 RGB coordinates. We will apply the described method, using different stopping criteria, to the corresponding 9260 coordinates (X, Y, Z), and compare the obtained results with the initial RGB values.

A smaller number of input data has generated good preliminary results of the chromaticity and luminance as we can see in Figure 4.



RGB measured	xy measured (P)	P(6)	||P-P(6)||
{150, 114, 222}	{0.238211, 0.178435}	{0.237616, 0.177402}	0.00119221
{222, 150, 114}	{0.46001, 0.366537}	{0.459551, 0.367301}	0.000891428
{246, 102, 138}	{0.486608, 0.282426}	{0.485899, 0.283513}	0.00129749
{54, 90, 66}	{0.261584, 0.425806}	{0.263327, 0.425545}	0.00176255
{150, 126, 114}	{0.371154, 0.345913}	{0.3719, 0.345883}	0.000746814

RGB measured	RGB obtained	Iterations
{150, 114, 222}	{150, 114, 221}	4
{222, 150, 114}	(221,151,113)	5
{246, 102, 138}	(244,104,137)	3
{54, 90, 66}	(54, 90, 65)	7
{150, 126, 114}	{149, 126, 114}	2

RGB measured	xy measured (P)	P(6)	||P-P(6)||
{150, 66, 30}	{0.601293, 0.362987}	{0.596861, 0.359555}	0.00560519
{114, 246, 126}	{0.264215, 0.526032}	{0.266775, 0.52416}	0.00317163
{198, 54, 114}	{0.499097, 0.249962}	{0.496406, 0.25482}	0.00555291
{78, 102, 30}	{0.373003, 0.550964}	{0.371467, 0.548853}	0.00261064
{102, 78, 222}	{0.177777, 0.116012}	{0.17579, 0.112613}	0.00393741

RGB measured	RGB obtained	Iterations
{150, 66, 30}	(151,66,34)	3
{114, 246, 126}	(119,245,127)	7
{198, 54, 114}	(195,59,112)	4
{78, 102, 30}	(77,102,31}	7
{102, 78, 222}	(103,77,226)	4

RGB measured	xy measured (P)	P(6)	||P-P(6)||
{174, 18, 54}	{0.626028, 0.287795}	{0.617277, 0.29718}	0.0128317
{30, 186, 30}	{0.244618, 0.692386}	{0.254722, 0.678603}	0.0170896
{54, 246, 186}	{0.194573, 0.429994}	{0.206681, 0.425796}	0.0128156
{6, 162, 78}	{0.213521, 0.575475}	{0.228894, 0.57077}	0.0160763
{18, 42, 90}	{0.1365, 0.128646}	{0.145565, 0.122745}	0.0108175
{126, 150, 18}	{0.408498, 0.556402}	{0.403706, 0.548295}	0.00941718
{78, 198, 210}	{0.185331, 0.311377}	{0.192844, 0.30827}	0.00812968

RGB measured	RGB obtained	Iterations
{174, 18, 54}	(170,36,51)	13
{30, 186, 30}	{47,185,34}	7
{54, 246, 186}	(79,244,185)	7
{6, 162, 78}	(40,160,75)	7
{18, 42, 90}	(29,39,94)	6
{126, 150, 18}	(126,150,31)	5
{78, 198, 210}	(90,196,209)	6

Figure 4: First results on chromaticity and luminance.

6 Conclusions

Based on a recent precise direct model for display characterization we have proposed in this paper an iterative method to solve the inverse model problem. The main advantage of the proposed method is its ability to converge in very few iterations to a very accurate result. As future work, we aim at testing the model in a range of different displays and compare extensively with methods in the state of the art.

References

[1] Berns, Methods for Characterizing CRT Displays, Displays (1996): https://doi.org/10.1016/0141-9382(96)01011-6

[2] M. Fairchild, D. Wyble, Colorimetric characterization of the Apple studio display (Flat panel LCD), RIT 1998: http://scholarworks.rit.edu/article/920/

[3] Kim, J.M., Lee, S.W., Universal color characterization model for all types of displays. *Optical Engineering*, 54(10), 103103, 2015. https://doi.org/10.1117/1.oe.54.10.103103

[4] Malo,J., Luque, M.J., ColorLab: the Matlab toolbox for Colorimetry and Color Vision. Univ. Valencia, 2002, http://isp.uv.es/code/visioncolor/colorlab.html

[5] P. Capilla, M. Diez, M.J. Luque, J. Malo, Corresponding-pair procedure: a new approach to simulation of dichromatic color perception. *JOSA A*, 21(2): 176-186, 2004.

[6] Westland, S., Ripamonti, C., Cheung, V. (2012). Computational colour science using MATLAB. John Wiley & Sons.

[7] R. Penrose, On best approximate solution of linear matrix equations, Proc. Camb. Philos. Soc. 52, 17–19 (1956).

36

Application of polynomial algebras to non-linear equation solvers

J. Canela$^\flat$ and D. Pérez-Palau$^{\natural 1}$

($\flat$) Insitut Universitari de Matemàtiques i Aplicacions de Castelló and Departament de Matemàtiques,
Universitat Jaume I,
Av. Vicent Sos Baynat, s/n 12071 Castelló de la Plana, Spain.
($\natural$) Escuela Superior de Ingeniería y Tecnología,
Universidad Internacional de la Rioja (UNIR),
Av. La Paz 137, 26006 Logroño, Spain.

1 Introduction

Polynomial algebras emerged during the decade of the 1990 [3] as a tool to propagate efficiently sets of points around an initial condition x_0. They were first used in the study of particle accelerators.

The main idea of a polynomial algebra approach is to approximate a function $g(x)$ by a polynomial $P(\xi)$ in such a way that $P(\xi) \approx g(x_0 + \xi)$. This technique is usually applied to functions like Poincaré maps, flow maps, etc. The polynomial approximation of functions is equivalent to the floating-point approximation of real values. As it is the case of the floating-point, where the real numbers are truncated up to a given number of digits, the polynomials are truncated up to a given order N. Therefore, Jets (or truncated polynomials) are used to approximate functions. The larger is the order N the better is the approximation of the function.

The use of polynomial arithmetics has been widely extended for the last decades. Some examples of its implementation can be found in [8, 9]. Using such arithmetics, given two truncated series S_a and S_b one can compute the jet of all the classic operations, like products $S_a \cdot S_b$, additions $S_a + S_b$, or divisions S_a/S_b (assuming that the independent term of S_b does not banish), as well as exponentials, logarithms or trigonometric operations, among other. For instance, if $S_a = \sum_{i=0}^{n} a_i \xi^i$ and $S_b = \sum_{i=0} b_i \xi^i$. Then, the coefficients of $S_c = S_a \cdot S_b = \sum_{i=0} c_i \xi^i$ and $S_d = S_a/S_b = \sum_{i=0} d_i \xi^i$ are given by:

$$c_i = \sum_{j=0}^{i} a_{i-j} b_j, \qquad c_n = \frac{a_n - \sum_{j=1}^{n} c_{n-j} b_j}{b_0}. \tag{1}$$

These techniques have been used in several applications. The main one is in the propagation of ordinary differential equations. In [3] the bases of the technique were introduced. Later on, [1,2,10] used it to study and propagate trajectories and its uncertainties in the evolution of spacecrafts and asteroids. In [4, 5] the same ideas were applied in the modification of a Kalman filter and in the orbit determination with uncertainties. From a theoretical point of view, [6] used the jet transport technique to compute periodic orbits of delayed differential equations and [7] computed high order expansions of invariant manifolds around tori of high dimension.

[1] daniel.perez@unir.net

As can be observed in the bibliography, the Jet Transport is widely used in the context of the propagation of differential equations. However, there are other fields where it can be used. One of those applications is in the solution of non-linear equations with parameters.

Let us consider a non-linear equation $f(x) = 0$. Usually, such equation can be solved numerically by means of iterative methods such as Newton's method:

$$x_{k+1} = N(x_k) = x_k - \frac{f(x_k)}{f_x(x_k)}.$$

If the initial seed x_0 is close enough to the solution $x^\star$ of $f(x) = 0$, the iteration converges to $x^\star$.

In the following, let us consider a function f which depends on a parameter c. Hence, we want to solve $f(x, c) = 0$. Therefore, we look for a solution $x(c)$ such that $f(x(c), c) = 0$. It follows from the Implicit Function Theorem that $x(c)$ exists under regularity conditions. We want to find $x(c) = \sum a_i(c_0 - c)^i$ such that $f(x(c), c) = 0$. However, such solution $x(c)$ may require an infinite sum. Hence, we will restrict to a finite order polynomial approximation $x_N(c) = \sum_{i=0}^{N} a_i(c_0 - c)^i$.

In Section 2 we show how to use the jet arithmetics to solve an example of such problem. Let $c = c_0 + \xi$, and assume that x_0 is a solution of $f(x_0, c_0)$. The first step is to apply Newton's method to the function $f(x, c_0 + \xi)$ with the initial seed x_0. We have that

$$x_{k+1}(\xi) = x_k(\xi) - \frac{f(x_k(\xi), c + \xi)}{f_x(x_k(\xi), c + \xi)}.$$

We show that at each step there is an increasing number of coefficients that correctly approximate $x(c)$. In Section 3 we give the main result of the paper. There, we proof that at each iterate of the procedure the number of vanishing coefficients of the Taylor expansion of $f(x_k(\xi), c_0 + \xi)$ doubles. Finally, in Section 5 we give some conclusions.

2 Preliminary example

Given the function $f(x, c) = x^2 + x + c$, where $c \in \mathbb{R}$, We want to find a parametric solution of $f(x, c) = 0$ around $c = c_0 = 0$. Notice that $x = 0$ is a solution of $f(x, c_0) = 0$ ($f(0, 0) = 0$). The exact solution is given by $x(c) = \dfrac{-1 + \sqrt{1 - 4c}}{2}$. The Taylor series of the previous function can be computed by derivation, obtaining

$$x(c_0 + \delta c) = -\delta c - \delta c^2 - 2\delta c^3 - 5\delta c^4 - 14\delta c^5 - 42\delta c^6 - 132\delta c^7 - 429\delta c^8 - 1430\delta c^9 + o(\delta c^{10}).$$

Instead of using derivation, we can obtain the Taylor series of x_c (up to a given order) applying the Newton Polynomial Technique

$$x_{k+1} = x_k - \frac{f(x_k, 0 + \xi)}{f_x(x, 0 + \xi)}.$$

We look for a functional solution of $f(x, c(\xi)) = 0$, where $c(\xi) = c_0 + \xi = 0 + \xi$. We take as initial approximation of the constant solution $x_0(\xi) = 0$, since $f(x_0(\xi), c(0)) = 0$. After iterating the procedure once, we obtain

$$x_1(\xi) = x_0(\xi) - \frac{f(x_0(\xi), 0 + \xi)}{f_x(x_0(\xi), 0 + \xi)} = 0 + \frac{0^2 + 0 + 0 + \xi}{2 \cdot 0 + 1} = -\xi.$$

Repeating the procedure and applying the polynomial algebra formulas 1 to the fraction we get

$$x_2(\xi) = x_1(\xi) - \frac{f(x_1(\xi), 0 + \xi)}{f_x(x_1(\xi), 0 + \xi)} = 0 + \frac{(-\xi)^2 - \xi + 0 + \xi}{2(-\xi) + 1} = -\xi - \xi^2 - 2\xi^2 + o(\xi^3)$$

Counting the constant term, the power series of $x_1(\xi)$ has 2 exact terms. After applying the procedure, $x_2(\xi)$ has 4 exact terms. As we will see, at each step of new Newton Polynomial Technique we double the number of exact terms. Indeed, for $x_3(\xi)$ we get

$$x_3(\xi) = -\xi - \xi^2 - 2\xi^3 - 5\xi^4 - 14\xi^5 - 42\xi^6 - 132\xi^7 + o(\xi^8).$$

3 Main Theorem

In this section we formalize the results discussed up to now. This is the content of Theorem A. Before stating the theorem, we define the concept of polynomial zero of a (sufficiently smooth) function $f(x)$.

Definition 1. *A polynomial $P(\xi)$ is a* **polynomial zero of order** *n of a function f if*

$$f(P(\xi)) = 0 + \mathcal{O}(\xi^n).$$

Theorem A. *Let $f(x, c)$ be a sufficiently smooth function depending on a parameter $c \in \mathbb{R}$. Let $N(x)$ denote the Newton method using polynomial iterates,*

$$P_{k+1}(\xi) = N(P_k(\xi)) = x - \frac{f(P_k(\xi), c_0 + \xi)}{f_x(P_k(\xi), c_0 + \xi)},$$

and let $P_0(\xi) = x_0$ be the initial seed with $f(x_0, c_0) = 0$. Then, the iterate $P_k(\xi)$ is a polynomial zero of order 2^k of $f(x, c + \xi)$.

Proof. We will prove the result by induction. Notice that the (constant) polynomial $P_0(\xi) = x_0$ is a polynomial zero of order $1 = 2^0$. Indeed, by hypothesis we have that $f(x_0, c_0 + \xi) = 0 + \mathcal{O}(\xi)$.

For the sake of clarity, next prove that P_1 is a polynomial zero of order 2^1. We have:

$$\begin{aligned} P_1(\xi) &= P_0(\xi) - \frac{f(P_0(\xi), c_0 + \xi)}{f_x(P_0(\xi), c_0 + \xi)} \\ &= x_0 - \frac{f(x_0, c_0 + \xi)}{f_x(x_0, c_0 + \xi)}. \end{aligned}$$

Evaluating f using its Taylor expansion with respect to x around x_0 we get:

$$\begin{aligned} f(P_1(\xi), c_0 + \xi) =& f\left(x_0 - \frac{f(x_0, c_0 + \xi)}{f_x(x_0, c_0 + \xi)}, c_0 + \xi\right) \\ =& f(x_0, c_0 + \xi) - f_x(x_0, c_0 + \xi)\left(\frac{f(x_0, c_0 + \xi)}{f_x(x_0, c_0 + \xi)}\right) \\ &+ \frac{1}{2} f_{xx}(x_0, c_0 + \xi)\left(-\frac{f(x_0, c_0 + \xi)}{f_x(x_0, c_0 + \xi)}\right)^2 + \mathcal{O}\left(\left(-\frac{f(x_0, c_0 + \xi)}{f_x(x_0, c_0 + \xi)}\right)^3\right). \end{aligned}$$

Using that $f(x_0, c_0 + \xi) = 0 + \alpha_1 \xi + o(\xi)$ we get that

$$f(P_1(\xi), c_0 + \xi) = 0 + \mathcal{O}(\xi^2).$$

We conclude that $P_1(\xi)$ is a polynomial zero of order $2 = 2^1$.

We will use induction to prove the general case. The **induction hypothesis** is that the polynomial P_k is a polynomial zero of order 2^k, that is:

$$f(P_k(\xi), c_0 + \xi) = \mathcal{O}\left(\xi^{2^k}\right).$$

The polynomial P_{k+1} is given by

$$P_{k+1}(\xi) = P_k(\xi) - \underbrace{\frac{f(P_k, c_0+\xi)}{f_x(P_k, c_0+\xi)}}_{\tilde{P}_{k+1}(\xi)}.$$

Evaluating $f(x, c_0+\xi)$ at P_{k+1} and using the Taylor expansion of $f(x, c_0+\xi)$ with respect to x around $P_k(\xi)$ we get

$$\begin{aligned} f(P_{k+1}(\xi), c_0+\xi) =& f\left(P_k(\xi) - \tilde{P}_{k+1}(\xi), c_0+\xi\right) \\ =& f(P_k(\xi), c_0+\xi) - f_x(P_k(\xi), c_0+\xi)\frac{f(P_k(\xi), c_0+\xi)}{f_x(P_k, c_0+\xi)} \\ &+ \frac{1}{2} f_{xx}(P_k(\xi), c_0+\xi)(-\tilde{P}_{k+1}(\xi))^2 + \mathcal{O}\left(\left(\tilde{P}_{k+1}(\xi)\right)^3\right). \end{aligned}$$

By induction hypothesis we have $f(P_k, c_0+\xi) = \mathcal{O}\left(\xi^{2^k}\right)$ and, therefore

$$\tilde{P}_{k+1}(\xi) = \frac{f(P_k, c_0+\xi)}{f_x(P_k, c_0+\xi)} = \mathcal{O}\left(\xi^{2^k}\right).$$

Finally, we get

$$\begin{aligned} f(P_{k+1}(\xi), c_0+\xi) =& \frac{1}{2} f_{xx}(P_k, c_0+\xi)\left(\frac{f(P_k(\xi), c_0+\xi)}{f_x(P_k(\xi), c_0+\xi)}\right)^2 + \mathcal{O}\left(\left(\tilde{P}_{k+1}(\xi)\right)^3\right) \\ =& \frac{1}{2} f_{xx}(P_k(\xi), c_0+\xi)\left(\mathcal{O}\left(\xi^{2^k}\right)\right)^2 + \mathcal{O}\left(\left(\mathcal{O}\left(\xi^{2^k}\right)\right)^3\right) \\ =& \mathcal{O}\left(\xi^{2^{k+1}}\right). \end{aligned}$$

■

4 Conclusions

This paper studies methods for solving non-linear equations using a polynomial algebra. The method computes the solution, $x(c_0+\xi)$, of the non-linear equation $f(x, c_0+\xi) = 0$. This is done by approximating $x(c_0+\xi)$ as a jet, $P(\xi)$, and working with polynomial arithmetics in a Newton's method scheme. The main theorem of the paper shows that each iterate of the procedure, $P_k(\xi)$, multiplies by two the coefficients that vanish in $f(P_k(\xi), c_0+\xi)$.

References

[1] Elisa Maria Alessi, Ariadna Farres, Arturo Vieiro, Angel Jorba, and Carles Simó. Jet transport and applications to neos. In *Proceedings of the 1st IAA Planetary Defense Conference, Granada, Spain*, pages 10–11, 2009.

[2] Roberto Armellin and Pierluigi Di Lizia. Probabilistic optical and radar initial orbit determination. *Journal of Guidance, Control, and Dynamics*, 41(1):101–118, 2018.

[3] Martin Berz. *Modern map methods in particle beam physics.* Academic Press, 1999.

[4] Jianlin Chen, Josep Masdemont, Gerard Gomez, and Jianping Yuan. Analysis of jet transport-based geostationary trajectory uncertainty propagation. *Journal of Guidance, Control, and Dynamics*, 43, 04 2020.

[5] Jianlin Chen, Josep Masdemont, Gerard Gomez, and Jianping Yuan. An efficient statistical adaptive order-switching methodology for kalman filters. *Communications in Nonlinear Science and Numerical Simulation*, 93, 09 2020.

[6] Joan Gimeno and Angel Jorba. Using automatic differentiation to compute periodic orbits of delay differential equations. *Discrete & Continuous Dynamical Systems - B*, 22, 01 2017.

[7] Joan Gimeno, Angel Jorba, Begoña Nicolás, and Estrella Olmedo. Numerical computation of high-order expansions of invariant manifolds of high-dimensional tori. *SIAM Journal on Applied Dynamical Systems*, 21:1832–1861, 09 2022.

[8] Àlex Haro, Marta Canadell, Jordi-Lluis Figueras, Alejandro Luque, and Josep María Mondelo. *The Parameterization Method for Invariant Manifolds: From rigorous results to effective computations.* Springer, 2016.

[9] Angel Jorba and Maorong Zou. A software package for the numerical integration of odes by means of high-order taylor methods. *Experimental Mathematics*, 14(1):99–117, 2005.

[10] Monica Valli, Roberto Armellin, Pierluigi Di Lizia, and Michèle R Lavagna. Nonlinear filtering methods for spacecraft navigation based on differential algebra. *Acta Astronautica*, 94(1):363–374, 2014.

37

A Linear Quadratic Tracking Problem for Impulsive Controlled Stochastic Systems. The Infinite Horizon Time Case

V. Drğan^{a,b}, I-L. Popa^{c,d,1} and I.Ivanov^{e}

(a) Institute of Mathematics ”Simion Stoilow” of the Romanian Academy, P.O.Box 1-764,
RO-014700, Bucharest, Romania

(b) Academy of the Romanian Scientists, Bucharest, Romania

(c) Department of Computing, Mathematics and Electronics,
”1 Decembrie 1918” University of Alba Iulia, Alba Iulia, 510009 , Romania

(d) Faculty of Mathematics and Computer Science, Transilvania University of Braşov,
Iuliu Maniu Street 50, 500091 Braşov, Romania

(e) Faculty of Economics and Business Administration, Sofia University ”St. Kl. Ohridski”,
125 Tzarigradsko chaussee blvd., bl. 3, Sofia 1113, Bulgaria

1 Introduction

A modern approach to investigate the linear quadratic tracking control is the reinforcement learning approach [5–7]. An online learning algorithm to investigate the linear quadratic tracking problem for partially-unknown continuous-time systems is developed in [7]. A tracking algebraic Riccati equation to solve the problem is derived. The reinforcement of on-line learning algorithm is applied to the Riccati equation. The solution of optimal tracking of nonlinear affine and non-affine systems is investigated in [6]. The optimal tracking control problem for linear systems subject to multiple false-data-injection attacks is considered in [5]. This is considered as a problem for solving an algebraic game theoretic Riccati equation. A Q-learning algorithm to solve the Riccati equation is proposed. It is worth mentioning that the above studies have been done for time-invariant control systems. Generally speaking, in practical examples, as mentioned above, various factors as temperature varying, disturbance, unpredictable factors, lead us to systems that should be time-varying. This is why, recently many researchers considered the time-varying case. See for example [3], [9], [11], and the reference therein.

Moreover, the subject and applications of the backward jump matrix Riccati type differential equations is intensively analysed by researches. The backward jump matrix linear differential equations with Riccati type jump is a strong tool to investigate the control problems [1,2,4,8]. The solvability of this type of the Riccati equation which is related with an optimal control problem with random coefficients is analysed in [10]. The optimal control problem has an unique optimal control in an affine state feedback form if and only if a jump Riccati type matrix differential equation admits a stabilizing solution.

In this paper we consider a stochastic system described by Itô differential equations controlled by impulses. The paper's aim is to design a control $u(t)$ to minimize the deviation of the controlled

[1] lucian.popa@uab.ro

output $z(t)$ from the reference signal $r(t)$ under the given performance criterion. We prove the equivalence between the solvability of a stochastic linear differential equation with jumps and the solvability of a discrete-time linear equation. A definition of the concept of the mean square stabilizability by impulses is introduced and applied in the investigation. We consider a backward jump matrix linear differential equation (BJMLDE) with impulses driven by Riccati type operator as a power tool of investigation.

We derive necessary and sufficient conditions which guarantee the existence of the bounded and stabilizing solution of the BJMLDE. We prove several properties of the solution. Finally, we derive the explicit formula of the optimal control which solves the introduced optimal control problem.

Let us consider the controlled system described by:

$$dx(t) = A_0(t)x(t)dt + A_1(t)x(t)dw(t), \;\; kh < t \leq (k+1)h \tag{1a}$$

$$x(kh^+) = A_{d0}(k)x(kh) + B_{d0}(k)u(k) + w_d(k)(A_{d1}(k)x(k) + B_{d1}(k)u(k)), \; k \in \mathbb{Z}_+ = \{0, 1, \ldots\}, \tag{1b}$$

$$z(t) = C(t)x(t), \;\; t \in \mathbb{R}_+ = [0, \infty), \tag{1c}$$

where $x(t) \in \mathbb{R}^n$ are the state parameters, $z(t) \in \mathbb{R}^{n_z}$ is the controlled output and $u(k) \in \mathbb{R}^m$ are the control parameters; $\{w(t)\}_{t\geq 0}$ is an 1-dimensional standard Wiener process defined on the probability space $(\Omega, \mathfrak{F}, \mathcal{P})$, $\{w_d(k)\}_{k\in\mathbb{Z}_+}$ is a sequence of independent random variable with zero mean and variance 1. The system (1a)-(1b) is a system of Itô differential equations controlled by impulses, where the period between two impulses is $h > 0$, fixed.

Let $r(\cdot) : \mathbb{R}_+ \to \mathbb{R}^{n_z}$ be a bounded and continuous vector valued function. In the sequel, $r(\cdot)$ will be called "reference signal" or simply "reference". Roughly speaking, our aim is to design a control $u(k)$, $k \in \mathbb{Z}_+$, which to minimize the derivation of the controlled output $z(t)$ from the reference signal $r(t)$. For a rigorous setting of the output problem, let us describe the class of admissible controls and to introduce a performance criterion. First, we make the following assumption referring to the noises which affect the coefficients of the system (1).

(A1)

(a) The Wiener process, $w(t)$, $t \geq 0$ has the properties: $w(0) = 0$, $\mathbb{E}[w(t)] = 0$, $\mathbb{E}[w^2(t)] = t$, for all $t > 0$.

(b) The sequence of random variables $\{w_d(k)\}_{k\geq 0}$ has the properties: $\mathbb{E}[w_d(k)] = 0$, $\mathbb{E}[w_d(k)w_d(l)^\top] = \delta_{kl}$, for all $k, l \in \mathbb{Z}_+$, where δ_{kl} being the Kronecker symbol.

(c) $\{w(t)\}_{t\geq 0}$ and $\{w_d(k)\}_{k\in\mathbb{Z}_+}$ are independent stochastic processes.

Here and in the sequel $\mathbb{E}[\cdot]$ stands for the mathematical expectation. For each $t > 0$, $\mathcal{F}_t \subset \mathfrak{F}$ is the $\sigma-$algebra generated by the random variables $w(s)$, $w_d(k)$ for all $0 < s \leq t$, $k \in \mathbb{Z}_+$ such that $kh < t$ augmented with all events $\Theta \in \mathfrak{F}$ such that $\mathcal{P}(\Theta) = 0$.

For each $x_0 \in \mathbb{R}^n$, the set of admissible controls $\mathcal{U}_{ad}(x_0)$ consists of all sequences $\mathbf{u} = \{u(k)\}_{k\in\mathbb{Z}_+}$ where for each k, $u(k) : \Omega \to \mathbb{R}^m$ is a random vector $\mathcal{F}_{kh}-$measurable and, additionally we have

$$\sup_{k\in\mathbb{Z}_+} \mathbb{E}[|u(k)|^2] < \infty \tag{2a}$$

$$\sup_{t\geq 0} \mathbb{E}[|x(t; x_0, \mathbf{u})|^2] < \infty, \tag{2b}$$

with $x(\cdot; x_0, \mathbf{u})$ being the solution of the stochastic linear differential equation with jumps (JSLDE) (1a)-(1b) corresponding to the input $\mathbf{u}$ and satisfying $x(0; x_0, \mathbf{u}) = x_0$. To measure the quality of the tracking of the reference signal $r(t)$ by the controlled output $z(t)$ of the system (1) determined by an admissible control $\mathbf{u}$ we introduce the performance criterion:

$$J(x_0;\mathbf{u}) = \lim_{n\to\infty} \sup \left(\frac{1}{Nh} \int_0^{Nh} \mathbb{E}[|z(t;x_0,\mathbf{u}) - r(t)|^2]dt + \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}[u^\top(k)R(k)u(k)] \right) \tag{3}$$

where $z(\cdot;x_0,\mathbf{u}) = C(\cdot)x(\cdot;x_0,\mathbf{u})$ is the controlled output of the system (1) determined by the control $\mathbf{u} \in \mathcal{U}_{ad}(x_0)$. The optimal control problem which we want to solve in this work ask for finding an admissible control $\tilde{\mathbf{u}} \in \mathcal{U}_{ad}(x_0)$ with the property that

$$J(x_0,\tilde{\mathbf{u}}) = \min_{\mathbf{u}\in\mathcal{U}_{ad}(x_0)} \{J(x_0;\mathbf{u})\} \tag{4}$$

2 A class of backward jump matrix linear differential equations with Riccati type jumping operators

Based on the matrices involved in (1) and (3) we consider the following backward jump matrix linear differential equation (BJMLDE) on the Hilbert space $\mathcal{S}_n$:

$$-\dot{Y}(t) = A_0^\top(t)Y(t) + Y(t)A_0(t) + A_1^\top(t)Y(t)A_1(t) + \frac{1}{h}C^\top(t)C(t), \ \ kh \le t < (k+1)h \tag{5a}$$

$$\begin{aligned} Y(kh^-) = {}& \sum_{j=0}^{1} A_{dj}^\top(k)Y(kh)A_{dj}(k) - \left(\sum_{j=0}^{1} A_{dj}^\top(k)Y(kh)B_{dj}(k) \right) \cdot \\ & \cdot \left(R(k) + \sum_{j=0}^{1} B_{dj}^\top(k)Y(kh)B_{dj}(k) \right)^{-1} \left(\sum_{j=0}^{1} B_{dj}^\top(k)Y(kh)A_{dj}(k) \right), \ \ k \in \mathbb{Z}_+. \end{aligned} \tag{5b}$$

One sees that the right hand side of (5b) has the form of a operator arising in a discrete-time Riccati equation. That is way in the sequel (5) will be named BJMLDE with Riccati type jumping operator.

In this work we are interested in solutions $Y(\cdot) : \mathbb{R}_+ \to \mathcal{S}_n$ of (5) which are bounded functions and satisfy a condition of the form:

$$\inf_{k\in\mathbb{Z}_+} \left| \det \left(R(k) + \sum_{j=0}^{1} B_{dj}^\top(k)Y(kh)B_{dj}(k) \right) \right| > 0. \tag{6}$$

To derive the optimal control of the problem of the optimal control (4) a crucial role will be played by the bounded and stabilizing solution of (5).

We give a proof that the the BJMLDE (5) has a bounded and stabilizing solution.

3 Results

We present our approach to solve the tracking problem. We derive the explicit formula of the control $\tilde{\mathbf{u}} \in \mathcal{U}_{ad}(x_0)$ which solves the optimal control problem (4). The main tools are the bounded and stabilizing solution of the BJMLDE with Riccati type jumping operator (5) and the unique bounded on $\mathbb{R}_+$ solution of the following BJLDE:

$$-\dot{\varphi}(t) = A_0^\top(t)\varphi(t) + \frac{1}{h}C^\top(t)C(t), \ \ kh \le t < (k+1)h \tag{7a}$$

$$\varphi(kh^-) = (A_{d0}(k) + B_{d0}(k)\tilde{F}(k))^\top \varphi(kh), \ \ k \in \mathbb{Z}_+, \tag{7b}$$

where $\{\tilde{F}(k)\}_{k\geq 0}$ is associated to the bounded and stabilizing solution $\tilde{Y}(\cdot)$ of (5)

The main result of this work. It provides the control which solves the optimization problem (4).

Theorem 15. *Assume:*

*(a) the assumptions **(A1)** and **(A2)** hold;*

(b) the BJMLDE with Riccati type jumping operator (5) has a bounded and stabilizing solution $\tilde{Y}(\cdot)$.

Under these conditions for each $x_0 \in \mathbb{R}^n$, the optimal control problem which is requiring the minimization of the cost function (3) along the trajectories of the controlled system (1) determined by the set of admissible controls $\mathcal{U}_{ad}(x_0)$ has a unique optimal control $\tilde{\mathbf{u}} = \{\tilde{u}(k)\}_{k\in\mathbb{Z}_+}$ The minimal value of the cost function is

$$J(x_0;\tilde{\mathbf{u}}) = \lim_{N\to\infty}\sup\left(\frac{1}{Nh}\int_0^{Nh}|r(t)|^2dt - \frac{1}{N}\sum_{k=0}^{N-1}\tilde{\varphi}^\top(kh)B_{d0}(k)\Pi_d^{-1}(k,\tilde{Y}(kh))B_{d0}^\top(k)\tilde{\varphi}(kh)\right). \quad (8)$$

4 Conclusions

In this paper the linear quadratic tracking problem has been investigated for a class of impulsive controlled stochastic systems. The infinite horizon case is taken into account. First, the concept of mean square stability by impulses is introduced. Then, the proposed linear quadratic tracking problem was transformed to a special optimal control problem for a backward jump matrix linear differential equations with Riccati type jumping operators. Finally, the explicit solution of the tracking problem was derived based on the bounded and stabilizing solution of the considered optimal control problem. In addition, for the particular cases of periodic systems as well as time-invariant systems the explicit formula of the tracking problem in derived.

It is worth mentioning that our approach can be extended to the finite/infinite horizon tracking problems with preview for impulsive controlled stochastic systems.

References

[1] S. Chen, X. Li, X. Y. Zhou, Stochastic linear quadratic regulators with indefinite control weight costs, SIAM J. Control Optim., 36 (1998) 1685–1702.

[2] K. Du, Solvability conditions for indefinite linear quadratic optimal stochastic control problems and associated stochastic Riccati equations, SIAM J. Control Optim., 53 (2015) 3673–3689.

[3] N. Jin, S. Liu, State feedback control for stochastic regular linear quadratic tracking problem with input time delay, Int J Robust Nonlinear Control. 3 (3) 2021, 1324–1339.

[4] M. Kohlmann, S. Tang, Multidimensional backward stochastic Riccati equations and applications, SIAM J. Control Optim., 41 (2003) 1696–1721.

[5] H. Liu, H. Qiu, Optimal tracking control of linear discrete-time systems under cyber attacks, IFAC PapersOnLine, 53(2) (2020) 3545–3550.

[6] H. Modares, B. Kiumarsi, K. G. Vamvoudakis, F. L. Lewis, Adaptive H_∞ tracking control of nonlinear systems using using reinforcement learning. Adaptive Learning Methods for Nonlinear System Modeling, (2018) 313–333.

[7] H. Modares, F. L. Lewis, Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. IEEE Trans. Automat. Contr., 59(11) (2014) 3051–3056.

[8] Z. Qian, X. Y. Zhou, Existence of solutions to a class of indefinite stochastic Riccati equations. SIAM J. Control Optim., 51 (2013) 221–229.

[9] B. Wei, E. Tian, T. Zhang and X. Zhao, Probabilistic-Constrained H_∞ Tracking Control for a Class of Stochastic Nonlinear Systems Subject to DoS Attacks and Measurement Outliers, IEEE Transactions on Circuits and Systems I: Regular Papers, 68(10), 2021, 4381–4392.

[10] F. Zhang, Y. Dong, Q. Meng, Backward stochastic Riccati equation with jumps associated with stochastic linear quadratic optimal control with jumps and random coefficients. SIAM J. Control Optim., 58(1) (2020) 393–424.

[11] X. Zhao, C. Liu, J. Liu, E. Tian, Probabilistic-constrained reliable H_∞ tracking control for a class of stochastic nonlinear systems: An outlier-resistant event-triggered scheme. J Franklin Inst., 358(9) (2021) 4741–4760.

38

A new model for the spread of cyber-epidemics

E. Primo[♮,1], D. Aleja[♮,♢], G. Contreras-Aso[♮], K. Alfaro-Bittner[♮], M. Romance[♮,♢] and R. Criado[♮,♢]

(♮) Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, ESCET Universidad Rey Juan Carlos
c/Tulipán s/n, 28933 Móstoles, Madrid, Spain.

(♢) Data, Complex Networks and Cybersecurity Sciences Technological Center, Universidad Rey Juan Carlos, Pza. Manuel Becerra 14, 28028 Madrid, Spain.

1 Introduction

In an interconnected world like ours, a cyber-epidemic is a real risk. A cyber-virus that spreads from one device to another until it undermines the global Internet system with devastating consequences in terms of economic costs and social damage related to the shutdown of essential services is not so far of our reality. Since the first attempt to describe the spread of computer viruses [1], several other studies have attempted to adapt classical, global or networked compartmental models.

In particular, the SIR model [2, 3] is one of the simplest compartmental models, and describes the evolution of diseases resulting in the immunization or death of the infected individuals. The model assumes that, at each time, each individual can be in one of three possible compartments: susceptible (denoted by S), infected (I), or removed (R). The susceptible units of the network are those healthy individuals that can develop the disease if they are in contact with infected individuals. Once an individual contracts the infection, it moves into the infected (and infective) compartment, and then, after some time, into the removed compartment, which indicates that the individual cannot catch the disease anymore (or passes it on), due to a lasting resistance conferred by the recovery (or because it dies).

Starting from the work by Pastor–Satorras and Vespignani [4], the last years have seen a burst of activity on understanding the effects of the network topology on the rate and patterns of the disease spread. A lot of studies have tried indeed to predict the total number of infected individuals, the duration of an epidemic, or to estimate various relevant parameters such as the reproductive number, among others. Moreover, and especially in relation to the recent COVID19 world pandemic crisis, these models have been used to assess the effects of public health interventions and/or to quantify the efficiency of issuing a limited number of vaccines in a given population [5, 6].

In our work [7], we introduce a new model for studying the spreading of a malware and the awareness of its incidence through different waves which are evolving on the same graph structure (the global network of connected devices). We show the evolution of a cyber-epidemic from the moment a new virus is planted on the network, until some users(or one) realize that their devices have been damaged and, consequently, initiates a wave of awareness that finally ends with the development of a suitable antivirus software.

[1] eva.primo@urjc.es

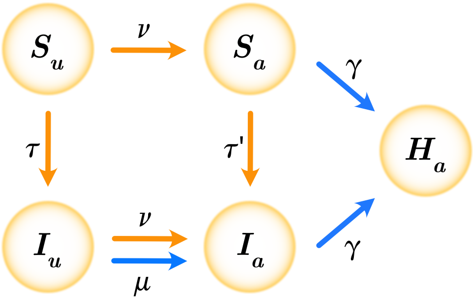


Figure 1: **The compartmental model for cyber-epidemics.** Units of the network can be in one of four possible states before recovery: susceptible-unaware (S_u), susceptible-aware (S_a), infected-unaware (I_u), and infected aware (I_a). The recovery or healed status (H_a) is by nodes that, after getting aware of the existence of the malware, applies a suitable operation on the system. Orange (blue) arrows stand for contact-based (localized spontaneous) transitions.
Source: Reprinted figure from Ref. [7].
©2022 with permission from Elsevier.

2 Results

In our model, we consider a wave of malware contagion as well as, but also a second wave of awareness and actions to counteract it. That is, we define a vector version of the SIR model, in which each susceptible or infected node is also aware or unaware of the existence of the virus. This means that there are four possible states before recovery: susceptible-unknown (S_u), susceptible-aware (S_a), infected-unknown (I_u), and infected-aware (I_a), see Fig. 1. Finally, the nodes (devices) whose users apply an appropriate operation on the system, for example, the installation of an antivirus, after learning of the existence of the virus go to the recovery or cured state (H_a).

The transition probabilities between such states are mediated by four fundamental parameters, whose meaning is directly linked to specific processes occurring during a cyber-epidemic:

- τ is the equivalent of the standard infection parameter in the SIR model. Here, it accounts for the contact-based transition rate from the susceptible to the infected status. In the case of S_a individuals, a different infection rate is used ($\tau' = \tau/10$),
- ν is the rate of spreading of awareness due to network contacts;
- μ accounts for the individual awareness parameter;
- γ is the recovery/healing parameter.

Symbol	Description	Range	Chosen value
τ	Infection rate	[0,1]	0.0055
ν	(Contact-based) awareness parameter	[0,1]	0.011
μ_0	(Spontaneous or local) awareness parameter	[0,1]	0.011
γ	Recovery parameter	[0,1]	0.03
ρ_0	Fraction of population initially infected	[0,1]	0.01

Table 1: Summary of the parameters involved in the model.

Two types of transitions may therefore occur in our model: contact-based transitions and individual (or contact-independent) ones. In individual transitions, the change of state is fully independent

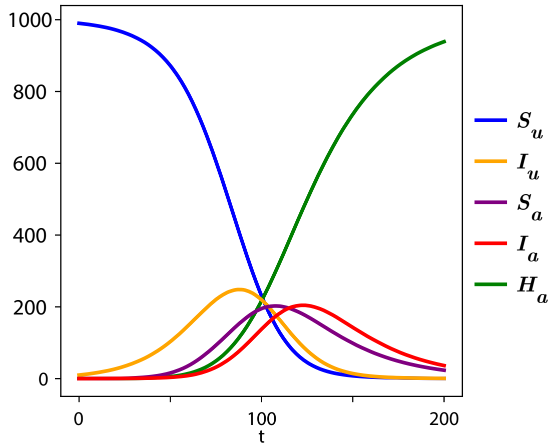


Figure 2: **The typical cycle of a cyber-epidemic.** The total number of S_u, S_a, I_u, I_a and H_a units (see legend for color code) vs. time, during a typical cyclic evolution of the virus. Simulations refer to an ensemble average over 1000 realization of an Erdős-Rényi random network with $N = 1000$ nodes, average degree $\langle k \rangle = 10$, threshold $\theta = 0.2$ and constant damage $d = 0.3$. It is possible to distinguish three different phases of the cycle: an initial phase of no awareness where more and more nodes pass from the S_u to the I_u state; a second stage, the awareness phase starting at around $t \sim 30$, when a given I_u node becomes I_a and ignites an awareness wave so that the total number of both I_a and S_a grows by contact transitions; a final healing stage where all nodes turn to the H_a state.
Source: Reprinted figure from Ref. [7].
©2022 with permission from Elsevier.

of the states of the rest of the devices, but is only due to the perception of the damage caused to the device (in the case of the passage from the state I_u to the state I_a at rate μ) and/or the user's interest in using an antivirus (in the case of the passage from any aware state to the state H_a at rate γ).

In our simulations, we initially prepare the network with all its units in the S_u state. Then, one node (or a small group of nodes) is turned to the state I_u, and contact propagations start, producing an initial spreading of the virus. At a second time, i.e., when one of the infected users gets aware of the damage produced by the cyber-virus in its device, a second wave (which spreads awareness in the same network) starts due to an initial local transition from the state I_u to the state I_a. Eventually, the entire cycle of the cyber-epidemic takes place, with an end state of the network where all its units are in the H_a state.

To control how damaging a cyber-epidemic can be in a network of devices, we introduce the d damage parameter in $[0, 1]$ caused when a device is infected by the virus, and we quantify the total damage caused to the system as the sum of the damage done to each of the devices. Finally, since the μ parameter is entirely related to the damage done to a device, it is convenient to define it as proportional to the d parameter. In addition, the μ parameter is activated depending on whether it exceeds a threshold $\theta \in [0, 1]$, to take into account the user's sensitivity to damage to the device. The latter leads to the following expression with $\mu_0 \in [0, 1]$:

$$\mu = \begin{cases} \mu_0(d - \theta) \ \text{ if } \ d \geq \theta, \\ \mu = 0 \ \text{ if } \ d < \theta. \end{cases}$$

Figure 2 reports the total number of S_u, S_a, I_u, I_a and H_a units vs. time during a typical cyclic evolution of the virus. From the figure, one can clearly distinguish the three different stages of the

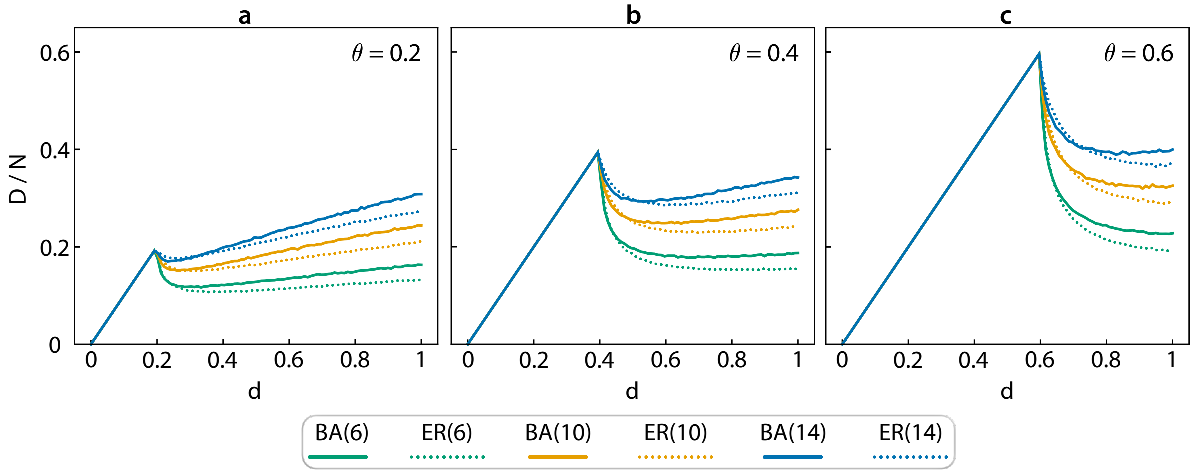


Figure 3: **Effects of network topology.** D/N vs. d, for different choices of networks (Barabási-Albert or Erdős-Rényi) of mean degree 6, 10 and 14 (see legend at the bottom of the figure for color code) and different values of the threshold θ: a) $\theta = 0.2$, b) $\theta = 0.4$, and c) $\theta = 0.6$.
Source: Reprinted figure from Ref. [7].
©2022 with permission from Elsevier.

cycle: an initial stage of no awareness where more and more nodes are infected (passing from the S_u to the I_u state), the beginning of the awareness phase (around $t \sim 30$) when a given I_u node becomes I_a and starts an awareness wave so that the total number of both I_a and S_a grows by contact transitions, and the final healing stage where all nodes turn to the H_a state.

To quantify the total injury produced in the system by the cyber-virus during its cycle of evolution, we introduce the quantity D/N accounting for the normalized sum of the individual damages suffered by each node. Its value is obtained with multiplying d by the total number of infected nodes during a cycle (regardless on whether they are in the state I_u or I_a), and dividing by N. When d is below the threshold θ the presence of the cyber threat is never detected and, as a consequence, the virus will propagate to all nodes in the network yielding $D/N = (N \cdot d)/N = d$. A non trivial behavior is instead observed for all values $d > \theta$, where the awareness mechanism is activated at a given time in the cycle.

As a first step, we face the problem of assessing how different topologies of the network react to a virus (causing a constant damage d when infecting a device) at different values of the threshold θ. With the aim of comparing homogeneous and heterogeneous topologies Fig. 3 illustrates the results of our simulations, and reports D/N vs. d for Barabási-Albert [8,9] scale free and Erdős-Rényi [10] random networks, at different values of the mean degree $\langle k \rangle$ and different values of θ. A first, even though rather trivial, evidence is that the global damage is higher for higher values of θ, indicating that the harder it is for nodes to become aware, the more free is the cyber threat to propagate without countermeasures. Finally, and more remarkably, the results shown in Fig. 3 allow to conclude that Barabási-Albert scale free (i.e., heterogenous) networks are more fragile, and more sensitive to the spreading of viruses causing a fixed damage than Erdős-Rényi (i.e., homogenous) ones, in line with what was already known about their structural fragility against intentional attacks [11].

We then move to consider the effects of a damage which is variable in time, that is, $d = d(t)$, where t is a discrete time measuring the number of iterations in the model (a time unit being the lapse from one to another iteration of the networks' nodes). Namely, starting at $d_0 = 0.1$, the

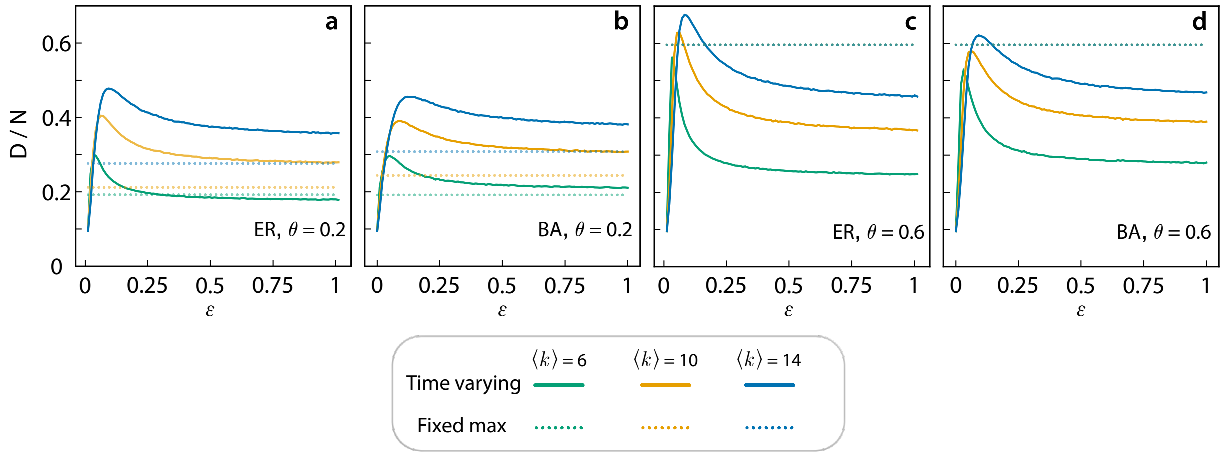


Figure 4: **Effects of time varying damage.** D/N vs. ε (see text for definition), for different choices of networks (Erdős-Rényi in the left column and Barabási-Albert in the right column) of mean degree 6, 10 and 14 (see legend at the bottom of the figure for color code) and different values of the threshold θ: a,b) $\theta = 0.2$, c,d) $\theta = 0.6$. In all panels the maximal injuries achieved by a fixed strength virus (the global maximum of the curves of Fig. 3) are reported with horizontal dashed lines.
Source: Reprinted figure from Ref. [7].
©2022 with permission from Elsevier.

damage is increased in times as

$$d(t) = \frac{d_0 e^{\varepsilon t}}{1 + d_0(e^{\varepsilon t} - 1)}, \tag{1}$$

for some $\varepsilon > 0$. Note that the function $d(t)$ is the logistic function with growth rate ε, initial population d_0 and carrying capacity 1, so that it displays an exponential behavior ($d(t) \approx d_0 e^{\varepsilon t}$) when $t \approx 0$ and approaches 1 when $t \to \infty$. Choosing the logistic function as the damage function is ideal in the model, as the intentions of the bad team is precisely that of spawning as much damage as possible before the virus is detected by users.

Fig. 4 reports D/N vs. ε, for both Erdős-Rényi (left column) and Barabási-Albert (right column) at different mean degrees and different values of θ. For comparison, in the same figure the maximum possible injury caused to the system by a fixed strength virus is reported by horizontal dashed lines. One can easily see that, as a function of ε, D/N displays an initial monotonic growth and an asymptotic behavior for $\varepsilon \to 1$, featuring a local maximum for some $0 < \bar{\varepsilon} < 1$. Remarkably, one can notice that for low values of the threshold, there is a range of ε for which the amount of injury inflicted to the system is actually higher than the maximum at constant base damage. On the contrary, at high values of θ (i.e., when security is not so demanding), viruses whose base damage is given by Eq. (1) cause an overall injury which is comparable with the maximal value at constant damage strength.

3 Conclusions

In conclusion, we have introduced a novel compartmental model that describes the spread of malware (and knowledge of its occurrence) in a given network of devices. The novelty of our approach is to consider vector compartments formed by two components the first describing the state of the device with respect to the spread of the virus, and the second the awareness of the device user about the presence of the cyber threat.

We then illustrate the global damage that a malware is capable of producing in Erdős-Rényi and scale-free architectures both for the case where the virus causes fixed damage on each device and for the case where, instead, the virus is that the virus is designed to mutate as a function of time.

References

[1] Winfried Gleissner. A mathematical theory for the spread of computer viruses. *Computers & Security*, 8(1):35–41, 1989.

[2] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control.* Oxford university press, 1992.

[3] James D Murray. *Mathematical biology. Second corrected edition.* Springer-Verlag, Berlin, 1993.

[4] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200, 2001.

[5] Stefano Boccaletti, William Ditto, Gabriel Mindlin, and Abdon Atangana. Modeling and forecasting of epidemic spreading: The case of covid-19 and beyond. *Chaos, Solitons, and Fractals*, 135:109794, 2020.

[6] Stefano Boccaletti, Gabriel Mindlin, William Ditto, and Abdon Atangana. Closing editorial: Forecasting of epidemic spreading: lessons learned from the current covid-19 pandemic. *Chaos, Solitons, and Fractals*, 139:110278, 2020.

[7] David Aleja, Gonzalo Contreras-Aso, Karin Alfaro-Bittner, Eva Primo, Regino Criado, Miguel Romance and Stefano Boccaletti. A compartmental model for cyber-epidemics, *Chaos, Solitons, and Fractals* 161:112310, 2022.

[8] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[9] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[10] Paul Erdös and Alfred Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[11] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

39

Bifurcation analysis in dryland vegetation models with discrete and distributed delays

I. Medjahdi[♭], F.Z. Lachachi[♭], M.A. Castro[♭] and F. Rodríguez[♭,♮,1]

(♭) Dept. Applied Mathematics, University of Alicante,
(♮) Multidisciplinary Institute for Environmental Studies (IMEM),
University of Alicante,
E-03080, Alicante, Spain.

1 Introduction

Mathematical modeling has become an essential tool in the study of dryland ecosystems, helping analyze threshold-type dynamics and identify critical factors that can affect degradation and restoration trajectories. Different types of dryland vegetation models have been proposed in the literature, with different methodological approaches, species and responses. In order to reflect the effect of past information of the system on the dynamic behavior of the models, it is often necessary to include time delays into the models. It is well known that the inclusion of delay terms in population models may fundamentally impact the dynamics of a system, which may result in stability changes and the presence of oscillations and periodic solutions (e.g., [1, 2]).

In this work, a mean field approximation to a cellular automata model that integrates basic ecological processes in drylands [3], including the effect of surrounding vegetation on favouring plant establishment and facilitating recovering from soil degradation, is considered. Both discrete and distributed time delay terms, to account for the effects of cumulated seed bank on plant establishment and of past continuous vegetation cover on soil recovery, are incorporated. Distributed delay terms with gamma type kernels are considered, and the so called *linear chain trick* is used in the analysis [4, 5]. The dynamics of the models is analysed in terms of equilibria, stability, and the presence of bifurcations, both depending on a parameter representing environmental conditions and on the effects of delays.

2 Methods and Results

The spatially explicit cellular automata model in [3] included three possible states, vegetated, empty and degraded, for each site in a rectangular mesh. The derived spatially implicit mean field approximation model [6] considers the frequencies for each type of state. Since the sum of frequencies total one, the model can be reduced to a system of two independent equations [3],

[1] f.rodriguez@ua.es

$$\begin{cases} x'(t) \equiv \dfrac{dx(t)}{dt} & = & -mx(t) + (1 - x(t) - y(t))x(t)(b - cx(t)) \\ \\ y'(t) \equiv \dfrac{dy(t)}{dt} & = & d(1 - x(t) - y(t)) - y(t)(r + fx(t)), \end{cases}$$

where x and y represent, respectively, the frequencies of vegetated and degraded states, so that $x, y \geq 0$ and $x + y \leq 1$, the frequency of the empty state being $1 - x - y$.

There are six parameters in the model, m (mortality), c (competition), b (plant establishment), d (degradation), r (regeneration), and f (facilitation). Vegetation may disappear at a rate m and recover, from empty sites and depending on seed production, at a maximum rate b, which is diminished by competition (c). Degraded sites are unable to sustain new vegetation, and they are originated, from empty sites, at a constant rate d, and recovered, into empty sites, at a constant rate r increased by facilitation (f) depending on surrounding vegetation.

There are two terms in this model where it is reasonable to expect and consider delay effects, revegetation of empty sites and vegetation-dependent regeneration of degraded soil. Production of seed is not immediate, and plant establishment may depend on seeds produced the previous season or cumulated during a certain period, leading to models with either discrete or distributed delays. The effect of vegetation on recovery is almost certainly cumulative, with maintenance of vegetation helping improve soil properties, and so a distributed delay would be a reasonable choice.

The inclusion of a discrete delay in the first equation of the model was considered in [7],

$$\begin{cases} x'(t) & = & -mx(t) + (1 - x(t) - y(t))x(t - \tau)(b - cx(t)) \\ y'(t) & = & d(1 - x(t) - y(t)) - y(t)(r + fx(t)), \end{cases}$$

showing that the presence of the delay could help recovery after disturbances. In this work distributed delays are introduced in the form of Gamma kernels,

$$g_a^p(t) := \frac{a^p t^{p-1} e^{-at}}{(p-1)!},$$

that is, the density function for a Gamma distribution with parameters $a > 0$ and $p = 1, 2...$, so that $\int_0^\infty g_a^p(t)dt = 1$.

We consider different models, including distributed delays in the revegetation term,

$$x'(t) = -mx(t) + (1 - x(t) - y(t))(b - cx(t)) \int_{-\infty}^{t} g_a^p(t - s)x(s)ds,$$

or in the vegetation-dependent facilitation,

$$y'(t) = d(1 - x(t) - y(t)) - y(t) \left(r + f \int_{-\infty}^{t} g_a^p(t - s)x(s)ds \right),$$

as well as combinations of either discrete or distributed delay in the first equation and distributed delay in the second one.

In particular, we use so called weak ($p = 1$) and strong ($p = 2$) Gamma kernels (Figure 1). In these type of kernels, the average delay is given by p/a, with variance p/a^2 (Figure 2), so that a discrete delay τ can be recovered as the limit case when $p \to \infty$ and $a = p/\tau$.

For these type of kernel, the so called *linear chain trick* [4,5] can be used to convert the system with delays into an extended ordinary differential equations system. Thus, for instance, the model

$$\begin{cases} x'(t) & = & -mx(t) + (1 - x(t) - y(t))x(t - \tau)(b - cx(t)) \\ y'(t) & = & d(1 - x(t) - y(t)) - y(t) \left(r + f \int_{-\infty}^{t} g_a^p(t - s)x(s)ds \right), \end{cases}$$

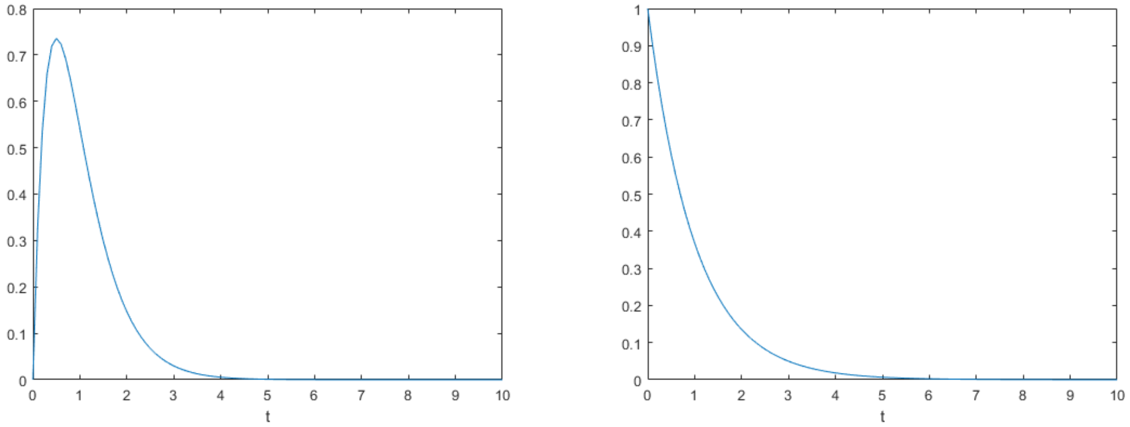


Figure 1: Weak ($p = 1$, left) and strong ($p = 2$, right) Gamma kernels.

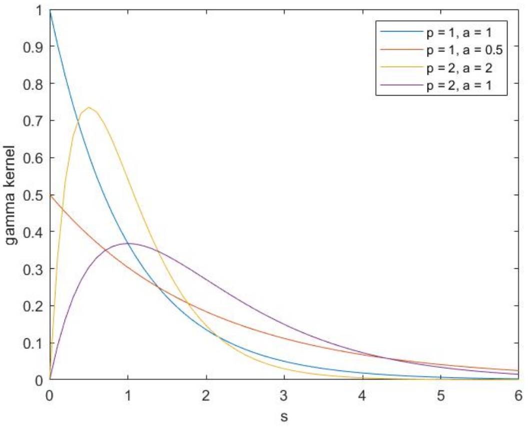


Figure 2: Weak and strong Gamma kernels with different average delays and variances.

with $p = 1$, introducing the new variable

$$z(t) = \int_{-\infty}^{t} ae^{(t-s)}x(s)ds,$$

can be transformed into the system

$$\left\{\begin{array}{rcl} x'(t) & = & -mx(t) + (1 - x(t) - y(t))z(t)(b - cx(t)), \\ y'(t) & = & d(1 - x(t) - y(t)) - y(t)(r + fz(t)), \\ z'(t) & = & ax(t) - az(t). \end{array}\right.$$

Similarly, for the same model with $p = 2$, defining two new variables,

$$\begin{array}{rcl} z(t) & = & \displaystyle\int_{-\infty}^{t} a^2(t-s)e^{(t-s)}x(s)ds, \\ w(t) & = & \displaystyle\int_{-\infty}^{t} ae^{(t-s)}x(s)ds, \end{array}$$

it can be transformed into the system

$$\begin{cases} x'(t) &=& -mx(t) + (1 - x(t) - y(t))x(t)(b - cx(t)), \\ y'(t) &=& d(1 - x(t) - y(t)) - y(t)(r + fz(t)), \\ z'(t) &=& aw(t) - az(t), \\ w'(t) &=& ax(t) - aw(t). \end{cases}$$

Equilibrium points, stability and bifurcations

For all the models considered here, with or without delays, there is one trivial equilibrium point with no vegetation ($x = 0, y = d/(d + r)$), called desert state, and possibly up to two equilibria with positive vegetation given by the solutions in $(0, 1]$ of the third degree equation

$$g(x) = cfx^3 + (-cf - bf + rc)x^2 + (-mf + bf - rc - br)x - md - mr + br.$$

Hence, for a certain range of parameter values, there can be two non-trivial equilibria (Figure 3, left), with two stable states (desert and vegetated) separated by an unstable equilibrium.

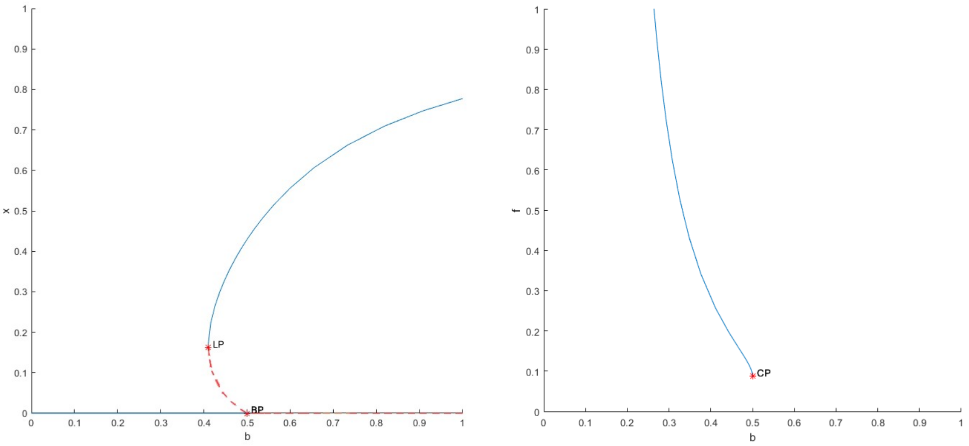


Figure 3: Left: Stable (blue) and unstable (dashed red) vegetation equilibria, and bifurcation points (LP and BP), as function of plant establishment (b). Right: Bifurcation analysis depending on parameters b and f.

.

Linearising at equilibria points, and with standard stability analyses, it can be shown that the presence of different types of delays do not alter the stability properties of equilibria in the different models. As a function of plant establishment, the parameter best representing environmental conditions, all the models exhibit a limit point (fold bifurcation) and a branch point (LP and BP in Figure 3 left), at the critical values of b defining the region of bistability. As function of the two parameters b and f, there is a cusp bifurcation (CP in Figure 3 right), showing that there needs to be a minimum level of facilitation for the model to exhibit bistability.

Recovery from perturbations

The presence of either discrete or distributed delays in the different models affect the capacity of the system to recover from random disturbances. As shown in Figure 4, after a decrease in vegetation from the equilibrium vegetated state, the presence of a delay allows the system to recover, but not below a certain critical value of the delay.

A more detailed analysis can be obtained by considering all possible perturbations from vegetated equilibrium, both in vegetation and degraded states frequencies. Different magnitudes of

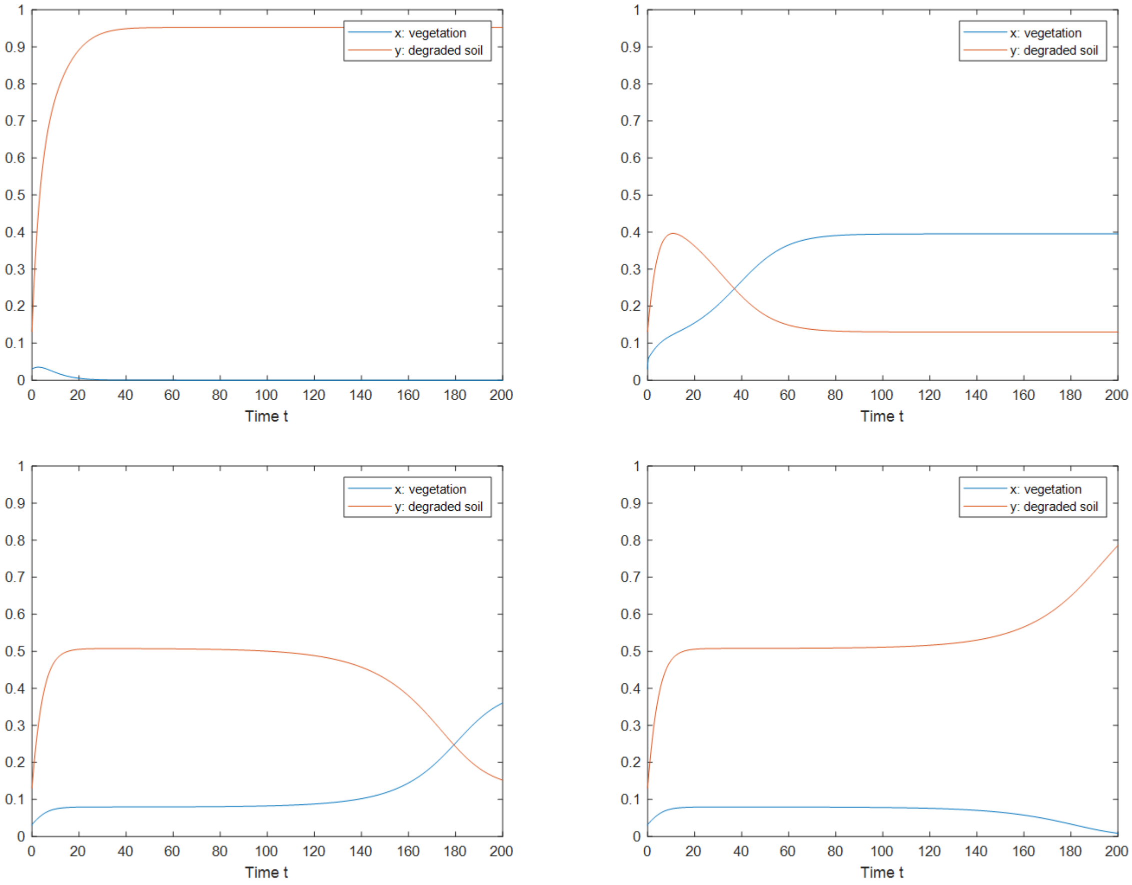


Figure 4: Dynamics of vegetation (blue) and degraded sites (ochre) after perturbation from vegetated equilibrium in the models without delay (top left) and with distributed delay in the revegetation term with $p = 2$ and different average delay values (top right: $a = 10$; bottom left: $a = 143$; bottom right: $a = 145$).

perturbations are simulated by displacing the system to new x and y values, and the final equilibrium state, with or without vegetation, is obtained. As shown in Figure 5, for a model with distributed delay in the vegetation-dependent facilitation term, the presence of a delay reduces the basin of attraction of the desert state, helping the recovery of vegetation.

3 Conclusions

The mean field dryland vegetation models considered in this work, both with and without discrete and/or distributed delays, may exhibit bistability for a range of environmental conditions, as represented by plant establishment (b), with the presence of a fold bifurcation for a critical value of b, also depending on a minimum level of facilitation.

The presence of a time delay affecting either revegetation or vegetation-dependent facilitation of recovery from the degraded state may increase the regeneration capacity of the system after random disturbances from equilibrium, reducing the basin of attraction of the desert state, with the need of a minimum critical level of delay for this effect to happen.

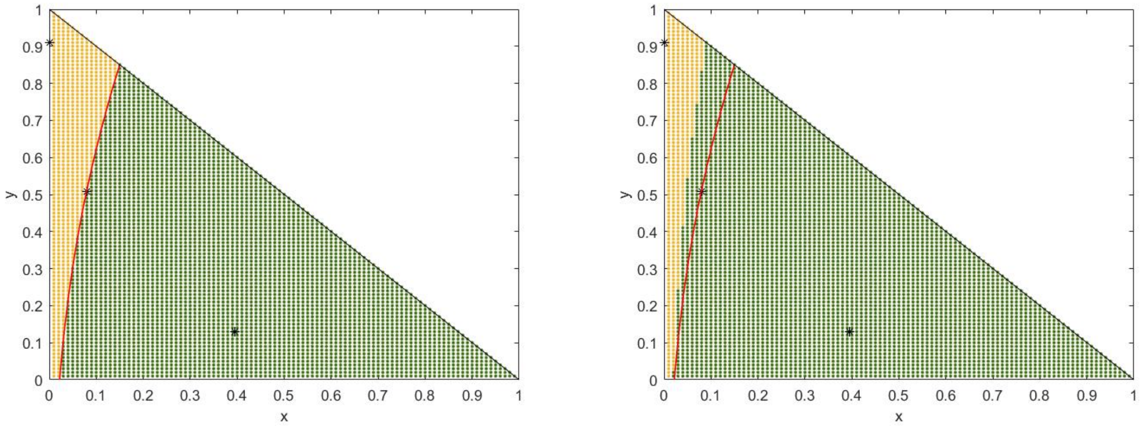


Figure 5: Final state of the system, vegetated (green) and desert (ochre) in the models without delay (left) and with a distributed delay in the vegetation-dependent facilitation term with $p = 1$ and $a = 0.5$ (right), after different levels of perturbations from vegetated equilibrium

References

[1] Ma, Z.P., Huo, H.F., Liu, C.Y.: Stability and Hopf bifurcation on a predator-prey model with discrete and distributed delays. *Nonlinear Analysis: Real World Applications*, 10: 1160–1172, 2009.

[2] Song, Y.L., Yuan, S.L.: Bifurcation analysis in a predatorprey system with time delay. *Nonlinear Analysis: Real World Applications*, 7: 265–284, 2006.

[3] Kéfi, S., Rietkerk, M., van Baalen, M., and Loreau, M., Local facilitation, bistability and transitions in arid ecosystems, *Theoretical Population Biology*, 71: 367–379, 2007.

[4] Cushing, J.M. *Integrodifferential Equations and Delay Models in Population Dynamics.* Springer, Heidelberg 1977.

[5] Smith, H. *An Introduction to Delay Differential Equations with Applications to the Life Sciences.* Springer, New York, 2011.

[6] Morozov, A., Poggiale, J.-C. From spatially explicit ecological models to mean-field dynamics: The state of the art and perspectives, *Ecological Complexity* 10: 1–11, 2012.

[7] Lachachi, F.Z., Medjahdi, I., Castro, M.A., Rodríguez, F. Stability, bifurcations, and recovery from perturbations in a mean-field semiarid vegetation model with delay. In: J.R. Torregrosa, J-C. Cortés, J. A. Hervás, A. Vidal-Ferràndiz and E. López-Navarro, eds. *Modelling for Engineering & Human Behaviour 2021.* I.U. de Matemàtica Multidisciplinar, Universitat Politècnica de València, 2021.

40

Relative research contribution towards railways superstructure quality determination from the vehicles inertial response

J. R. Sánchez[♭,1], F. J. Vea[♮], G. Muinelo[♭] and G. Mateo[◇]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(♮) Universidad Católica San Antonio de Murcia,
Av. de los Jerónimos 135, Guadalupe de Maciascoque, Murcia, Spain.
(◇)Idvia 2020 Horizonte 2020 SL.
Av. d'Aragó 30, València, Spain.

1 Introduction

Currently, prior to pavement maintenance, periodic monitoring campaigns are considered, together with visual inspections, to detect pavement deterioration (corrective/preventive maintenance). In these campaigns, the International Regularity Index (IRI) is measured, which indicates whether there are irregularities in the pavement that may negatively influence the driver's comfort and convenience and that may affect traffic safety. However, current high-precision systems are slow and/or costly procedures.

2 Methodology

The project has developed a solution that achieves: on the one hand to obtain an auscultation vehicle from a conventional vehicle capable of recording, at conventional operating speed (80km/h), the International Roughness Profile (IRI), all with high accuracy and on the other hand, as an added value, the system is able to provide a survey of the defects detected, a prediction of the evolution of the roughness and a maintenance plan, all automatically, continuously and in real time. The proposed system consists of 4 subsystems:

2.1 Hardware subsystem

- 2 triaxial accelerometers, which are installed one on the fixed mass of the vehicle and the other on the suspended mass, The accelerometers used are MEMS type with a measurement range of ±6 g and a bandwidth of 1000 Hz.

- A 6-channel analog data acquisition system (3 for each accelerometer) with high resolution (16 bits) and a sampling rate of 2000 Hz.

[1] juasanv6@upv.es

- A Global Positioning System (GPS).
- An On Board Diagnostics II (OBD-II) reader or interface.

2.2 Communication subsystem

To facilitate the installation of the device, the hardware has been divided into two units, the NODEBOX (in charge of taking the accelerometer measurements) and the COMMUNICATION BOX (in charge of taking the OBD and GPS measurements, storing the NODEBOX data and sending them to the server). The communication between the two units is done through a TCP socket via a Wi-Fi connection. Finally, through another TCP socket, the data is sent to the web server by means of 4G module.

2.3 Software subsystem

The SOFTWARE system: consists of four algorithms:

- Obtaining the IRI: algorithm based on the quarter car model obtains the roughness profile of the road.
- Defect detection: the algorithm uses a Wavelet Space Scale filter to extract the acceleration peaks due to the presence of defects, then by means of a shape analysis it classifies the defect into bumps or potholes and quantifies its intensity and size.
- IRI prediction: from periodic readings of the roughness of a road and its characteristics by means of a neural network, the algorithm is able to predict the evolution of the roughness of the road.
- Predictive maintenance: from the previous results, the algorithm is able to detect the pathology suffered in a section of the road and propose a series of actions for its correction.

Obtaining the IRI

The roughness profile is obtained from the measured accelerations and the vehicle position according to the following expression (developed from the quarter car model): [1]

$$W = Z_u + \frac{M_s\ddot{Z}_s + M_u\ddot{Z}_u}{K_t} \quad , where \quad Z_u = \int\int_0^T \ddot{Z}_u\, dt \tag{1}$$

The Golden Car simulation is carried out on the profile obtained [2]:

$$\left.\begin{array}{l} M_s\ddot{Z}_s + C_s(\dot{Z}_s - \dot{Z}_u) + K_s(Z_s - Z_u) = 0 \\ M_u\ddot{Z}_u - C_s(\dot{Z}_s + \dot{Z}_u) - K_s(Z_s - Z_u) + K_t(W - Z_u) = 0 \end{array}\right\} \rightarrow \dot{x} = Ax + B\, h_{ps} \tag{2}$$

Finally, with the accelerations obtained from the simulation, the IRI is obtained according to its definition:

$$IRI = \frac{1}{L}\int_0^{\frac{L}{v}} |\dot{Z}_s - \dot{Z}_u|\, dt \tag{3}$$

Defects detection

Defect detection is based on Wavelet transforms:

$$f(x) = \sum_{j=1}^{J} \sum_{k=1}^{K} \psi_{j,k}(x) W(j,k) \tag{4}$$

Specifically, the Scale Space Filter (SSF) is used, the mother function is the db2 of the Daubechy family of functions of order 2 and the spatial correlation of order $\Omega_2(j,k)$ is used as the identification signal. [3].

$$\Omega_2(j,k) = W(j,k)W(j+1,k) \tag{5}$$

The developed algorithm [4] follows the following steps to obtain the new filtered Wavelet coefficients $W_{new}(j,k)$ These coefficients are used to identify abrupt changes in acceleration due to the presence of potholes or bumps on the road.

Predictive maintenance

As for the future IRI prediction algorithm, a NARX [5] type neural network is used for this purpose, and the inputs are a characterization of the road and the previous measured IRI values.

In order to simplify maintenance, an algorithm for the recognition of the different defect patterns has been developed and a database of the most appropriate correction method for each of the typologies has been established [6].

3 Results

3.1 Results of IRI measures

To validate the system as a profilometer, a series of tests were carried out on the CV-101 and CV-240. Both sections are single carriageway roads with 2 lanes, one in each direction of approximately 2 km in length. A total of 4 passes were made on each lane in order to compare the different measurements. The IRI deviation obtained a maximum value of 0.3352 m/km for CV101 C2 and a mean value of 0.2837 m/km. Comparing these measures with the values obtained by a commercial system (Greenwood laser profilometer) obtained errors are less than 0.5 m/km, the average being around 0.39 m/km.

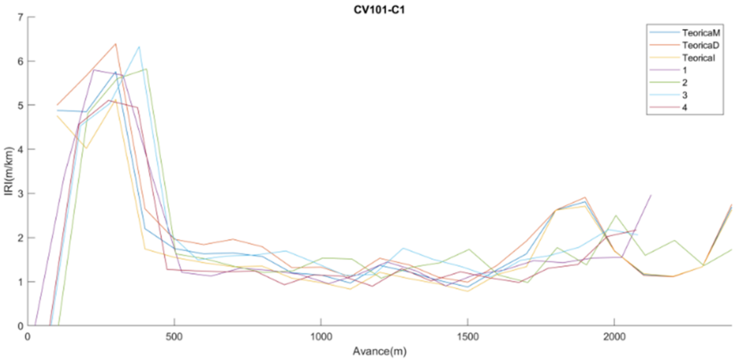


Figure 1: IRI measures over CV101

Mean errors (m/km)	CV101 C1	CV101 C2	CV240 C1	CV240 C2
P1	0.4520	0.4805	0.3118	0.2712
P2	0.4023	0.3250	0.3509	0.4183
P3	0.4813	0.6130	0.2770	0.3248
P4	0.4105	0.4823	0.3421	0.3410
Mean	0.4353	0.4635	0.3191	0.3348

Table 1: Mean error of different measures

CEDEX certification

The equipment was then subjected to the annual intercomparison test of regularity index measurement equipment organized by CEDEX. The test was carried out over A1 from pk 70 to 79.

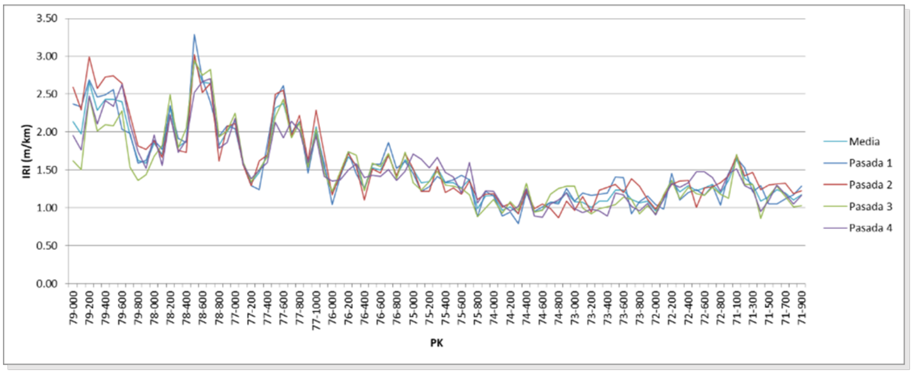


Figure 2: IRI measures over A1 pk 79 to 70

The results obtained for the section in question can be seen in the following graphs in both directions, for the measurements a deviation of 0.135 and 0.178 m/km was obtained.

According CEDEX results, the system got an average deviation of 0.07 mm/km between their runs, which is good parameter. But the problem appears when comparing it with the reference values, where an average error of 11.9% is obtained, this is increased by the number of passes made with a $\sqrt{3}$, where the value of 20.6% is obtained while the rest of the teams were around 10%. For this reason, the results were negative.

Improvements

In order to improve the results of IRI, some modifications were made to the hardware and software.

- Improve crystal precision with changes of temperature, change internal crystal of 4% stability by external with 25ppm (± 0,0025 %).
- News accelerometers developed by Microsensor are 5 times more accurate (Nonlinearity = 0.1% (0,5%) and Transverse sensitivity = 1% (5%).
- Install accelerometers on the rear axle (reduce influence of motor vibration)
- Calibrate with real IRI measures (nowadays is calibrated to improve repeatability)
- Calibrate or compare with obtained profile (calibration is done by only IRI measures)
- Filter acceleration with other family of filters (currents filters are based in FFT it presents distortions around localized disturbances “potholes and bumps”)

- Compare Golden-Car algorithm with a commercial software (Profile to IRI)
- Change quarter-car model by half-car model or complete model (reduce influence of vibration due to other wheels)

Results and improvements

Once improvements have been made, tests have been carried out with the Applus auscultation vehicle to check the improvement in the system's operation and to carry out the proposed software improvements. The test has been carried out according to the instructions followed during the CEDEX tests and the results have been analysed according to the same procedure. To obtain the reference measurements, a conventional Applus laser profilometer (certified by the CEDEX test) was used.

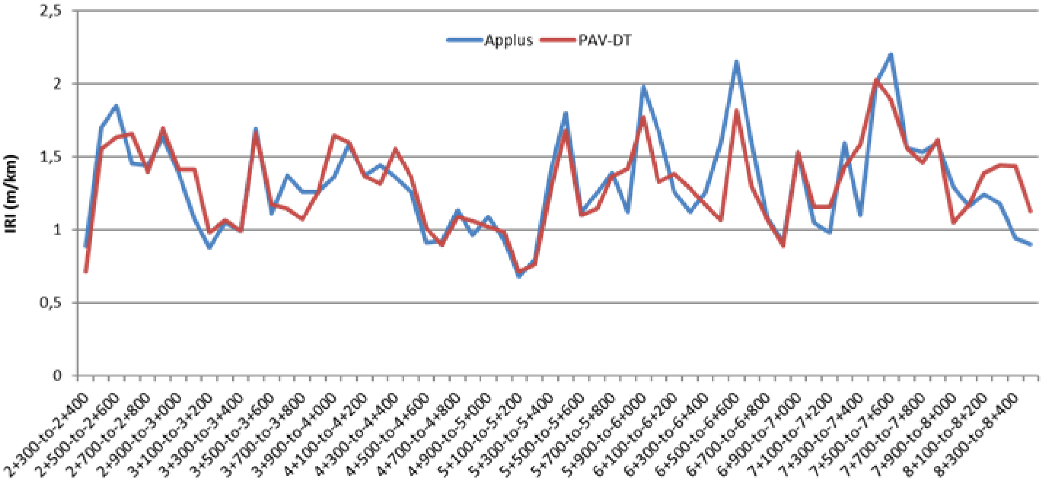


Figure 3: IRI measures over M30 pk 2 to 8

According to this analysis, an average error of 7.11% is obtained. The error obtained is much better than the one obtained in the CEDEX tests. It is estimated that the improvement in the results is due to the improvement in the acquisition times. The use of the higher precision external crystal allows less error in the positioning of the data.

Work continues to improve the software, especially with the use of alternative filters to the FFT. An improvement in the influence of anomalies has been observed with Hanning filtering.

With the current results and improvements being developed, the device is expected to be ready to pass the next CEDEX test.

3.2 Results of defect detection algorithm

A series of tests were then carried out to validate the defect detection and classification algorithm. Some tests were carried out to see the general functioning of the algorithm, the defects that were passed were marked manually. Here we can see in red triangle the defects identified manually, and the defects detected by the algorithm bumps (circles) and the potholes (squares).

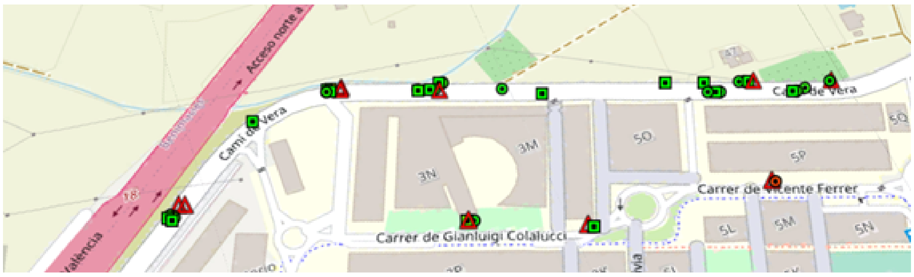


Figure 4: Defect detection results

4 Conclusion

- Nowadays, the system allows obtaining roughness profiles (IRI) with a mean standard deviation lower than 0.2 m/km.
- Several improvements have been implemented:
  - Improved location accuracy, (Start and end route identification)
  - Improved correlation between acceleration, GPS and OBD data.
  - Improved time accuracy
- Several improvements are currently being made to improve the accuracy of the equipment:
  - Filtering engine vibration
  - Improved compensation for changes in speed.
- Validation plan of the defect detection algorithm in a real environment have been developed
- With all these improvements, it is expected to achieve the objective precision

References

[1] Eshkabilov, S., Measuring and Assessing Road Profile by Employing Accelerometers and IRI Assessment Tools. *Journal of Traffic and Transportation Engineering.* 3, 2018

[2] Múčka, P. International Roughness Index Specifications around the World. *Road Materials and Pavement Design.* 2016

[3] Chui C.K. Wavelet Analysis and its applications: An introduction to Wavelets, 1992

[4] Ugweje O.C., Selective noise filtration of image signals using wavelet transform. *Measurement*, 36(3):279–287, 2004

[5] La Torre, F., Domenichini, L., & Darter, M. I. (1998). Roughness prediction model based on the artificial neural network approach. *4th Int. Conf. on Managing Pavements*, 2, 1998

[6] Wilde, W. J., Thompson, L., & Wood, T. J., Cost-Effective Pavement Preservation Solutions for the Real World. Department of Transportation, Research Services & Library, 2014.

41

Probabilistic analysis of the pseudo-n order adsorption kinetic model

C. Andreu-Vilarroig[♭], J.-C. Cortés [♭,1] A. Navarro-Quiles[♮] and S.-M. Sferle[♭]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València,
Camino de Vera s/n, 46022, Valencia, Spain.
(♮) Department of Statistics and Operational Research,
Universitat de València,
Dr. Moliner 50, 46100, Burjassot, Spain.

1 Introduction

One of the major problems of recent decades is environmental pollution. Among the various causes of contamination, human activity is one of the main ones. Intense industrial activity has brought with it multiple ecological damages, including the generation of polluting wastes and residues that have a negative impact on the environment. At present, one of the serious consequences of environmental problems is water pollution, especially the discharge of toxic industrial effluents such as lead, copper or organic dye waste into wastewater [2–5,5]. The effects of pollutants cause detrimental impacts on both the environment and human health [2,3]. These can be avoided by treating contaminated wastewater. In recent years, numerous techniques such as reverse osmosis [2,4], ultrafiltration [6], photocatalysis [6], electrochemical degradation [6], coagulation [2,4] and adsorption [2,3,6], among others, have been studied and applied to remove pollutants from wastewater. Adsorption has been recognized as superior to other techniques due to its high efficiency, simple design, easy operation and maintenance, and availability of different adsorbents at relatively low cost [2,4–6,6,8,9].

Adsorption is a surface phenomenon in which atoms, ions or molecules are trapped or deposited on the surface of a material. The efficiency of this process is studied by means of adsorption kinetics, since it provides important information on the rate and mechanism of adsorption, as well as on the time required for the adsorption process to complete [8–10]. In the literature we can find several models that have been developed to describe the adsorption kinetic process such as the pseudo-first-order model [11], the pseudo-second-order model [12], the Ritchie's equation [13] or the Elovich model [14], the first two being the most widely used. In the present work, we will introduce the pseudo-n-order (PNO) equation, whose form is as follows

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t} = k_n(q_{\mathrm{e}} - q(t))^n, \qquad n > 1, \tag{1}$$

where $q(t)$ represents the adsorption capacity at time t (mg/g), n the reaction order, q_{e} the adsorption capacity at equilibrium (mg/g) and k_n the rate ratio ($(\mathrm{mg/g})^{1-n}$ min^{-1}), which, in turn, has the following expression

[1] jccortes@mat.upv.es

$$k_n = \frac{r}{q_{\mathrm{e}}^{n-1}}, \tag{2}$$

where r is the adsorption rate (min^{-1}). For further details, see [10].

By integrating equation (1) yields the exact solution of the PNO model

$$q(t) = q_{\mathrm{e}} \left(1 - \left(\frac{1}{1 + t(n-1)k_n q_{\mathrm{e}}^{n-1}} \right)^{\frac{1}{n-1}} \right), \tag{3}$$

and substituting k_n for expression (2) the following expression is ensured

$$q(t) = q_{\mathrm{e}} \left(1 - \left(\frac{1}{1 + t(n-1)r} \right)^{\frac{1}{n-1}} \right), \tag{4}$$

where the parameters q_{e} and r are strictly positive and $n > 1$.

In many studies these parameters are considered deterministic. However, they are calculated from experimental data, so they have measurement errors, and depend on molecular properties that are not completely know and also on random external factors such as temperature or humidity, so they contain an intrinsic uncertainty [10]. Therefore, it is more convenient to consider them as random variables. Consequently, this leads to the equation (1) becoming a random differential equation, the solution of which will now be a stochastic process of the following form

$$q(t,\omega) = q_{\mathrm{e}}(\omega) \left(1 - \left(\frac{1}{1 + t(n(\omega)-1)r(\omega)} \right)^{\frac{1}{n(\omega)-1}} \right), \qquad \omega \in \Omega, \tag{5}$$

where $q_{\mathrm{e}}(\omega), r(\omega)$ and $n(\omega)$ are assumed to be absolutely continuous random variables defined on a common complete probability space $(\Omega, \mathcal{F}_{\Omega}, \mathbb{P})$, with a known joint probability density function (PDF) $f_0(q_{\mathrm{e}}, r, n)$. For simplicity, the ω-notation related to the sample dependence for random variables will be omitted.

This document is organized as follows. In section 2, we will determine the first probability density function (1-PDF), $f_1(q,t)$, of the solution (5). In section 3, the PNO random model is used to model the adsorption of cadmium on the tree fern, in order to apply the theoretical results to real case. Conclusions are drawn in the last section.

2 Computing the 1-PDF of the solution

This section is devoted to determine the 1-PDF of the solution of the PNO random model described in (5). For this purpose, we will apply the Random Variable Transformation (RVT) technique [10]. In this way, we will obtain a complete probabilistic information of the proposed model.

We set $t > 0$ and we apply the RVT technique using the following identification

$$\mathbf{v} = (q_{\mathrm{e}}, r, n), \quad f_{\mathbf{v}}(\mathbf{v}) = f_0(q_{\mathrm{e}}, r, n),$$
$$\mathbf{w} = (w_1, w_2, w_3) = \mathbf{r}(q_{\mathrm{e}}, r, n),$$

being the mapping $\mathbf{r} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ and its inverse $\mathbf{s} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ given, respectively, by

$$w_1 = r_1(q_{\mathrm{e}}, r, n) = q_{\mathrm{e}} \quad \Rightarrow \quad q_{\mathrm{e}} = s_1(w_1, w_2, w_3) = w_1,$$

$$w_2 = r_2(q_{\mathrm{e}}, r, n) = q_{\mathrm{e}} \left(1 - \left(\frac{1}{1 + t(n-1)r} \right)^{\frac{1}{n-1}} \right) \quad \Rightarrow \quad r = s_2(w_1, w_2, w_3) = \frac{\left(1 - \frac{w_2}{w_1}\right)^{1-w_3} - 1}{t(w_3 - 1)},$$

$$w_3 = r_3(q_\text{e}, r, n) = n \quad \Rightarrow \quad n = s_3(w_1, w_2, w_3) = w_3.$$

As $q_\text{e} = w_1 > 0$, $n = w_3 > 1$ and $0 \leq \frac{q(t)}{q_\text{e}} = \frac{w_2}{w_1} < 1$, then $\left(1 - \frac{w_2}{w_1}\right)^{1-w_3} > 1$ and $r = \frac{\left(1-\frac{w_2}{w_1}\right)^{1-w_3} - 1}{t(w_3-1)} > 0$, so it follows that the inverse mapping is well-defined in the conditional probability space $(\Omega, \mathcal{F}_\Omega, \mathbb{P}[\cdot|C])$, where $C = \{\omega \in \Omega : q_\text{e}(\omega) > q(t,\omega)\}$. But, due to the way the problem is defined, this condition is guaranteed so the probability of event C is 1 on the above conditional probability space.
The corresponding Jacobian is

$$J_3 = \frac{1}{tw_1(1 - \frac{w_2}{w_1})^{w_3}} > 0.$$

Thus, the joint PDF of the random vector $\mathbf{w} = (w_1, w_2, w_3)$ is

$$f_{w_1,w_2,w_3}(w_1, w_2, w_3) = f_0 \left(w_1, \frac{\left(1 - \frac{w_2}{w_1}\right)^{1-w_3} - 1}{t(w_3 - 1)}, w_3 \right) \frac{1}{tw_1(1 - \frac{w_2}{w_1})^{w_3}}.$$

Finally, considering an arbitrary t and taking into account that $q(t,\omega) = w_2$, we obtain the 1-PDF of the solution (5) as follows

$$f_1(q,t) = \int_0^\infty \int_1^\infty f_0 \left(q_\text{e}, \frac{\left(1 - \frac{q}{q_\text{e}}\right)^{1-n} - 1}{t(n-1)}, n \right) \frac{1}{tq_\text{e}(1 - \frac{q}{q_\text{e}})^n} \, \mathrm{d}n \, \mathrm{d}q_\text{e}. \tag{6}$$

From it, we can compute all the one-dimensional statistical moments of $q(t,\omega)$, such as the mean, $\mathbb{E}[q(t,\omega)] = \int_\mathbb{R} q f_1(q,t)\mathrm{d}q$, or variance, $\mathbb{V}[q(t,\omega)] = \int_\mathbb{R} q^2 f_1(q,t)\mathrm{d}q - (\mathbb{E}[q(t,\omega)])^2$, and other interesting features such as the probability that the solution lies within a particular interval of interest, $\mathbb{P}[\{\omega \in \Omega : a \leq q(t,\omega) \leq b\}] = \int_a^b f_1(q,t)\mathrm{d}q$, or the probability that the adsorbed amount exceeds a given quantity $\hat{q}$, $\mathbb{P}[\{\omega \in \Omega : q(t,\omega) > \hat{q}\}] = \int_{\hat{q}}^\infty f_1(q,t)\mathrm{d}q$.

3 An application

The aim of this section is to apply our proposed model together with the above theoretical results to a real case, particularly in a chemical environment, to study cadmium adsorption on tree ferns. In [16] they conclude that tree fern may be a suitable sorbate for sorption of metal cations due its polar and acid characters. The data are shown in Table 1.

t_i	0	4	5	10	15	20	30	45	60
q_i	0	7.172414	8.022989	9.724138	9.793103	10.551724	10.574713	11.103448	11.195402

Table 1: Adsorption capacity of cadmium ions on tree ferns, q_i, for different time instants t_i, $i \in \{1, 2, \ldots, 9\}$. Source [16].

To meet the objective of this section, it is first necessary to establish a joint PDF of the model parameters. Due to the physical significance of the variables, we can consider them independent, so that $f_0(q_\text{e}, r, n)$ in this case is the product of the marginals, $f_0(q_\text{e}, r, n) = f_{q_\text{e}}(q_\text{e})f_r(r)f_n(n)$. Therefore, a PDF must be assigned for each of the model parameters. This is a crucial step in practice, since it is necessary to find appropriate parameter distributions that best capture the data uncertainty.

To do so, we are going to use the Random Least Mean Square (RLMS) method [17]. Considering the positivity, boundedness and the available information of the random variables in the framework of this chemical application we assume

$$q_{\mathrm{e}} \sim \mathrm{Unif}(a,b), \quad a,b>0,$$
$$r \sim \mathrm{Ga}(\alpha,\beta), \quad \alpha,\beta>0,$$
$$n \sim \mathrm{N}_T(\mu,\sigma^2), \quad T=[1,3].$$

Nevertheless, the generality of the results presented in the previous section allows the choice of probability distributions different from those already chosen, which is an advantage on a practical level.

To know the 1-PDF of the solution, it is required to determine the parameters of the proposed parametric distributions, since it depends on these. To this end, we will minimize the mean squared error between the observed data, q_i, and the expectation of the solution $q(t_i,\omega;a,b,\alpha,\beta,\mu,\sigma)$ evaluated at the time instants t_i. This leads to the next optimization program

$$\min_{a,b,\alpha,\beta,\mu,\sigma,>0} \quad \sum_{i=1}^{9}\left(q_i - \mathbb{E}\left[q(t_i,\omega;a,b,\alpha,\beta,\mu,\sigma)\right]\right)^2, \tag{7}$$

where the expectation of the solution is calculated as indicated in section 2, with $f_1(q,t)$ being the expression given in (6).

The output of the minimization problem (7) are the following local optimal values obtained using the Nelder-Mead's algorithm implemented in Mathematica© software

$$a=10.9934,\ b=12.1161,\ \alpha=42.0172,\ \beta=0.009594,\quad \mu=1.8858,\ \sigma=0.3710.$$

Once the distributions of the model parameters have been estimated using the RLMS method, we obtain the 1-PDF of the solution of the PNO random model, the graph of which is shown in Figure 1.

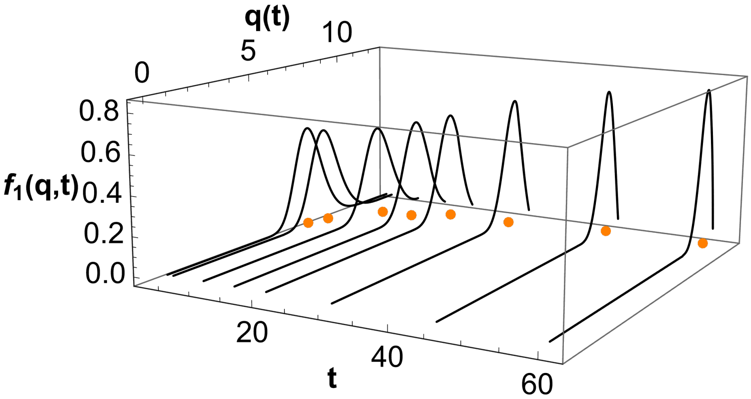


Figure 1: Visual display of the 1-PDF of the solution of the PNO random model, $f_1(q,t)$, given in equation (6), at time instants t_i for $i=2,...,9$, collected in Table 1.

It is observed that for each time instant, the PDF is concentrated approximately around the adsorbed quantity registered, following a morphology quite similar to that of a Gaussian. It is also observed that the density mass shifts with time towards quantities closer and closer to equilibrium levels, with some leptokurtic tendency and some negative asymmetry.

As mentioned above, from the 1-PDF we can get a broader analysis of the solution, thus obtaining its expectation together with the 95% confidence interval, whose representation is shown in Figure 2.

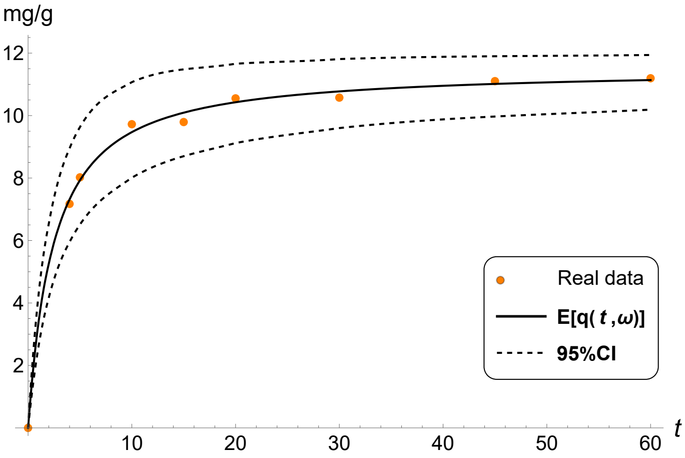


Figure 2: Visual display of the expectation (solid line) together with the 95% confidence interval (dashed lines) of the solution of the PNO random model, which represents the probabilistic fit of the data (points) indicated in Table 1.

Note that the expected value fits the data series well at all time points and that its evolution is increasing over time and stabilizes around equilibrium. As for the 95% confidence interval, we note that it captures the uncertainty coming from our data.

4 Conclusions

In this work we have proposed a randomization of a kinetic model formulated by differential equations to describe the chemical adsorption process. For this purpose, we have treated all the model parameters as random variables with arbitrary distributions. Then, we have solved it probabilistically by determining its respective first probability density function, thus obtaining a full probabilistic description of the solution at each time instant. Given real data, we have shown an uncertainty quantification technique for assigning appropriate distributions to all model parameters when applying the PNO model to real-world data.

Acknowledgement

This work has been partially supported by the Spanish Ministerio de Economía, Industria y Competitividad (MINECO), the Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional (FEDER UE) grant PID2020–115270GB–I00, the Generalitat Valenciana (grant AICO/2021/302), el Fondo Social Europeo y la Iniciativa de Empleo Juvenil EDGJID/2021/185 and la ayuda PRE2021-101090, financiado por MCIN/AEI/10.13039/501100011033 y por el FSE+.

References

[1] Salman Naeem, M., Javed, S., Baheti, V., Wiener, J., Javed, M. U., Ul Hassan, S. Z., Mazari, A., Naeem, J. Adsorption kinetics of acid red on activated carbon web prepared from acrylic fibrous waste. *Fibers and Polymers*, 19(1):71–81, 2018.

[2] Fideles, R. A., Ferreira, G. M. D., Teodoro, F. S., Adarme, O. F. H., da Silva, L. H. M., Gil, L. F., Gurgel, L. V. A. Trimellitated sugarcane bagasse: A versatile adsorbent for removal of cationic dyes from aqueous solution. Part I: Batch adsorption in a monocomponent system. *Journal of colloid and interface science*, 515:172–188, 2018.

[3] Sun, P., Xu, L., Li, J., Zhai, P., Zhang, H., Zhang, Z., Zhu, W. Hydrothermal synthesis of mesoporous $Mg_3Si_2O_5(OH)_4$ microspheres as high-performance adsorbents for dye removal. *Chemical Engineering Journal*, 334:377–388, 2018.

[4] Ahmad, R., Mirza, A. Synthesis of Guar gum/bentonite a novel bionanocomposite: Isotherms, kinetics and thermodynamic studies for the removal of Pb(II) and crystal violet dye. *Journal of molecular liquids*, 249:805–814, 2018.

[5] Romero-Cano, L. A., García-Rosero, H., Gonzalez-Gutierrez, L. V., Baldenegro-Pérez, L. A., Carrasco-Marín, F. Functionalized adsorbents prepared from fruit peels: Equilibrium, kinetic and thermodynamic studies for copper adsorption in aqueous solution. *Journal of cleaner production*, 162:195–204, 2017.

[6] Berenjian, A., Maleknia, L., Fard, G. C., Almasian, A. Mesoporous carboxylated Mn_2O_3 nanofibers: Synthesis, characterization and dye removal property. *Journal of the Taiwan Institute of Chemical Engineers*, 86:57–72, 2018.

[7] Aichour, A., Zaghouane-Boudiaf, H., Iborra, C. V., Polo, M. S. Bioadsorbent beads prepared from activated biomass/alginate for enhanced removal of cationic dye from water medium: Kinetics, equilibrium and thermodynamic studies. *Journal of Molecular Liquids*, 256:533–540, 2018.

[8] Musah, M., Azeh, Y., Mathew, J. T., Umar, M. T., Abdulhamid, Z., Muhammad, A. I. Adsorption Kinetics and Isotherm Models: A Review. *CaJoST*, 4(1):20–26, 2022.

[9] Wang, J., Guo, X. Adsorption kinetic models: Physical meanings, applications, and solving methods. *Journal of Hazardous materials*, 390:122156, 2020.

[10] Andreu-Vilarroig, C., Cortés, J. C., Navarro-Quiles, A., Sferle, S. M. (to appear). Statistical Analysis of a General Adsorption Kinetic Model with Randomness in Its Formulation. An application to Real Data. *MATCH Communications in Mathematical and in Computer Chemistry.*

[11] Lagergren, S. K. About the theory of so-called adsorption of soluble substances. *Sven. Vetenskapsakad. Handingarl*, 24:1–39, 1898.

[12] Ho, Y. S., Wase, D. A. J., Forster, C. F. Removal of lead ions from aqueous solution using sphagnum moss peat as adsorbent. *Water SA*, 22(3):219–224, 1996.

[13] Ritchie, A. G. Alternative to the Elovich equation for the kinetics of adsorption of gases on solids. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 73:1650–1653, 1977.

[14] Elovich, S. Y., Larinov, O. G. (1962). Theory of adsorption from solutions of non electrolytes on solid (I) equation adsorption from solutions and the analysis of its simplest form, (II) verification of the equation of adsorption isotherm from solutions. *Izv. Akad. Nauk. SSSR, Otd. Khim. Nauk*, 2(2):209–216, 1962.

[15] Soong, T. T., Random Differential Equations in Science and Engineering. New York, Academic Press, 1973.

[16] Ho, Y. S. Second-order kinetic model for the sorption of cadmium onto tree fern: a comparison of linear and non-linear methods. *Water research*, 40(1):119–125, 2006.

[17] Smith, R. C., Uncertainty quantification: theory, implementation, and applications. Siam, 2013.

42

Approximating Fixed Points by New and Fast Iterative Schemes

Puneet Sharma [♭,1] Vinay Kanwar[♮], Ramandeep Behl[♢] and Mithil Rajput[♠]

(♭) Department of Mathematics, Goswami Ganesh Dutta Sanatan Dharma College,
Chandigarh 160030, India.
(♮) University Institute of Engineering and Technology,
Panjab University, Chandigarh 160014, India.
(♢) Department of Mathematics, King Abdulaziz University,
Jeddah 21589, Saudi Arabia.
(♠) Indian Institute of Technology,
Kharagpur 721302, India.

1 Introduction

Let $(E, \|.\|)$ be a real Banach space, D be a nonempty closed and convex subset of E, $T : D \to D$ a self map of D and $x_0 \in D$. Suppose that $F_T = \{p \in D \mid T(p) = p\}$ is the set of all fixed points of T in D.

Starting with a suitable initial guess x_0, the recurrence process

$$x_{n+1} = Tx_n, \quad n \geq 0, \tag{1}$$

is known as the classical Picard's iteration. This iteration is locally convergent if $|T^{'}(x)| < 1$ for all $x \in [a, b]$.
Mann [1] defined the Mann iteration as

$$u_{n+1} = (1 - \alpha_n)u_n + \alpha_n \, Tu_n, \tag{2}$$

where $\{\alpha_n\}_{n=0}^{\infty}$ is a sequence of real numbers in $[0, 1]$. For $\alpha_n = 1$, (2) reduces to the well-known Picard iteration.
Imoru and Olatinwo [2] proved the stability of Picard and Mann iterative schemes using the following contractive condition: there exists $\delta \in [0, 1)$ and a monotone increasing and continuous function $\phi : [0, \infty) \to [0, \infty)$ with $\phi(0) = 0$ such that for each $x, y \in E$,

$$\|Tx - Ty\| \leq \phi(\|x - Tx\|) + \delta \, \|x - y\|. \tag{3}$$

Very recently, Kanwar et al. [3] have proposed a new geometrically constructed iterative scheme defined as follows:

$$x_{n+1} = \frac{mx_n + Tx_n}{m + 1}, \tag{4}$$

[1] puneetsharmamaths@gmail.com

where $m \geq 0$ is a real number. When $m = 0$, it reduces to the well-known Picard's iteration scheme. The basic idea here is to approximate the nonlinear function $y = T(x)$ by a linear approximation. Geometrically, the fixed point occurs where the graph of linear approximation

$$y = \frac{mx + T(x)}{m+1}, \tag{5}$$

intersects with the straight line $y = x$.
Khan [6] developed the Picard-Mann hybrid itertive process as:

$$\begin{cases} u_{n+1} = Tv_n, \\ v_n = (1 - \alpha_n)u_n + \alpha_n\, Tu_n, \end{cases} \tag{6}$$

where $\{\alpha_n\}_{n=0}^{\infty}$ is a sequence of real numbers in $(0, 1)$.

2 Development of new iterative scheme

There is a stiff competition among researchers to develop and identify the fastest fixed point iteration schemes. This tough race has made significant advancement to the fixed point iterations and fixed point theory in general. There are very few high quality iterative schemes available in the literature.

With this aim, we propose a new family of iterative schemes having higher precision than some of the widely used schemes in practice. The idea is to compute the fixed point using two nonlinear curves rather than taking a straight line and a nonlinear curve.
In what follows, equations $y = T(x)$ and (5) gives

$$T(x) = \frac{mx + T(x)}{m+1}. \tag{7}$$

Letting x_0 as the initial guess and x_1 as the first approximation to the required fixed point, one gets a new recursive process from (1) as

$$x_1 = \frac{(m+1)Tx_0 - Tx_1}{m}, \tag{8}$$

where $m > 0$ is a real number.
We can't predict x_1 in the left hand side unless x_1 in the right hand side is known. In order to get round off this difficulty, we approximate x_1 in the right hand side of equation (2) by our own scheme (4).
Without loss of generality, the general form of the above iterative scheme can be written as

$$x_{n+1} = \frac{1}{m}\left((m+1)Tx_n - Tx^*_{n+1}\right), \quad n \geq 0, \tag{9}$$

where $x^*_{n+1} = \dfrac{mx_n + Tx_n}{m+1}$ and $m > 0$ is a real number.

2.1 Two-step iterative schemes

We propose the two different new two-step iterative scheme as follows:
(i) First iteration scheme:
For x_0 in D, the sequence $\{x_n\}$ in D is defined by

$$\begin{cases} x_{n+1} = Ty_n, \\ y_n = \dfrac{1}{m}\left((m+1)Tx_n - T\left(\dfrac{mx_n + Tx_n}{m+1}\right)\right), \end{cases} \tag{10}$$

where $m \geq 1$ is a real number.
(ii)Second iteration scheme:
For x_0 in D, the sequence $\{x_n\}$ in D is defined by

$$\begin{cases} x_{n+1} = T\left(\dfrac{m^{'}y_n + Ty_n}{m^{'}+1}\right), \\ y_n = \dfrac{1}{m}\left((m+1)Tx_n - T\left(\dfrac{mx_n + Tx_n}{m+1}\right)\right), \end{cases} \tag{11}$$

where $m \geq 1$ and $m^{'} > 0$ are real numbers.

Theorem 16. *Let D be a nonempty closed and convex subset of real Banach space E and $T : D \to D$ be a mapping satisfying* (3) *with $\delta < \frac{1}{3}$ and p is a fixed point of T. Let $\{x_n\}_{n=0}^{\infty}$ be defined by the iteration process* (11) *and $x_0 \in D$, where $m \geq 1$ and $m^{'} > 0$ are real numbers. Then, $\{x_n\}_{n=0}^{\infty}$ converges strongly to a unique fixed point p of T.*

Theorem 17. *Let D be a nonempty closed and convex subset of real Banach space E and $T : D \to D$ be a mapping satisfying* (3) *with $\delta < \frac{1}{3}$ and p is a fixed point of T. Let $\{x_n\}_{n=0}^{\infty}$ be defined by the iteration process* (11)*, where $m \geq 1$ and $m^{'} > 0$ are real numbers. Then, the iterative scheme* (11) *is T-stable.*

Theorem 18. *Let D be a nonempty closed and convex subset of real Banach space E and $T : D \to D$ be a mapping satisfying* (3) *with $\delta < \frac{1}{3}$and p is a fixed point of T. Let $\{\alpha_n\}_{n=0}^{\infty}$ be a sequence of real numbers in* $(0, 1)$ *satisfying $\alpha_0 \leq \alpha_n < 1$ for all $n \in \mathbb{N}$ and $m \geq 1, m^{'} > 0$ are real numbers. For given $u_0 = x_0 \in D$, consider the iterative schemes $\{x_n\}_{n=0}^{\infty}$ and $\{u_n\}_{n=0}^{\infty}$ defined by the iteration processes* (11) *and* (6)*, respectively. Then, $\{x_n\}_{n=0}^{\infty}$ converges to p faster than $\{u_n\}_{n=0}^{\infty}$.*

2.2 Three-step iterative schemes

In this section, we propose the following new three-step iterative scheme as follows:
(i) First iteration scheme:
For x_0 in D, the sequence $\{x_n\}$ in D is defined by

$$\begin{cases} x_{n+1} = Ty_n, \\ y_n = T\left(\dfrac{m^{'}z_n + Tz_n}{m^{'}+1}\right), \\ z_n = \dfrac{1}{m}\left((m+1)Tx_n - T\left(\dfrac{mx_n + Tx_n}{m+1}\right)\right), \end{cases} \tag{12}$$

where $m \geq 1$ and $m^{'} > 0$ are real numbers.
(ii) Second iteration scheme:
For x_0 in D, the sequence $\{x_n\}$ in D is defined by

$$\begin{cases} x_{n+1} = T\left(\dfrac{m^{''}y_n + Ty_n}{m^{''}+1}\right), \\ y_n = T\left(\dfrac{m^{'}z_n + Tz_n}{m^{'}+1}\right), \\ z_n = \dfrac{1}{m}\left((m+1)Tx_n - T\left(\dfrac{mx_n + Tx_n}{m+1}\right)\right), \end{cases} \tag{13}$$

where $m \geq 1$, $m^{'} > 0$ and $m^{''} > 0$ are real numbers.

3 Numerical examples

In order to check the effectiveness of our schemes, we consider two different nonlinear equations which are illustrated in examples (1) and (2). We denote the iteration schemes (10), (11), (12) and (13) by ***Itr*** (10), ***Itr*** (11), ***Itr*** (12) and ***Itr*** (13), respectively. In Table 1, we have compared different two-step iterative schemes mentioned in [4], [5] and [6] with new two-step iterative schemes ***Itr*** (10) and ***Itr*** (11) . In Table 2, we have compared various three-step iterative schemes with new three-step iterative schemes ***Itr*** (12) and ***Itr*** (13). In Table 1 and 2, we have mentioned the results after fourteen iterations (i.e. $k = 14$). We have considered $\alpha_n = \beta_n = \gamma_n = \frac{3}{4}$, $m = 9$ and $m^{'} = m^{''} = \frac{1}{10}$. Computations are performed with the package *Mathematica* 12.0 with multiple precision arithmetic. The $a(\pm b)$ stands for $a \times 10^{\pm b}$. E.C. and R.E. stand for errors between two consecutive iterations and residual errors in the corresponding function by using the obtained fixed point, respectively.

Example 1. *Let us consider the following problem*

$$f(x) = \sin x - 10(x - 1). \tag{14}$$

Let $E = \mathbb{R}$ and $D = [0, 2]$. We can easily obtained the following fixed point iterative method $T : D \to D$ based on expression (14) *as*

$$T(x) = 1 + \frac{\sin x}{10}. \tag{15}$$

The required zero of expression (14) and fixed point for (15) is $p = 1.08859775239789$. We select $x_0 = 2$ as the initial guess for comparison.

Example 2. *Assume another test problem as following*

$$f(x) = x + 2 - e^x. \tag{16}$$

Let $E = \mathbb{R}$ and $D = (-\infty, 0]$. The corresponding fixed point iterative method $T : D \to D$ is given as follows:

$$T(x) = e^x - 2. \tag{17}$$

The required zero of expression (16) and fixed point for (17) is $p = -1.84140566043696$ with initial guess $x_0 = -1$.

Table 1: **Comparison of two-step iterative schemes on examples** (1) **and** (2) **with** $k = 14$

Examples	E.C. R.E.	*Ishikawa* [4]	*S iteration* [5]	*Picard-Mann* [6]	***Itr*** (10)	***Itr*** (11)
(1)	$\|x_{k+1} - x_k\|$	4.3(−9)	1.4(−24)	3.0(−27)	**2.2(−39)**	**1.2(−51)**
	$\|f(x_k)\|$	5.5(−8)	1.3(−23)	2.9(−26)	**2.1(−38)**	**1.1(−50)**
(2)	$\|x_{k+1} - x_k\|$	2.3(−8)	9.8(−16)	6.4(−18)	**6.8(−23)**	**1.1(−31)**
	$\|f(x_k)\|$	2.8(−8)	9.0(−16)	5.7(−18)	**5.9(−23)**	**8.9(−32)**

Table 2: **Comparison of three-step iterative schemes on examples** (1) **and** (2) **with** $k = 14$

Examples	E.C. R.E.	*Noor* [8]	*CR Iteration* [9]	*Piri et al.* [10]	***Itr*** (12)	***Itr*** (13)
	$\|x_{k+1} - x_k\|$	$4.1(-9)$	$3.2(-32)$	$1.5(-53)$	$\mathbf{2.5(-70)}$	$\mathbf{1.4(-82)}$
(1)	$\|f(x_k)\|$	$5.3(-8)$	$3.1(-31)$	$1.4(-52)$	$\mathbf{2.4(-69)}$	$\mathbf{1.3(-81)}$
	$\|x_{k+1} - x_k\|$	$1.5(-8)$	$8.8(-22)$	$3.8(-35)$	$\mathbf{6.8(-43)}$	$\mathbf{1.1(-51)}$
(2)	$\|f(x_k)\|$	$1.7(-8)$	$7.6(-22)$	$3.2(-35)$	$\mathbf{5.7(-43)}$	$\mathbf{9.0(-52)}$

4 Conclusions

We propose new classes of fixed point iterative schemes and prove that they converge faster than a number of existing iterations schemes. Backed by numerical examples, a comparative study with some existing ones manifests that our proposed schemes produce approximations of greater accuracy and excellent error estimates. Further, our proposed two-step iterative schemes perform better than some of the three-step iterative schemes available in the literature. Therefore, this paper further enriches and develops methods for computing fixed points and their applications in related fields.

References

[1] Mann, W.R., Mean value methods in iteration, *Proceedings of the American Mathematical Society* Vol. 4, 506–510 (1953).

[2] Imoru, C.O., Olatinwo, M.O., On the stability of Picard and Mann iteration processes, *Carpathian Journal of Mathematics*, Vol. 19, 155–160 (2003).

[3] Kanwar, V., Sharma, P., Argyros, I.K., Behl, R., Argyros, C., Ahmadian, A., Salimi, M., Geometrically constructed family of the simple fixed point iteration method, *Mathematics* Vol. 9, 1–13, (2021).

[4] Ishikawa, S., Fixed points by a new iteration method. *Proceedings of the American Mathematical Society* Vol. 44, 147–150 (1974).

[5] Agarwal, R., O'Regan, D., Sahu, D., Iterative construction of fixed points of nearly asymptotically nonexpansive mappings, *Journal of Nonlinear and Convex Analysis* Vol. 8, 61–79 (2007).

[6] Khan, S.H., A Picard-Mann hybrid iterative process, *Fixed Point Theory and Applications* Vol. 2013, 1–10, 69 (2013).

[7] Berinde, V., Iterative Approximation of Fixed Points. Berlin, Springer, (2007).

[8] Noor, M.A.: New approximation schemes for general variational inequalities, *Journal of Mathematical Analysis and Applications* Vol. 251, 217–229 (2000).

[9] Chugh, R., Kumar, V., Kumar, S., Strong convergence of a new three step iterative scheme in Banach spaces, *American Journal of Computational Mathematics* Vol. 2, 345–357 (2012).

[10] Piri, H., Daraby, B., Rahrovi, S., Ghasemi, M., Approximating fixed points of generalized α-nonexpansive mappings in banach spaces by new faster iteration process, *Numerical Algorithms*, Vol. 81, 1129-1148 (2019).

Higher-order multiplicative derivative iterative scheme to solve the nonlinear problems

G. Singh $^\flat$ S. Bhalla$^{\natural}$,[1] and R. Behl$^{\diamondsuit}$

($\flat$,$\natural$) Chandigarh University,
Department of Mathematics, Chandigarh University, Gharuan, Mohali, India.
($\diamondsuit$) King Abdulaziz University,
Department of Mathematics Sciences, King Abdulaziz University, Jeddah 21589, Saudi Arabia.

1 Introduction

In numerical analysis, solving nonlinear equations effectively is one of the most interesting tasks due to its wide application in engineering and sciences. Sometimes it is not possible to solve these problems analytically. Therefore, to solve these problems, the iterative numerical methods are used. The first well known method is Newton's method [1] used to approximate the root of nonlinear equation $g(x) = 0$ and defined as

$$x_{q+1} = x_q - \frac{g(x_q)}{g'(x_q)}, \quad q = 0, 1, 2, 3 \ldots, \tag{1}$$

where $g'(x_q)$ is the first-order ordinary derivative of function g at the point x_q of the qth iteration. For the simple root of the nonlinear equation the Newton's method has quadratic convergence. Several variants of Newton's method have been developed in order to obtain a faster convergence rate. Some of the well-known cubically convergent methods are Halley method [2], Euler's method [3], Super-Halley method [4], Weerakoon and Fernando [5] etc. All of the above mentioned methods consist of second order derivatives except Weerakoon and Fernando. From 1964 to 2012, researchers [6–9] has developed fourth order methods to find the roots of the nonlinear equations such as Traub and Ostrowski [6], King [7], Cordero and Torregrosa [8], Kanwar et al. [9] etc. Out of them, Kanwar et. al. [9] introduced a method which consists of second order derivatives while other listed methods has first-order derivative. Sometime it is difficult and time consuming to evaluations of the second order derivative of the function.

In the 70s of the 20th century multiplicative calculus introduced by Grossman and Katz [10] and several authors applied the multiplicative calculus in various branches like in 2008 Bashirov et al. [11] discussed the theoretical foundations as well as various applications of multiplicative calculus. Florack and Van Assen [12] used multiplicative calculus in biomedical image analysis, Filip and Piatecki [13] used it to investigate economic growth, Mısırlı and Gurefe [14], Riza et al. [15], and Özyapıcı and Mısırlı [16] used multiplicative calculus to develop multiplicative numerical methods and Bashirov et al. [17] used it to develop multiplicative differential equations. Bashirov and Riza [18] and Uzer [19] extended the multiplicative calculus to include complex valued functions of complex variables, which was previously applicable only to positive real valued functions of real

[1] sonia.e8843@cumail.in

variables.
In last few years, researchers use multiplicative derivatives for solving the nonlinear equations. With this same concept we constructed the fourth order Multiplicative King's method to approximate the root of function $g(x) = 0$.

2 The Proposed Method and Analysis of Convergence

The proposed King's iterative method in multiplicative derivative reprsented as

$$y_q = x_q - \frac{ln\ g(x_q)}{ln\ g^*(x_q)},$$

$$x_{q+1} = y_q - \left(\frac{log\ g(x_q) + \beta log\ g(y_q)}{log\ g(x_q) + (\beta-2) log\ g(y_q)}\right)\left(\frac{log\ g(y_q)}{log\ g^*(x_q)}\right). \tag{2}$$

Where q is iteration step, $g^*(x)$ is multiplicative derivative, and β is a free parameter.
Theorem: For an open interval I, let $r \in I$ be a multiplicative zero of a sufficiently multiplicative differential function $g : I \subseteq \mathbb{R} \to \mathbb{R}^+$, then multiplicative King's method has fourth order of convergence with error

$$x_{q+1} = (b_2^3 + 2\beta b_2^3 - b_2 b_3)e_q^4 + \mathcal{O}(e_q^5).$$

3 Numerical Examples

In this section, we solve the nonlinear equation $g(x) = 0$ using ordinary King's method [7] denoted as (KM), Chun method [21] denoted as (CM), Jnawali method [22] denoted as (JM) and the proposed multiplicative King's method denoted as (PM). The results obtained using these methods are presented in Table 1. All computations have done in Mathematica version 11.1.1 software and the stopping criteria $|x_{q+1} - x_q| < \epsilon$ and $\epsilon = 10^{-200}$ is used. We used a free parameter $\beta = 3$, 0.5 and -1. Moreover, the Approximated computational order of convergence(ACOC) is computed by using the following.

$$\rho \cong \frac{ln| \left|\dfrac{x_{q+1} - r}{x_q - r}\right|}{ln\left|\dfrac{x_q - r}{x_{q-1} - r}\right|}. \tag{3}$$

Numerical results indicate in the Table 1 that the proposed method execute less number of iterations and reduce the computational time.

Remark: The meaning of expression $m(\pm n)$ is $m \times 10^{\pm n}$ and d represents that scheme is divergent in all the tables.

Example 1. *We consider the population growth model that formulate the following nonlinear equation*

$$g(x) = \frac{1000}{1564}e^x + \frac{435}{1564}(e^x - 1) - 1.$$

In this model we evaluate the birth rate denoted as x, if in a specific local area has 1000 *thousand people at first and* 435 *thousand move into the local area in the first year. Likewise, assume*

1564 *thousand individuals toward the finish of one year. The computed results towards the root* $x_r = 0.1009979\ldots$ *are displayed in Table 1. Clearly, the method* PM *shows better results in terms of consecutive error and number of iteration in comparison of existing ones.*

β	*Method*	q	$\lvert x_q - x_{q-1}\rvert$	$\lvert g(x_q)\rvert$	ρ	*No. of iteration*	*C.P.U time*

3	KM	2
		3
		4
			4.4(−5)
			2.4(−18)
			2.1(−71)
				3.7(−5)
				2.0(−18)
				1.8(−71)
					4.000	5	0.32

3	PM	2
		3
		4
			2.7(−18)
			4.8(−74)
			4.6(−297)
				1.000	4.000	4	0.32

0.5	KM	2
		3
		4
			4.1(−7)
			3.8(−27)
			3.0(−107)
				3.5(−7)
				3.3(−27)
				2.5(−107)
					4.000	5	0.39

0.5	PM	2
		3
		4
			2.7(−19)
			2.4(−78)
			1.5(−314)
				1.000	4.000	4	0.26

−1	KM	2
		3
		4
			1.8(−5)
			1.8(−20)
			1.9(−80)
				1.5(−5)
				1.6(−20)
				1.6(−80)
					4.000	5	0.39

−1	PM	2
		3
		4
			2.2(−20)
			5.5(−83)
			2.0(−333)
				1.000	4.000	4	0.29

	CM	2
		3
		4
			1.8(−5)
			4.8(−20)
			2.5(−78)
				1.5(-5)
				4.1(−20)
				2.1(−78)
					4.000	5	0.37

	JM	2
		3
		4
			4.4(−6)
			8.8(−23)
			1.5(−89)
				3.7(−6)
				7.5(−23)
				1.3(−89)
					4.000	5	0.34

Table 1: Results of population growth model with initial guess $x_0 = 1$

4 Conclusions

By deriving multiplicative root finding method better approximations can be achieved with less computational time and complexity. In this paper, we developed the multiplicative King's method. We tested the proposed method for nonlinear equations and compared it to ordinary King's method,Chun method, and Jnawali method. The proposed method works efficiently in transcendental equations.

References

[1] B. Bradie, A Friendly Introduction to Numerical Analysis, Pearson Education Inc.: New Delhi, India, (2006).

[2] V. Kanwar and S.K. Tomar, Modified families of Newton, Halley and Chebyshev methods, Appl. Math. Comput. 192, 20–26 (2007).

[3] S. Amat, S. Busquier, and J. M. Gutierrez,Geometric Construction of Iterative Functions to Solve Nonlinear Equations, J. Comput. Appl. Math. 157(1), 197–205 (2003).

[4] J. Gutierrez and M. Hernandez, An accerlation of Newton's method, Appl. Math. Comput. 117(2-3), 223–239 (2001).

[5] S. Weekaroon and T.G.I. Fernando, A variant of Newton's method with accelerated third-order convergence, Appl. Math. Lett. 13, 87–93 (2002).

[6] J. F. Traub, Prentice-Hall, Iterative methods for the solution of equations, Englewood Cliffs, New Jersey, (1964).

[7] R. F. King, A Family of Fourth Order Methods for Nonlinear Equations. SIAM Journal on Numerical Analysis, 10(5), 876–879 (1973).

[8] A. Cordero, J. L. Hueso, E. M. Juan, R. Torregrosa, Steffensen type methods for solving nonlinear equations, J. Comput. Appl. Math. 236(12), 3058–3064 (2012).

[9] V. Kanwar, R. Behl, K. K. Sharma, Simply constructed family of a Ostrowski's method with optimal order of convergence, Comput. Math. with Appl., 62(11), 4021–4027(2011).

[10] M. Grossman, R. Katz, Non-Newtonian Calculus, Pigeon Cove, Lee Press, Massachusats (1972).

[11] A. E. Bashirov, E. M. Kurpınar, A. Özyapıcı, Multiplicative calculus and its applications, J. Math. Anal. Appl. 337(1), 36–48 (2008).

[12] L. Florack and H. V. Assen, Multiplicative calculus in biomedical image analysis, J. Math. Imaging Vis. 42(1), 64–75 (2012).

[13] D. A. Filip, C. Piatecki, A non-Newtonian examination of the theory of exogenous economic growth, CNCSIS-UEFISCSU (project number PNII IDEI 2366/2008) and Laboratoire d.Economie d.Orleans (LEO), (2010).

[14] E. Mısırlı, Y. Gürefe, Multiplicative adams bashforth-moulton methods, Numer. Algo. 57(4), 425–439 (2011).

[15] M. Riza, A. Özyapıcı, E. Mısırlı, Multiplicative finite difference methods. Q. Appl. Math. 67(4),745–754 (2009).

[16] E. Mısırlı, A. Özyapıcı, Exponential approximations on multiplicative calculus. In Proc. Jangjeon Math. Soc 12(2), 227–236 (2009).

[17] A.E. Bashirov, E. Mısırlı, Y. Tandoğdu, On modeling with multiplicative differential equations. Appl. Math. J. Chin. Univ. 26(4), 425–438 (2011).

[18] A. E. Bashirov, S. Norozpour, On complex multiplicative integration. J. Appl. Eng. Math. 7(1), 51–61 (2017).

[19] A. Uzer, Multiplicative type complex calculus as an alternative to the classical calculus, Comput. Math. with Appl. 60(10), 2725–2737 (2010).

[20] I. Cumhur, A. Gokdogan, E. Unal, Multiplicative Newton's Methods with Cubic Convergence. New Trends in Mathematical Science. 3, 299–307 (2017).

[21] C. Chun, M. Y. Lee, B. Neta, and J. Dzunic, On optimal fourth-order iterative methods free from second derivative and their dynamics. Appl. Math. Comput. 218(11), 6427-–6438 (2012).

[22] J. Jnawali, A Newton Type Iterative Method with Fourth-order Convergence. J. Inst. Engineering. 12(1), 87–95 (2017).

[23] Chicharro, F.I., Cordero, A., Torregrosa, J.R.:Drawing dynamical and parameters planes of iterative families and methods. Sci. World J. **11**, Article ID 780153 (2013).

A Seventh Order Steffensen type Iterative Method for Solving Systems of Nonlinear Equations and Applications

Sana Sultana$^\flat$ and Fiza Zafar$^{\flat,1}$

($\flat$) Centre for Advanced Studies in Pure and Applied Mathematics,
Bahauddin Zakariya University
Multan, 60800, Pakistan.

1 Introduction

Numerical analysis is a process in which we develop, analyze, and investigate numerous methods and algorithms for numerically solving problems in diverse domains such as electrical engineering, chemical engineering, mathematics, physics, and applied sciences. The most significant and difficult task is to find an efficient and accurate solution to systems of non-linear equations $F(X) = 0$ with $F : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$. For this purpose, many higher order iterative methods are developed that are efficient and have less computational cost such as [1, 2]. The simplest and most powerful root finding method to solve systems of non-linear equations is Newton-Raphson method, defined as:

$$X^{(n+1)} = X^{(n)} - F'(X^{(n)})^{-1}F(X^{(n)}),$$

where $F'(X^{(n)})^{-1}$ is the Jacobian matrix for the function F evaluated at the nth iterations. One of the major problems of numerical analysis is the construction of efficient iterative methods for solving nonlinear equations as well as systems of equations. There are different ways to construct iterative schemes for nonlinear systems: The natural way is to construct a method for scalar equations and then extend to the multidimensional case. Some researchers like [3] adopted this way to construct higher order schemes for nonlinear systems. However, the main practical difficulty associated with these methods is to calculate the first-order Fréchet derivative at each step of the computation. There are several real-life situations where the calculation of Fréchet derivatives is expensive and/or it requires a great deal of time for them to be given or calculated. So, it is obvious that all the modifications or variants of Newton's method has the same drawback. To overcome such problems, a class of derivative-free methods Steffensen type processes were introduced that replaces F' with operator:

$$[\varpi\left(X^{(k)}\right), X^{(k)}, F],$$

where ϖ: $\mathbb{R}^n \to \mathbb{R}^n$. Using these divided differences, some non-optimal and derivative-free higher-order convergent methods have also been proposed by researchers, for example, by Sharma and Arora [4], Wang and Zhang [5] etc.

[1] fizazafar@bzu.edu.pk

We, in here, present a new seventh-order Steffensen-type family of methods based on weight functions using four function evaluations that make it valuable for solving nonlinear systems. Moreover, we show its applicability chemical equilibrium problem, neurophysiology application, and economics modeling application.

1.1 Formulation of the Seventh Order Scheme for Systems of Nonlinear Equations

We develop a three-step seventh-order method to solve systems of non-linear equations, which is described:

$$\begin{aligned} Y^{(n)} &= X^{(n)} - (F'(X^{(n)}))^{-1}(F(X^{(n)})), \\ Z^{(n)} &= Y^{(n)} - H((\theta^{(n)}))F[X^{(n)}, Y^{(n)}]^{-1}F(Y^{(n)}), \\ X^{(n+1)} &= Z^{(n)} - (G(\theta^{(n)}) + L(\nu^{(n)}))F[X^{(n)}, Y^{(n)}]^{-1}F(Z^{(n)}). \end{aligned} \tag{1}$$

For multidimensional vector-valued function F, we can write $\theta^{(n)}$ as:

$$\theta^{(n)} = I - (F'(X^{(n)}))^{-1}F[X^{(n)}, Y^{(n)}].$$

Similarly, for multidimensional vector-valued function F, we can write $\nu^{(n)}$ as:

$$\nu^{(n)} = I - H(\theta^{(n)})(F[X^{(n)}, Y^{(n)}])^{-1}F[Y^{(n)}, Z^{(n)}].$$

where H, $G : M_{n\times n}(R) \to \Gamma(R^n)$, $L : N_{n\times n}(R) \to \Gamma(R^n)$, such that $M_{n\times n}$, $N_{n\times n}$ are the set of $n \times n$ square matrices respectively and $\Gamma(R^n)$, is the set of linear operators from R^n to R^n.

Theorem 19. *Let us suppose that Ψ is the root of the function, $F : \Omega \subseteq R^n \to R^n$, be a Frechet differentiable function in the neighborhood of closed interval Ω of its root Ψ. Let F' be continuous and non-singular at Ψ. Furthermore, the convergence of the described scheme is confirmed if we consider that the initial guess $X^{(0)}$ is close to Ψ. Then our defined method has convergence order seven with the following defined conditions on the weigth functions:*

$$\begin{aligned} H_0 &= H(0) = I,\ H_1 = H'(0) = I, \\ G_0 &= G(0) = I,\ G_1 = G'(0) = I,\ G_2 = G''(0) = (2 + a)I, \\ L_0 &= L(0) = 0,\ L_1 = L'(0) = I \\ \textit{where } H_2 &= H''(0) < \infty,\ G_3 = G'''(0) < \infty, L_2 = L''(0) < \infty, \end{aligned}$$

The expression for the final error equation:

$$\begin{aligned} E^{(n+1)} = &\frac{1}{12}C_2^2(2C_3 - 6C_2^2 + H''(0)C_2^2) \times (3H''(0)C_2^2 - 48C_2^2 + G'''(0)C_2^2 - C_2^2H'''(0) + 18C_2)E^{(n)^7} \\ &+ ... + O(E^{(n)^9}). \end{aligned}$$

From Theorem 1, several cases of our proposed scheme (1) are obtained for various choices of the weight functions. We consider some specific cases of weight functions for the proposed multidimensional scheme:

Case 1 WF1: We get a seventh-order scheme for multidimensional case by taking the weight functions as:

$$H(\theta^{(n)}) = [I - (\theta^{(n)})]^{-1},$$

$$
\begin{aligned}
G(\theta^{(n)}) &= [I + \theta^{(n)} + (\theta^{(n)})^2], \\
L(\nu^{(n)}) &= [\nu^{(n)} + \frac{1}{2}(\nu^{(n)})^2].
\end{aligned}
$$

Case 2 WF2: We get another seventh-order scheme for multidimensional case by taking the weight functions as:

$$
\begin{aligned}
H(\theta^{(n)}) &= [I + \theta^{(n)}], \\
G(\theta^{(n)}) &= [I + (\theta^{(n)})^2]^{-1}, \\
L(\nu^{(n)}) &= [\nu^{(n)} + \frac{1}{2}(\nu^{(n)})^2].
\end{aligned}
$$

Case 3 WF3: We consider a seventh-order scheme for multidimensional case by taking the weight functions as:

$$
\begin{aligned}
H(\theta^{(n)}) &= [I - (\theta^{(n)})]^{-1}, \\
G(\theta^{(n)}) &= [I - \theta^{(n)} + \frac{3}{2}(\theta^{(n)})^2]^{-1}, \\
L(\nu^{(n)}) &= [\nu^{(n)} + \frac{1}{2}(\nu^{(n)})^2].
\end{aligned}
$$

1.2 Numerical Results

In this section, we consider some examples of a system of non-linear equations and solve the systems by using proposed schemes WF1, WF2, and WF3. In order to check the efficiency of our proposed schemes, we also compare these results with a Steffensen type method developed by Abad et al. [3], denoted by AC having seventh order of convergence. Tables show the comparison of seventh order scheme results by listing errors in consecutive iteration $\left\|X^{(n)} - X^{(n-1)}\right\|_\infty$, and the residual error in function $\left\|H\left(X^{(n)}\right)\right\|_\infty$. Software MATLAB R2014a is used to perform all the computations for the multivariate case.

Example 1. *Let us consider a chemical equilibrium problem from [6], which is given below in the form of system of nonlinear equations:*

$$
\begin{aligned}
x_1x_2 + x_1 - 3y_5 &= 0, \\
2x_1x_2 + x_1 + 2R_{10}x_2^2 + x_2x_3^2 + R_7x_2x_3 + R_9x_2x_4 + R_8x_2 - Rx_5 &= 0, \\
2x_2x_3^2 + R_7x_2x_3 + 2R_5x_3^2 + R_6x_3 - 8x_5 &= 0, \\
R_9x_2x_4 + 2x_4^2 + 4Rx_5 &= 0, \\
x_1\left(x_2 + 1\right) + R_{10}x_2^2 + x_2x_3^2 + R_7x_2x_3 + R_9x_2x_4 + R_8x_2 + R_5x_3^2 + R_6x_3 + x_4^2 - 1 &= 0,
\end{aligned}
$$

where

$$
\begin{aligned}
R = 10,\ R_5 = 0.193,\ R_6 = \frac{0.002597}{\sqrt{40}},\ R_7 = \frac{0.003448}{\sqrt{40}} \\
R_8 = \frac{0.00001799}{40},\ R_9 = \frac{0.0002155}{\sqrt{40}},\ R_{10} = \frac{0.00003846}{40}.
\end{aligned}
$$

The system of above equations has four real, twelve complex, and an infinite number of solutions at infinity. One of real solution of our problem is $\Psi = (0.0031141, 34.5979245, 0.0650417, 0.8593780, 0.0369518)$ *and our initial root is:* $X^{(0)} = (0.001, 35, 0.05, 0.5, 0.01)$.

Table 1: Numerical Solution of CE problem

Cases	n	$\left\|X^{(n)} - X^{(n-1)}\right\|_\infty$	$\left\|H\left(X^{(n)}\right)\right\|_\infty$
WF1	1	0.7481	0.6858
	2	0.3461	0.3423
	3	3.3322(−8)	0.3232(−7)
WF2	1	0.3460	4.54(−2)
	2	0.2466	0.3085(−11)
	3	1.024(−10)	0.4318(−43)
WF3	1	0.3866	0.0013
	2	0.1544(−1)	0.1741(−16)
	3	3.2487(−16)	0.3004(−46)
AC	1	0.7481	0.6858
	2	0.3461	0.3423
	3	3.3322(−8)	0.3232(−7)

Example 2. *Consider another problem which is a neurophysiology application from [7] which considers the following nonlinear system:*

$$\begin{aligned} x_1^2 + x_3^2 &= 1, \\ x_2^2 + x_4^2 &= 1, \\ x_5 x_3^3 + x_6 x_4^3 &= c_1, \\ x_5 x_1^3 + x_6 x_2^3 &= c_2, \\ x_5 x_1 x_3^2 + x_6 x_2 x_4^2 &= c_3, \\ x_5 x_3 x_1^2 + x_6 x_4 x_2^2 &= c_4, \end{aligned}$$

where c_i, $i = 1, ..., 4$ *can be taken arbitrarily, we considered* $c_i = 0$, *for each* i. *The desired solution of our problem is* $(-0.8282192, 0.5446434, -0.0094437, 0.7633676, 0.0199325, 0.1466452)$ *and our initial root is:* $X^{(0)} = (-0.5, 0, -0.005, 0.5, 0.02, 0.10)$.

Table 2: Numerical Solution of NP problem

Cases	n	$\left\|X^{(n)} - X^{(n-1)}\right\|_\infty$	$\left\|H\left(X^{(n)}\right)\right\|_\infty$
WF1	1	0.5001	0.0290
	2	0.2895(−1)	0.7917(−15)
	3	7.9169(−16)	0.1593(−57)
WF2	1	5.0143(2)	326.6471
	2	4.5155(2)	49.9890
	3	4.9989(1)	0.4660(−21)
WF3	1	0.5001	0.0290
	2	0.2895(−1)	0.7917(−15)
	3	7.9169(−16)	0.1593(−57)
AC	1	9.7416(1)	9.7416(1)
	2	9.5409(1)	1.4101(2)
	3	1.9066	1.9104

Example 3. *Lastly, we consider another example that is an economics modeling application. The modeling problem is considered as difficult and can be scaled up to arbitrary dimensions [8]. The*

problem is given by the following system of nonlinear equations:

$$\left(x_k + \sum_{i=1}^{n-k-1} x_i x_{i+k}\right) x_n - c_k = 0, \qquad 1 \le k \le n-1,$$
$$\sum_{l=1}^{n-1} x_i + 1 = 0.$$

The constants c_k can be randomly chosen. We considered the value 0 for the constants in our experiment and the case of four equations. So, for $n = 4$, our four equations becomes:

$$\begin{aligned} (x_1 + x_1x_2 + x_2x_3)x_4 - c_1 &= 0, \\ (x_2 + x_1x_2)x_4 - c_2 &= 0, \\ x_3x_4 - c_3 &= 0, \\ x_1 + x_2 + x_3 + 1 &= 0. \end{aligned}$$

The desired solution of our problem is $\Psi = (-0.1639324, -0.3813209, 0.2242448, -0.0755094)$ and our initial root is: $X^{(0)} = (-0.10, -0.35, 0.20, -0.05)$.

Table 3: Numerical Solution of EM problem

Cases	n	$\left\|X^{(n)} - X^{(n-1)}\right\|_\infty$	$\left\|H\left(X^{(n)}\right)\right\|_\infty$
WF1	1	8.3974(7)	1.6580(22)
	2	2.3511(6)	0.2289(−9)
	3	1.4901(−8)	0.5798(−59)
WF2	1	9.8997(8)	2.3443(25)
	2	1.0546(7)	0.7887(−7)
	3	3.9577(−8)	0.4557(−55)
WF3	1	8.3974(7)	1.6578(22)
	2	2.3509(6)	0.1132(−9)
	3	1.4901(−8)	0.9589(−42)
AC	1	7.9364(1)	7.9348(1)
	2	1.6727(1)	1.0444(5)
	3	1.8102(−14)	0.1810(−13)

The numerical results show that our newly developed seventh-order method is better or equally efficient as that of already existing methods.

References

[1] Ostrowski, A. M., Solution of Equations and Systems of Equations, Academic Press, New York, 1960.

[2] Traub, J. F., Iterative Methods for Solution of Equations, Prentice-Hall, New Jersey, 1964.

[3] Abad, M., Cordero, A., Torregrosa, J. R., A Family of Seventh-Order Schemes for Solving Nonlinear Systems, Bull. Math. Soc. Sci. Math., 57: 133-145, 2014.

[4] Sharma, J.R., Arora, H., A novel derivative free algorithm with seventh order convergence for solving systems of nonlinear equations, Numer. Algor. 67: 917–933, 2014.

[5] Wang, X. and Zhang, T., A family of Steffensen type methods with seventh-order convergence, Numer. Algor. 62(3): 429–444, 2013.

[6] Meintjes, K., Morgan, A.P., Chemical equilibrium systems as numerical test problems, ACM Trans. Math. Softw., 16(2): 143–151, 1990.

[7] Verschelde, J., Verlinden, P., Cools, R., Homotopies exploiting Newton polytopes for solving sparse polynomial systems, SIAM J. Numer. Anal., 31(3): 915–930, 1994.

[8] Morgan, A.P., Solving polynomial systems using continuation for scientific and engineering problems, Englewood Cliffs, NJ: Prentice-Hall, 1987.

Modifying Kurchatov's method to find multiple roots

A. Cordero$^\flat$, N. Garrido$^\flat$, J.R. Torregrosa$^\flat$ and P. Triguero-Navarro$^{\flat,1}$

($\flat$) Instituto de Matemática Multidisciplinar, Universitat Politècnica de València,
Camino de Vera, s/n, 46022-Valencia, Spain,
acordero@mat.upv.es, neugarsa@mat.upv.es, jrtorre@mat.upv.es, ptrinav@doctor.upv.es

1 Introduction

In many problems in engineering or applied mathematics, it is necessary to solve nonlinear equations $f(x) = 0$. They cannot always be solved exactly, which is why iterative methods appear to solve them. A well-known one is Newton's method, which has the following expression:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \text{ for } k = 0, 1, \ldots$$

To ensure the convergence of Newton's method, the derivative of the function evaluated in the solution must be non-zero, that is, the solution must be a simple root of $f(x) = 0$. This is not always the case. For this reason, iterative methods appear that allow us to obtain solutions with a multiplicity greater than 1.

One of them is the following modification of Newton's method, that can be find in [6], where m is the multiplicity of the solution of the equation.

$$x_{k+1} = x_k - m\frac{f(x_k)}{f'(x_k)}, \text{ for } k = 1, 2, 3 \ldots$$

In order to be able to apply this method, we must know the multiplicity of the solution in a priori.

To avoid the need to know the multiplicity in advance, iterative methods for multiple roots are designed that do not use this multiplicity in their iterative expression, see [2].

In this way, we propose the following iteratives methods based on Kurchatov's method.

To estimate the roots of $f(x) = 0$, we define the following method, denoted by KM,

$$x_{k+1} = x_k - \frac{g(x_k)}{g[2x_k - x_{k-1}, x_{k-1}]}, \qquad k = 0, 1, 2, \ldots$$

where $g(x) = \frac{f(x)}{f'(x)}$ and $g[y, z](y - z) = g(y) - g(z)$.

To calculate the expression of $g(x)$ in the previous method we use the derivative of the function to be solved. We can replace this derivative by a divided difference operator, so that to estimate the roots of $f(x) = 0$, we define the following method, denoted by KMD,

$$x_{k+1} = x_k - \frac{g(x_k)}{g[2x_k - x_{k-1}, x_{k-1}]}, \qquad k = 0, 1, 2, \ldots$$

[1] ptrinav@doctor.upv.es

where $g(x) = \frac{f(x)}{f[x+f(x),x]}$.

In this paper, we will analyse the order of convergence of the proposed methods and we will also perform numerical experiments to illustrate the behaviour of them.

2 Convergence analysis

Theorem 20. *Let $f : \mathbb{C} \longrightarrow \mathbb{C}$ be a sufficiently differentiable function in an neighbourhood of α, which we denote by $D \subset \mathbb{C}$, such that α is a multiple zero of $f(x) = 0$ with unknown multiplicity $m \in \mathbb{N}-\{1\}$. Then, taking an estimate x_0 close enough to α, the sequence of iterates $\{x_k\}$ generated by method KM converges to α with order 2.*

Proof. We first obtain the Taylor expansion of $f(x_k)$ around α where $e_k = x_k - \alpha$:

$$f(x_k) = \frac{f^{(m)}(\alpha)}{m!}\left(e_k^m + C_1 e_k^{m+1} + C_2 e_k^{m+2} + C_3 e_k^{m+3}\right) + O(e_k^{m+4}),$$

being $C_j = \frac{m!}{(m+j)!}\frac{f^{(m+j)}(\alpha)}{f^{(m)}(\alpha)}$ for $j = 1, 2, \ldots$

Calculating the derivative of the above expression we obtain

$$f'(x_k) = \frac{f^{(m)}(\alpha)}{m!}\left(me_k^{m-1} + (m+1)C_1 e_k^m + (m+2)C_2 e_k^{m+1} + (m+3)C_3 e_k^{m+2}\right) + O(e_k^{m+3}).$$

Now. we calculate $g(x_k)$

$$g(x_k) = \frac{f(x_k)}{f'(x_k)} = \frac{1}{m}\left(e_k - \frac{1}{m}C_1 e_k^2 + \frac{(m+1)C_1^2 - 2mC_2}{m^2}e_k^3\right) + O(e_k^4).$$

In an equivalent way, we obtain the following expressions for $g(x_{k-1})$ and $g(2x_k - x_{k-1})$

$$g(x_{k_1}) = \frac{f(x_{k-1})}{f'(x_{k-1})} = \frac{1}{m}\left(e_{k-1} - \frac{1}{m}C_1 e_{k-1}^2 + \frac{(m+1)C_1^2 - 2mC_2}{m^2}e_{k-1}^3\right) + O(e_{k-1}^4),$$

$$\begin{aligned} g(2x_k - x_{k-1}) = {} & \frac{f(x_k)}{f'(2x_k - x_{k-1})} = \\ & = \frac{1}{m}\left(2e_k - e_{k-1} - \frac{1}{m}C_1(2e_k - e_{k-1})^2 + \frac{(m+1)C_1^2 - 2mC_2}{m^2}(2e_k - e_{k-1})^3\right) + O_4(e_k, e_{k-1}), \end{aligned}$$

with $e_{k-1} = x_{k-1} - \alpha$.

From the above relations, we obtain

$$\begin{aligned} g[2x_k - x_{k-1}, x_{k-1}] &= \frac{g\left(2x_k - x_{k-1}\right) - g\left(x_{k-1}\right)}{2(x_k - x_{k-1})} \\ &= \frac{1}{m}\left(1 - \frac{2}{m}C_1 e_k + \frac{(m+1)C_1^2 - 2mC_2}{m^2}\left(4e_k^2 - 2e_j e_{k-1} + e_{k-1}^2\right)\right) + O_3(e_k, e_{k-1}). \end{aligned}$$

Thus, applying the above relationship, the following error equation is obtained:

$$\begin{aligned} x_{k+1} - \alpha &= x_k - \alpha - \frac{g(x_k)}{g[2x_k - x_{k-1}, x_{k-1}]} \\ &= \frac{-1}{m}C_1 e_k^2 + \frac{(m+1)C_1^2 - 2mC_2}{m^2}\left(-5e_k^3 + 2e_k^2 e_{k-1} - e_k e_{k-1}^2\right) + O_4(e_k, e_{k-1}). \end{aligned}$$

We have some different possibilities for the behaviour of e_{k+1} respect to e_k and e_{k-1}.

By the expression, we only are going to take into account if the behaviour is like e_k^2 or $e_ke_{k-1}^2$, because e_k^3 and $e_k^2e_{k-1}$ converge faster to 0 than e_k^2.

Then,

$$e_{k+1} \sim \frac{-1}{m}C_1e_k^2 - \frac{(m+1)C_1^2 - 2mC_2}{m^2}e_ke_{k-1}^2.$$

- If $e_{k+1} \sim e_k^2$, then the order of convergence is 2.
- Now, we suppose that $e_{k+1} \sim e_ke_{k-1}^2$. We assume that the method has R-order p, that means,

  $$e_{k+1} \sim e_k^p.$$

  In the same way, $e_k \sim e_{k-1}^p$. From the above relations, we get

  $$e_{k+1} \sim e_{k-1}^{p^2}.$$

  Then, the error equation is

  $$e_{k+1} \sim e_ke_{k-1}^2 \sim e_{k-1}^{p+2}.$$

  By equating the exponents of e_{k-1} of the above relations, we obtain the following polynomial $p^2 - p - 2 = 0$, whose only positive root is $p = 2$, then the order of convergence of the method is 2.

■

Theorem 21. *Let $f : \mathbb{C} \longrightarrow \mathbb{C}$ be a sufficiently differentiable function in an neighbourhood of α, which we denote by $D \subset \mathbb{C}$, such that α is a multiple zero of $f(x) = 0$ with unknown multiplicity $m \in \mathbb{N}-\{1\}$. Then, taking an estimate x_0 close enough to α, the sequence of iterates $\{x_k\}$ generated by method KMD converges to α with order 2.*

Proof. We first obtain the Taylor expansion of $f(x_k)$ around α where $e_k = x_k - \alpha$:

$$f(x_k) = \frac{f^{(m)}(\alpha)}{m!}\left(e_k^m + C_1e_k^{m+1}\right) + O(e_k^{m+2}).$$

being $C_j = \frac{m!}{(m+j)!}\frac{f^{(m+j)}(\alpha)}{f^{(m)}(\alpha)}$ for $j = 1, 2, \ldots$

In the same way,

$$f(x_k + f(x_k)) = \frac{f^{(m)}(\alpha)}{m!}\left((e_k + f(x_k))^m + C_1\left(e_k + f(x_k)\right)^{m+1}\right) + O(e_k^{m+2}).$$

Then,

$$f(x_k + f(x_k)) - f(x_k) = \frac{f^{(m)}(\alpha)}{m!}\left((e_k + f(x_k))^m - e_k^m + C_1\left((e_k + f(x_k))^{m+1} - e_k^{m+1}\right)\right) + O(e_k^{m+2}).$$

Using Newton's binomial and the Taylor expansion of $f(x_k)$ around α we obtain that

$$\frac{f(x_k + f(x_k)) - f(x_k)}{x_k + f(x_k) - x_k} = \frac{f^{(m)}(\alpha)}{m!}\left(me_k^{m-1} + (m+1)C_1e_k^m\right) + O(e_k^{m+1}).$$

We then calculate $g(x_k)$ from the above expressions.

$$g(x_k) = \frac{f(x_k)}{f[x_k + f(x_k), x_k]} = \frac{e_k^m + C_1e_k^{m+1} + O(e_k^{m+2})}{me_k^{m-1} + (m+1)C_1e_k^m + O(e_k^{m+1})}$$

$$= \frac{1}{m}\left(e_k - \frac{1}{m}C_1 e_k^2\right) + O(e_k^3).$$

In an equivalent way, we obtain the following expressions for $g(x_{k-1})$ and $g(2x_k - x_{k-1})$

$$g(x_{k-1}) = \frac{1}{m}\left(e_{k-1} - \frac{1}{m}C_1 e_{k-1}^2\right) + O(e_{k-1}^3),$$

$$g(2x_k - x_{k-1}) = \frac{1}{m}\left(2e_k - e_{k-1} - \frac{1}{m}C_1(2e_k - e_{k-1})^2\right) + O_3(e_k, e_{k-1}),$$

with $e_{k-1} = x_{k-1} - \alpha$.

Then, appyling the above relations, we obtain that

$$\begin{aligned} g[2x_k - x_{k-1}, x_{k-1}] &= \frac{g(2x_k - x_{k-1}) - g(x_{k-1})}{2(x_k - x_{k-1})} \\ &= \frac{1}{m}\left(1 - \frac{2}{m}C_1 e_k\right) + O_2(e_k, e_{k-1}). \end{aligned}$$

Thus, the following error equation is obtained

$$\begin{aligned} x_{k+1} - \alpha &= x_k - \alpha - \frac{g(x_k)}{g[2x_k - x_{k-1}, x_{k-1}]} \\ &= -\frac{1}{m}C_1 e_k^2 + e_k O_2(e_k, e_{k-1}) + O(e_k^3). \end{aligned}$$

We have some different possibilities for the behaviour of e_{k+1} respect to e_k and e_{k-1}.

By the expression, we only are going to take into account if the behaviour is like e_k^2 or $e_k e_{k-1}^2$, because e_k^3 and $e_k^2 e_{k-1}$ converge faster to 0 than e_k^2.

Then

- If $e_{k+1} \sim e_k^2$, then the order of convergence is 2.
- If we assume that $e_{k+1} \sim e_k e_{k-1}^2$. Then, we assume that the method has R-order p, that means,

  $$e_{k+1} \sim D_{k,p} e_k^p.$$

  At the same time, $e_k \sim e_{k-1}^p$, then we obtain that

  $$e_{k+1} \sim e_{k-1}^{p^2}.$$

  From the error equation and the last relation, we obtain that

  $$e_{k+1} \sim e_k e_{k-1}^2 \sim e_{k-1}^{p+2}.$$

  By equating the exponents of e_{k-1} of the last two equation, we obtain the following polynomial $p^2 - p - 2 = 0$, whose only positive root is $p = 2$, then the order of convergence of the method is 2.

■

3 Numerical experiments

Now, we perform a numerical experiment to see the behaviour of the two proposed iterative methods. For the computational calculations, we use Matlab R2020b with an arithmetic precision of 500 digits iterating from an initial estimate x_0 until it is verified that the absolute value of the function evaluated in the iteration is less than 10^{-50}, that is,

$$|f(x_k)| < 10^{-50}.$$

The numerical results we are going to compare the methods in the different examples are:

- the approximation obtained,
- the norm of the equation evaluated in that approximation,
- the norm of the distance between the last two approximations,
- the number of iterations necessary to satisfy the required tolerance,
- the computational time and the approximate computational convergence order (ACOC), defined by Cordero and Torregrosa in [3], which has the following expression

$$p \approx ACOC = \frac{\ln(|x_{k+1} - x_k|/|x_k - x_{k-1}|)}{\ln(|x_k - x_{k-1}|/|x_{k-1} - x_{k-2}|)}.$$

The equation we try to solve is $f(x) = (x^2 - 1)^3$, which has two roots with multiplicity 3.

Table 1: Results for $(x^2 - 1)^3 = 0$.

	x_0	x_{-1}	$\|x_{k+1} - x_k\|$	$\|g(x_{k+1})\|$	Iter	ACOC
KM	0.5	0.1	3.3307e-16	0	7	2.0058
KMD	0.5	0.1	9.7478e-14	0	9	1.7006

The results obtained for each of the methods for the function to be solved are shown in Table 1. We can see from the Tables that in all cases the ACOC is close to the theoretical convergence order shown above. It can be seen that the best results for this numerical experiment are obtained with the KM method. Both methods give good approximations to the solution, although the KM method performs less iterations to verify the stopping criterion.

4 Conclusions

In this work, we have studied two iterative methods for multiple roots with memory, obtaining that the order of convergence of them is 2. These iterative methods do not use the multiplicity of the root in their iterative expression, so it is not necessary to know this multiplicity before applying the iterative method. In the numerical experiments, we have verified the theoretical results concerning the order of convergence of the methods.

Financial disclosure

This research was partially supported by Spanish Ministerio de Ciencia, Innovación y Universidades PGC2018-095896-B-C22, by the internal research project ADMIREN of Universidad Internacional de La Rioja (UNIR) and was partially supported by Universitat Politécnica de Valéncia Contrato Predoctoral PAID-01-20-17 (UPV).

References

[1] F.I. Chicharro, R. Contreras & N. Garrido. A Family of Multiple-Root Finding Mathematics, 8, 2020.

[2] A. Cordero, B. Neta & J.R. Torregrosa. Memorizing Schröder's Method as an Efficient Strategy for Estimating Roots of Unknown Multiplicity. Mathematics, 9, 2021.

[3] A. Cordero, A. & J.R. Torregrosa. Variants of Newton's method using fifth-order quadrature formulas. Appl. Math. Comput., 190. 2007.

[4] B. Neta, C. Chun & M. Scott. On the development of iterative methods for multiple roots. Appl. Math. Comput., 224. 2013.

[5] S.M. Shakhno. On a Kurchatov's method of linear interpolation for solving nonlinear equations. PAMM Proc. Appl. Math. Mech. 4, 650–651, 2004.

[6] E. Schröder, Über unendlich viele Algorithmen zur Auflösung der Gleichungen, Math. Ann. 2, 317–365, 1870.

[7] F. Zafar, A. Cordero, J.R. Torregrosa & M. Penkova. A family of optimal fourth-order methods for multiple roots of nonlinear equations. Math Meth Appl Sci, 43. 2020.begindocument

The Relativistic Anharmonic Oscillator within a Double-Well Potential

Michael M. Tung ♭[1] and Frederic Rapp ♮

(♭) Instituto de Matemática Multidisciplinar,
Universitat Politècnica de València,
Camino de Vera, s/n, 46022 Valencia, Spain.

(♮) Fakultät 6: Luft- und Raumfahrttechnik und Geodäsie,
Universität Stuttgart,
Pfaffenwaldring 21, 70569 Stuttgart, Germany.

1 Introduction

Non-harmonic oscillations described by a double-well potential allow to model various physical phenomena and provide insight into their properties without overly simplifying the problem, thereby extending the more elementary harmonic case. Important applications of the anharmonic oscillator range from classical instantons and field theory to quantum theory and particle physics [1–3].

This presentation focuses on the treatment of the anharmonic oscillator in connection with the double-well potential including the non-negligible relativistic effects of a strong and uniform gravitational field. It extends previous work of the relativistic harmonic oscillator to the anharmonic model [4].

Our analysis parts from a relativistic action principle, where the dynamics is described as usual by the corresponding Euler-Lagrange equations. However, complications arise on how to handle the relativistic potential of the strong gravitational field. We proceed by deriving an integral equation for this potential following an approach—a method introduced by Goldstein & Bender for the brachistochrone problem [5] and more recently applied to the pendulum [6, 7].

Finally, a numerical simulation of the relativistic results is carried out to examine the dynamics of the model and compare it with the purely classical calculations. A detailed study of the phase space for both models will enable us to detect very distinct features among them, and thereby confirms the necessity for including relativistic corrections to yield reliable predictions for the strong-gravity case.

2 Methods

2.1 Lagrangian approach

Lagrangian mechanics is based on a Hamilton's principle (see e.g. Ref. [8]). The Lagrangian $L = T - V$ consists of the kinetic energy T and the potential energy V. In this model, the potential

[1] mtung@mat.upv.es

energy consists of the standard gravitational potential and also the double-well potential given by

$$V_{\text{double-well}} = \beta x^4 - \alpha x^2, \tag{1}$$

where α and β are positive parameters to control the typical double-well shape of the potential. In particular, the ratio α/β determines the shape of the double-well. The behaviour around the origin is shown in Figure 1, where the red curve represents a high α/β-ratio, the blue curve a low ratio, and the green curve a balanced ratio, respectively.

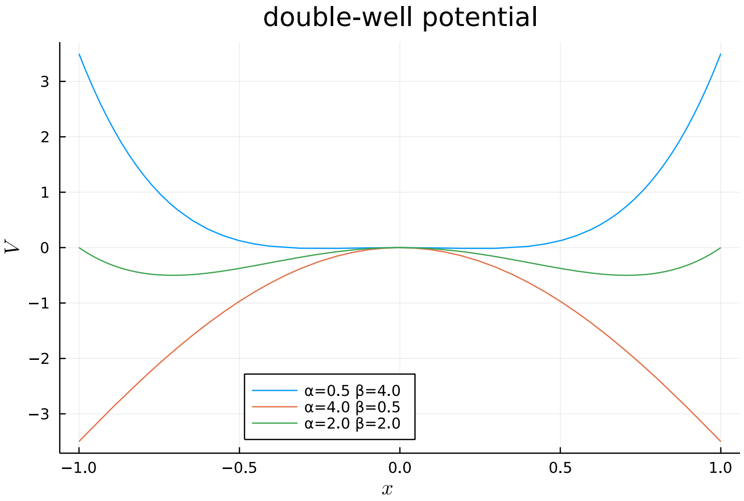


Figure 1: Double-well potential, *viz.* Eq. (1), for different parameters $\alpha, \beta > 0$.

Substituting for the potential Eq. (1), the deterministic equations of motion will result from the following variational principle by varying over all possible paths $x(t)$, while keeping the two end points fixed:

$$\delta \int dt\, L(x, \dot{x}) = \delta \int dt \left[\frac{1}{2} m_0 \dot{x}^2 - m_0 g x - \beta x^4 + \alpha x^2\right] = 0. \tag{2}$$

Here, as usual, g is the gravitational constant appearing in the gravitational potential, $V_g(x) = m_0 g x$, with inertial mass m_0. Thus, we consider a one-dimensional dynamical system, and the corresponding Euler-Lagrange equation are derived from

$$\left(\frac{d}{dt}\frac{\partial}{\partial \dot{x}} - \frac{\partial}{\partial x}\right) L = 0, \tag{3}$$

which then leads to the equations of motion for the classical case:

$$m_0 \ddot{x} + m_0 g + 4\beta x^3 - 2\alpha x = 0. \tag{4}$$

2.2 Relativistic generalization

For the generalization of Eq. (2) to special relativity, it is essential to regard the mass as a relativistic quantity, which thus will depend on the actual velocity of the body in motion $\dot{x}(t)$. The relation between rest mass, m_0, and the relativistic mass, m, is expressed by $m = \gamma m_0$, with the relativistic factor $\gamma = 1/\sqrt{1 - \dot{x}^2/c^2}$ and the speed of light c. Hence, we postulate the following variational principle for the relativistic Lagrangian:

$$\delta \int dt \left[-m_0 c^2 \sqrt{1 - \frac{\dot{x}^2}{c^2}} - \beta x^4 + \alpha x^2 - V_g(x)\right] = 0, \tag{5}$$

where the relativistically corrected gravitational potential, V_g, is still to be determined. The Lagrangian, integrand of Eq. (5), does not explicitly depend on time, so that total energy is conserved and serves as one of the integrators of the equation of motion.

According to the procedure first established in Ref. [5] and also explained in Ref. [4], we are able to obtain a integral equation for V_g and solve it analytically, thus yielding

$$\begin{aligned}V_g(x) = \left(1 - e^{-\frac{g}{c^2}x}\right)\left[m_0c^2 - 24\beta\left(\frac{c^2}{g}\right)^4 + 2\alpha\left(\frac{c^2}{g}\right)^2\right]\\ +24\beta x\left(\frac{c^2}{g}\right)^3 - 12\beta x^2\left(\frac{c^2}{g}\right)^2 + 2x\Big[2\beta x^2 - \alpha\Big]\left(\frac{c^2}{g}\right) - \beta x^4 + \alpha x^2.\end{aligned} \tag{6}$$

Note that taking $\alpha = -k/2$ and $\beta = 0$ reproduces the result for the harmonic oscillator calculated in Ref. [4]. This completes all the required data necessary for entirely describing the relativistic motion of the anharmonic oscillator in the gravitational field. The governing equation resulting from Eq. (5) is then provided by the second-order differential equation

$$\frac{m_0\ddot{x}}{\left(1 - \frac{\dot{x}^2}{c^2}\right)^{\frac{3}{2}}} + 4\beta x^3 - 2\alpha x + V_g'(x) = 0, \tag{7}$$

where the exact derivative of Eq. (6) is readily obtained, the analytic integration of Eq. (7), however, is impossible to carry out. Note that the direct integration of the equation of motion for the classical case, Eq. (4), is considerably complicated, producing elliptical functions and beyond. So computing a result in fully symbolic form for the more involved relativistic case, *viz.* Eq. (7), is hopeless.

3 Results

3.1 Numerical integration

For accurate estimates with precise predictions, we require to numerically integrate the differential equation in Eq. (7). For a state-of-the-art approach we employ the Julia programming language with the library `DifferentialEquations.jl`, see Refs. [9, 10].

Observe that for a forthright coding Eq. (7) can be recast into the convenient form

$$A(x)\ddot{x} + B(x)g = 0, \tag{8}$$

where the factors $A(x)$ and $B(x)$ are lengthy expressions containing simple polynomials in the variable x and also the gravitational potential $V_g(x)$ given by Eq. (6). The natural choice for the boundary conditions is $x(0) = 0$ and $\dot{x}(0) = 0$. For the integration using Julia, we adopt the default first-order interpolation algorithm, with at least a stepsize of $\Delta t = 0.005\,\mathrm{s}$ for the time domain, and a relative tolerance of 10^{-12} for high accuracy.

3.2 Simulation of phase space

Figure 2 depicts the phase-space trajectories for the classical case (red curve) and the relativistic case (blue curve) of the anharmonic oscillator within a gravitational field and a double-well potential. For mechanical systems, the phase space represents all physically possible values of the position and velocity variables (or alternatively momentum). For illustration purposes, we have chosen $\alpha = 2.0$ [kg/s^2] and $\beta = 0.5$ [kg/m^{2}s^2] for the shape parameters in the double-well potential, *viz.* Eq. (1). Obviously both oscillatory phenomena are periodic, which is made manifest by the closed trajectories. As is well-known, the phase space of the plain oscillator (not suspended in a gravitational field) acquires a perfectly elliptical form. However, the asymmetric shape of the red curve in Figure 2 is caused by gravity, which introduces a directional preference. Moreover, incorporating relativistic effects, further distorts the pear-shaped classical trajectory into a bullet shaped trajectory—clearly shown in Figure 2 for this extreme case. The pronounced elongation of the blue, relativistic prediction in x-direction is due to the dynamic quality of the relativistic mass—a variable mass instead of the fixed inertial mass in the classical phase-space simulation. A heavier mass certainly will be more affected by gravity. Additionally, the slimmer shape of the blue curve in $\dot{x}$-direction results from the increasing inertia resisting further acceleration with increasing velocity. This effect is purely relativistic and absent in any classical model. The reduced velocity range for the blue curve, taking values for $\dot{x}$ in the interval $[-1, 1]$ instead of $[-3, 3]$, is thus explained by the physical upper speed limit of relativistic dynamics.

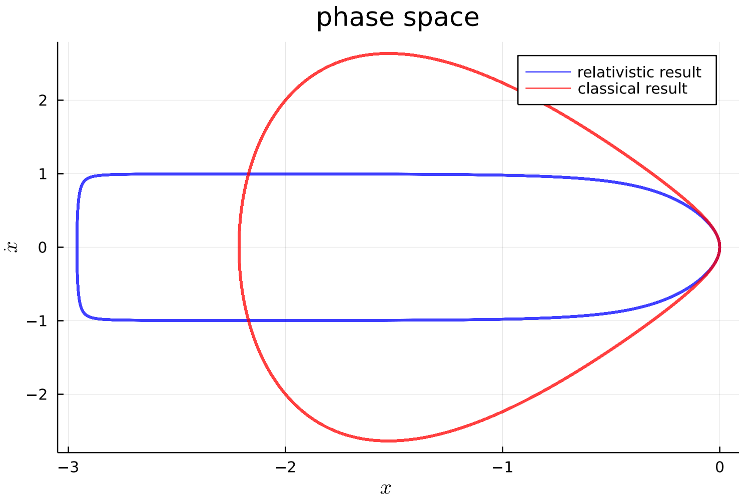


Figure 2: Phase-space diagram of the anharmonic oscillator in a uniform gravitational field, corresponding to the classical and relativistic case for the double-well potential, Eq. (1), with $\alpha = 2.0$ [kg/s^2] and $\beta = 0.5$ [kg/m^{2}s^2].

4 Conclusions

Starting from a Lagrangian approach for a classical anharmonic oscillator in a gravitational field and with a double-well potential, we generalized the discussion to include relativistic effects. By neglecting insignificant tidal effects on a small scale, it was not necessary to employ the fully developed general relativistic framework, but it sufficed to resort to special relativity.

The governing equations of motion for the classical model and its corresponding relativistic extension had to be integrated numerically. We used the Julia programming language as a state-of-the-art tool for obtaining numerical estimates for both models with high accuracy and precision. This allowed us to compute and visualize the phase space for the classical and relativistic case, and then compare both models.

A careful analysis of the phase-space trajectories showed that the differences between the predictions for the classical and the relativistic model are significant. More importantly, these differences are not merely quantitative but show that their origin is of fundamental nature—an immediate consequence of the very different concept of mass in the classical and relativistic approach. This reflects itself in the noticeably distinct shapes of the classical and relativistic trajectories taken in phase space.

In summary, our observations confirm the necessity for including relativistic corrections in our model to produce reliable predictions for the strong-gravity case or when velocities comparatively close to the speed of light are reached.

Acknowledgements

This work has been supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València under Grant PAID-11-21.

References

[1] Bender, C.M. and Wu, T.T., Anharmonic oscillator, *Phys. Rev.*, **184**(5):1231–1260, 1969.

[2] Bender, C.M. and Wu, T.T., Anharmonic oscillator II: a study of perturbation theory in large order, *Phys. Rev. D*, **7**(6):1620–1636, 1973.

[3] Liang, J.-Q. and Müller-Kirsten, H.J.W., Periodic instantons and quantum-mechanical tunneling at high energy *Phys. Rev. D*, **64**(19):4685–4690, 1992.

[4] Tung, M.M., The relativistic harmonic oscillator in a uniform gravitational field, *Mathematics*, **9**(4):1–12, 2021.

[5] Goldstein, H.F., Bender, C.M., Relativistic brachistochrone, *J. Math. Phys.*, **27**(2):507–511, 1986.

[6] Erkal, C., The simple pendulum: a relativistic revisit, *Eur. J. Phys.*, **21**:377–384, 2000.

[7] Torres, P.J., Periodic oscillations of the relativistic pendulum with friction, *Phys. Lett. A*, **372**:6386–6387, 2008.

[8] Lanczos, C., The Variational Principles of Mechanics, Mineola, NY, Dover Publications, 1986.

[9] Bezanson, J., Edelman, A., Karpinski, S., and Shah, V.B., Julia: A fresh approach to numerical computing, *SIAM Review*, **59**(1):65–98, 2017.

[10] Rackauckas, C. and Nie, Q., Differentialequations.jl—a performant and feature-rich ecosystem for solving differential equations in Julia, *J. Open Res. Softw.*, **5**(1):15, 2017.

47

Real-valued preconditioners for complex linear systems arising from the nuclear reactor noise equations

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

(♭) I. U. de Matemática Multidisciplinar,
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain.
(♮) I. U. de Seguridad Industrial Radiofísica y Medioambiental,
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain.

1 Introduction

The evolution of the neutron field and thus the heat field inside a nuclear reactor core can be modelled by the neutron transport equation, approximated by the time dependent neutron diffusion equation. First, the steady state of the reactor is calculated and then the time evolution against any perturbation is simulated. For a small oscillating perturbation that does not change the average state of the reactor, the frequency domain neutron noise diffusion equation can be worked out. This equation resolves is useful to simulate the effect of mechanical vibrations of core internals, periodical anomalies in coolant inlets, flow blockages...

To numerically solve this equation in a nuclear reactor, a large, sparse and complex value linear system must be solved. These types of systems also arise from the discretization of differential equations of interest in wave propagation, electromagnetism, structural dynamics, quantum mechanics, etc. This work is dedicated to efficiently solve the complex-valued linear systems that arise from the frequency-domain neutron noise equation using real-valued formulations.

The neutron flux in a nuclear reactor in the multigroup diffusion approximation is given by

$$\mathcal{V}\frac{\partial \Phi}{\partial t} + \mathcal{L}\Phi = (1-\beta)\mathcal{M}\Phi + \sum_{k=1}^{K} \lambda_k \mathcal{X}_k \mathcal{C}_k, \tag{1}$$

where the concentrations of the neutron precursors is

$$\frac{\partial \mathcal{C}_k}{\partial t} = \beta_k \mathcal{F}\Phi - \lambda_k^d \mathcal{C}_k, \qquad\qquad k = 1, \ldots, K. \tag{2}$$

The matrices for the usual 2 energy group approximation are

$$\Phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \quad \mathcal{F} = \begin{pmatrix} \nu\Sigma_{f1} & \nu\Sigma_{f2} \end{pmatrix}, \quad \mathcal{L} = \begin{pmatrix} -\vec{\nabla}\cdot\left(D_1\vec{\nabla}\right) + \Sigma_{a1} & 0 \\ -\Sigma_{s12} & -\vec{\nabla}\cdot\left(D_2\vec{\nabla}\right) + \Sigma_{a2} \end{pmatrix},$$

[1] anvifer2@upv.es

$$\mathcal{M} = \begin{pmatrix} \chi_1^p \nu\Sigma_{f1} & \chi_1^p \nu\Sigma_{f2} \\ \chi_2^p \nu\Sigma_{f1} & \chi_2^p \nu\Sigma_{f2} \end{pmatrix}, \quad \mathcal{V} = \begin{pmatrix} 1/\mathrm{v}_1 & 0 \\ 0 & 1/\mathrm{v}_2 \end{pmatrix}, \quad \mathcal{X}_k = \begin{pmatrix} \chi_1^{d,k} \\ \chi_2^{d,k} \end{pmatrix}.$$

For a given state, a steady-state configuration of the reactor can be found through the following eigenvalue problem:

$$\mathcal{L}_0\Phi_0 = \frac{1}{\lambda}\mathcal{M}_0\Phi_0, \tag{3}$$

The first-Order neutron noise equation can be found if we split each term $X(\vec{r}, t)$, into its mean value, X_0, which is considered as the steady-state solution; and its fluctuation around the mean value, δX.

$$X(\vec{r},\, t) = X_0(\vec{r}) + \delta X(\vec{r},\, t). \tag{4}$$

The fluctuations are assumed to be small compared to the mean values

$$|\delta X(\vec{r},\, t)| \lll |X_0(\vec{r})|. \tag{5}$$

The different magnitudes are then split as,

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_0 + \delta\mathcal{L}, \\ \mathcal{M} &= \mathcal{M}_0 + \delta\mathcal{M}, \\ \mathcal{F} &= \mathcal{F}_0 + \delta\mathcal{F}., \\ \Phi &= \Phi_0 + \delta\Phi, \\ \mathcal{C}_k &= \mathcal{C}_{k,0} + \delta\mathcal{C}_k \end{aligned}$$

Finally, we perform a Fourier Transform to analyse the problem in the frequency domain. This leads to the frequency-domain neutron noise diffusion equation,

$$\Big(i\omega\mathcal{V} + \mathcal{L}_0 - \gamma\mathcal{M}_0\Big)\delta\Phi = \Big(-\delta\mathcal{L} + \gamma\delta\mathcal{M}\Big)\Phi_0,$$

$$\gamma = (1-\beta) + \sum_{k=1}^{K}\frac{\lambda_k\beta_k}{i\omega + \lambda_k}.$$

In the usual 2 energy groups approximation:

$$\mathcal{C}\delta\Phi = \mathcal{D}\Phi_0 = d,$$

where

$$\mathcal{C} = \begin{bmatrix} \frac{i\omega}{\mathrm{v}_1} - \vec{\nabla}D_1\vec{\nabla} + \Sigma_{a1}^0 + \Sigma_{12}^0 - \gamma\nu\Sigma_{f1}^0 & -\gamma\nu\Sigma_{f2}^0 \\ -\Sigma_{12}^0 & \frac{i\omega}{\mathrm{v}_2} - \vec{\nabla}D_2\vec{\nabla} + \Sigma_{a2} \end{bmatrix},$$
$$\mathcal{D} = \begin{bmatrix} -\delta\Sigma_{a1} - \delta\Sigma_{12} + \gamma\delta\nu\Sigma_{f1} & +\gamma\delta\nu\Sigma_{f2} \\ \delta\Sigma_{12} & -\delta\Sigma_{a2} \end{bmatrix}.$$

To summarize, we need to solve a complex system as:

$$\mathcal{C}\delta\Phi = d,$$

$$\begin{pmatrix} -\vec{\nabla}D_1\vec{\nabla} + \sigma_{R11} + i\sigma_{i11} & \sigma_{R12} + i\sigma_{12} \\ \sigma_{R21} + i\sigma_{21} & -\vec{\nabla}D_2\vec{\nabla} + \sigma_{R22} + i\sigma_{22} \end{pmatrix}\begin{pmatrix} \delta\phi_1 \\ \delta\phi_2 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix},$$

We discretize the equation using the continuous Galerkin Finite Element Method.

To sum up, we must study preconditioners to the problem:

$$\left(-\vec{\nabla}D_1\vec{\nabla} + \sigma_R + i\sigma_i\right)\delta\Phi = d$$

2 Real equivalent formulations

If we consider the complex linear system

$$Cz = d, \tag{6}$$

where $C = A + iB$, $z = x + iy$ and $d = b + ic$. The most usual real formulation is:

$$\begin{pmatrix} A & -B \\ B & A \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}, \tag{7}$$

However, the choice of the real and imaginary part of the coefficient matrix is *largely* arbitrary and other equivalent real formulation exist:

$$\begin{pmatrix} A & B \\ B & -A \end{pmatrix} \begin{pmatrix} x \\ -y \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}, \tag{8}$$

$$\begin{pmatrix} B & A \\ -A & B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ -b \end{pmatrix}, \tag{9}$$

$$\begin{pmatrix} B & A \\ A & -B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ b \end{pmatrix}. \tag{10}$$

The performance of direct linear solvers applied to the real formulations degrades in comparison to the original complex case. Also, the effectiveness of preconditioners related to sparse direct solvers, like ILU, deteriorates similarly [1].

Then, we centre our efforts in equation (7). It must be noted that the elements of A and B must not be not stored twice. We also look for preconditioners that do not double storage requirements. Indeed matrix-free methodology is applied to non-diagonal blocks to minimize the memory requirements.

3 Block Preconditioners for real formulations

As in the neutron noise problem, A is dominant, the simplest preconditioner is to use a block Jacobi preconditioner:

$$P_{BJ} = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}. \tag{11}$$

Another, usual choice will be the the Block Gauss-Seidel preconditioner.

$$P_{GS} = \begin{pmatrix} A & B \\ 0 & A \end{pmatrix}.$$

This preconditioner is effective when $A^{-1}B \approx 0$ as it is the case in neutron noise simulations.

Another possible choice of a preconditioner is the shifted skew symmetric and the Hermitian/skew-Hermitian splitting (HSS) preconditioner,

$$P_{HSS} = \begin{pmatrix} A + \alpha I & 0 \\ 0 & A + \alpha I \end{pmatrix} \begin{pmatrix} \alpha I & -B \\ B & \alpha I \end{pmatrix}$$

where $\alpha > 0$.

$$P_{\alpha} = \begin{pmatrix} \alpha I & A \\ -A & \alpha I \end{pmatrix}, \qquad P_{HSS} = \begin{pmatrix} B + \alpha I & 0 \\ 0 & B + \alpha I \end{pmatrix} \begin{pmatrix} \alpha I & A \\ -A & \alpha I \end{pmatrix}.$$

Both preconditioners require the solution of systems $(A^2 + \alpha^2 I)u = r$.

Table 1: Numerical results with FE = 1.

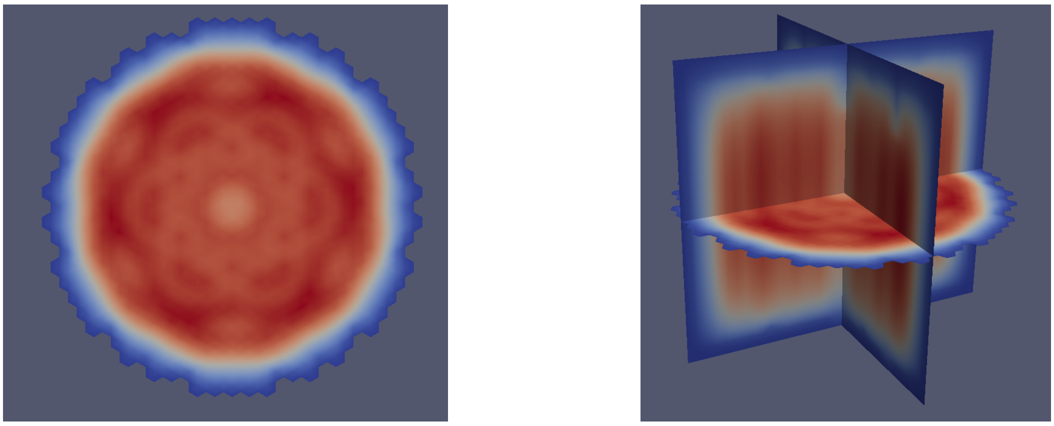
Algebra	Precond	N° of its	CPU Time(s)	Memory (Mb)
Complex system	ILU(0)	110 its	7.0 s	500
Complex system	Jacobi	798 its	20.6 s	500
Complex system	None	1079 its	26.89 s	500
Real system	ILU(0)	193 its	11.4 s	1000
Real system	Jacobi	798 its	24.6 s	500
Real system	None	1079 its	32.7 s	500

Table 2: Results with FE = 1

4 Numerical Results

To study the different practitioners proposed, we use a hexagonal reactor, the VVER-440. Figure 1 use the . First, linear polynomials in the Finite Element Method are studies. In this case, each energy-group block of the real formulation has a size of 65.611 × 65.611. Thus, the total size of the matrix is 262.444 × 262.444.

If block structures are not used, complex algebra is faster to equivalent real formulations, as Table 1 shows. However, we use the real value and imaginary value separation, that solves a 2x2 block structure. Here we can use matrix-free methodology for the non diagonal blocks that greatly reduces the memory demands. The results are displayed in Tables 3. The block Gauss Seidel preconditioners shows the best results using the real equivalent formulation in terms of CPU time and memory. Figure 2 shows the convergence graph of the different practitioners studied with $FE = 1$.



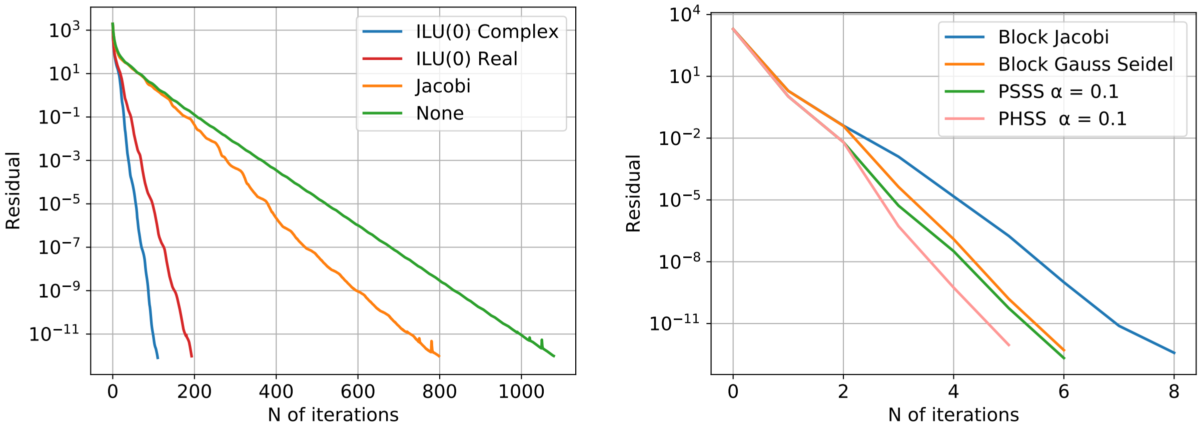
(a) Radial Plane

(b) Axial View

Figure 1: Solution of the VVER-440 reactor.

Table 3: Numerical results with FE = 1.

Algebra	Precond	N° of its	CPU (s)	Memory (Mb)
Real system	Block Jacobi	8 its	29.4 s	135
Real system	Blok Gauss Seidel	6 its	15.2 s	135
Real system	PSS	6 its	20.2 s	135
Real system	PHSS	5 its	18.4 s	135



(a) Complete matrix

(b) Block Preconditioning

Figure 2: Convergence graph of the different practitioners studied with FE = 1.

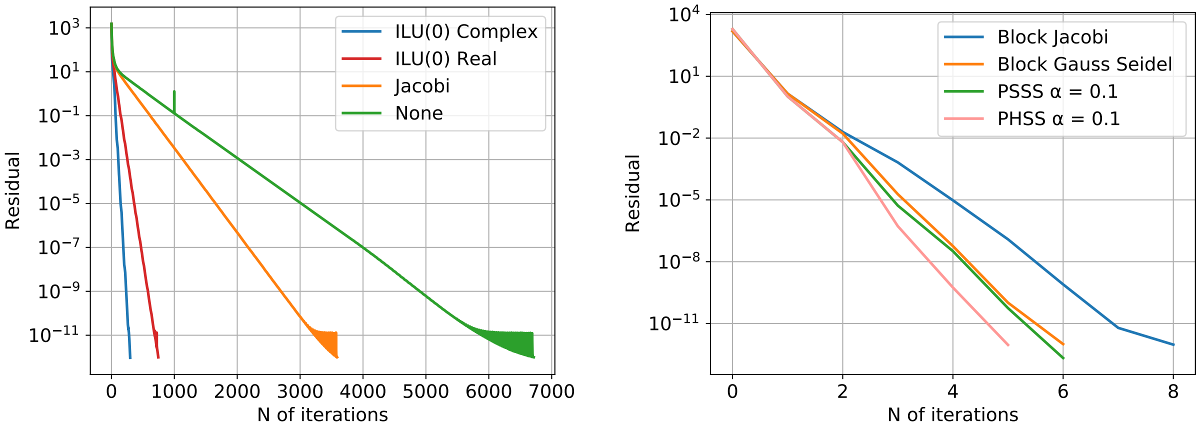
Now, we study the same problem using second degree polynomials in finite element method. In real formulations, each block has a size of 504.691.691 and the total size of the matrix is 2.018.764 × 2.018.764. The results are displayed in Tables 4 and 5. Again, if we do not use real equivalent formulations, the complex system is faster, but it uses more memory to store the matrix. The block Gauss Seidel preconditioners shows the best results using the real equivalent formulation in terms of CPU time and memory. Figure 3 shows the convergence graph of the different practitioners studied with FE = 2.

5 Conclusions

Sparse complex linear systems arise from the discretization of the neutron noise diffusion equations. In this work, real equivalent formulations are a practical way to solve them using libraries for

Table 4: Numerical results with FE = 2.

Algebra	Precond	N° of its	CPU (s)	Memory (Mb)
Complex system	ILU(0)	298 its	250 s	4000
Complex system	Jacobi	3588 its	1245 s	4000
Complex system	None	6715 its	2300 s	4000
Complex system	ILU(0)	745 its	786 s	8000
Complex system	Jacobi	3588 its	1618 s	4000
Complex system	None	6715 its	2990 s	4000



(a) Complete preconditioning

(b) Block Preconditioning

Figure 3: Convergence graph of the different practitioners studied with FE = 2.

Table 5: Numerical results with FE = 2.

Algebra	Precond	N° of its	CPU (s)	Memory (Mb)
Real system	Block Jacobi	8 its	372 s	1065
Real system	Blok Gauss Seidel	6 its	275 s	1065
Real system	PSS	6 its	455 s	1065
Real system	PHSS	5 its	375 s	1065

real-valued matrices. Preconditioners based on matrix factorizations, as ILU, deteriorate in real equivalent formulations.

However, block preconditioning permit to use matrix-free techniques and parallelism. For the neutron noise problem, Block Gauss-Seidel preconditioner show the best agreement between memory usage and CPU time.

Future works will be centred in multigrid preconditioner that have shown good results in similar problems.

References

[1] Benzi M. and D. Bertaccini, Block preconditioning of real-valued iterative algorithms for complex linear systems,*IMA Journal of Numerical Analysis*, 28(3), 598–618, July 2008.

[2] Carreño, A., A. Vidal-Ferràndiz, D. Ginestar, and G. Verdú, Frequency-domain models in the SPN approximation for neutron noise calculations. *Progress in Nuclear Energy*, 148 (14233), 1–11, 2022. DOI: 10.1016/j.pnucene.2022.104233

[3] Vidal-Ferràndiz, A., A. Carreño, D. Ginestar and G. Verdú, Edge-wise perturbations to model vibrating fuel assemblies in the frequency-domain using FEMFFUSION: development and verification. *Annals of Nuclear Energy* (175). DOI: 10.1016/j.anucene.2022.109246

Mathematical model for heat transfer and stabilization of LED lamps for measurements in a laboratory

Carlos Velásquez[a,b,c,1] M. Ángeles Castro[a], Francisco Rodríguez[a] and Francisco Espín[b,d]

(a) Dept. of Applied Mathematics,
University of Alicante,
Apdo. 99, 03080, Alicante, Spain
(b) Instituto de Investigación Geológico y Energético,
Av. de la República E7-263 y Diego de Almagro, Quito, Ecuador
(c) Universidad Central del Ecuador, Modalidad en Línea
Quito, Ecuador
(d) Departamento de Luminotecnia, Luz y Visión, Universidad Nacional de Tucumán
Tucumán, Argentina

1 Introduction

The use of LED lamps as an energy-saving policy is a practice that occurs in several countries [4,7]. However, in emerging countries the metrological control is not deepened. There are standardized methods for measuring the properties of LED luminaires such as ANSI/IES LM 79-19 Approved Method: Optical and Electrical Measurements of Solid-State Lighting Products [5], or CIE S 025/E:2015 Test Method for LED Lamps, LED Luminaires, and LED Modules [6].

These methods suggest, among others, the use of an integrating sphere plus a spectroradiometer as the main elements for measuring the properties of the lamp, as well as the power supply and electrical meter to measure the electrical magnitudes [8]. Figure 1 shows a typical scheme for measuring a LED lamp.

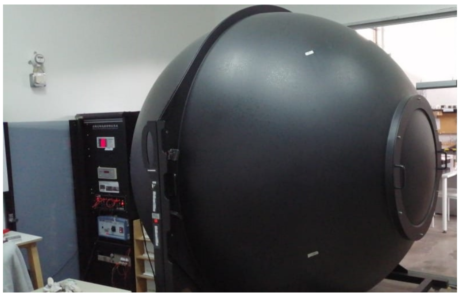


Figure 1: Integrating sphere for luminous flux detection according to IES LM 79

These methods are demanding in their conditions. In particular, the ambient temperature condition for LM 79 is $25°C \pm 1.2°C$ [5]. The device under test (DUT), in general, must be

[1] cavf1@alu.ua.es

stabilized for a time that is usually close to 2 hours so that its photometric magnitudes are stable. In this process, there is a phenomenon of heat transmission from the lamp to the environment.

An ISO/IEC 17025 [4] accredited laboratory has the necessary equipment to meet this condition. However, the problem is in taking data from a significant number of test items. A given DUT can use the complete experimental scheme for just over 2 hours (its stabilization and data collection), and this would imply an approximate maximum of 7 test items per day.

On the other hand, an efficiency policy consistent with the quality of imported items requires market surveillance, which depending on the volume of brands in the market could reach a need of up to 40 tests per day [2].

This is a problem that can be solved technically. If the longest time (stabilization) can be done outside the integrating sphere, the experimental scheme would be used only for data collection, allowing the tests required for market surveillance to be carried out.

The stabilization outside the integrating sphere would represent 40 lamps turned on simultaneously that would act as heat sources in a stabilized environment [1]. For the construction of an isolated stabilization bench that does not disturb the thermal stabilization of the environment, it is necessary to find an adequate distance between the lamps.

Some studies indicate that, although an LED lamp has many components, its thermal behavior and heat transmission to the environment can be studied through a suitable simulation [9, 10].

2 Methods

The construction of the stabilization bench starts from the resolution of the equation to find the heat transfer from a heat source to the environment,

$$\rho c_p \frac{\partial T}{\partial t} + \rho c_p u \nabla T = \nabla(k\nabla T) + Q + Q_{vh} + W_p, \tag{1}$$

where ρ is density, c_p is the specific heat capacity at constant pressure, T is temperature, t, time, Q_{vh} is the viscous dissipation, W_p is the thermodynamic work, u is the velocity vector, $k\nabla T$ is the heat flux by conduction, and Q is the heat source, which is computed as

$$Q = \frac{P_{tot}}{V}, \tag{2}$$

where P_{tot} is the power of the lamp and V is the volume. Additionally, the contour of the bank must be isolated, so that the insulation will be modeled by

$$-n(k\nabla T) = 0, \tag{3}$$

where n is the normal vector. In this model, convection is considered to be negligible.

3 Results

LED chips were modeled as solid bulk aluminum plates surrounded by air for the development of the simulation. 40 lamps of 10W were taken into account. First, the 2D system was solved by locating the lamps in different distributions, finding that the location in matrix form is the most appropriate, with a separation of 30cm between lamps (Figure 2).

Heat transfer was calculated for a period of 2 hours, that is, 7200 seconds, and it was supervised that the temperature distribution between the lamps did not exceed $30°C$. Then a 3D simulation was carried out, varying the number of planes in which to place the lamps. However, the specifications of the laboratory and the space available for the construction of the stabilization bench means that the bench should not be too wide, although it can have large dimensions both in length and height.

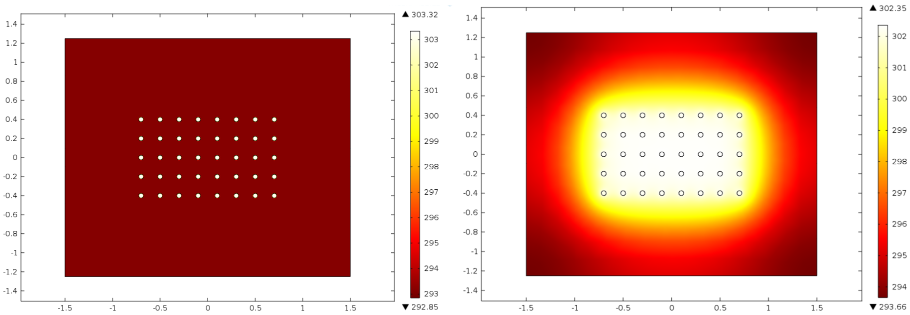


Figure 2: On the left is the simulation at time $t = 0s$ and on the right the simulation at time $t = 7200s$.

The final diagram used that meets the real spatial requirements as well as the thermal conditions can be seen in the Figure 3.

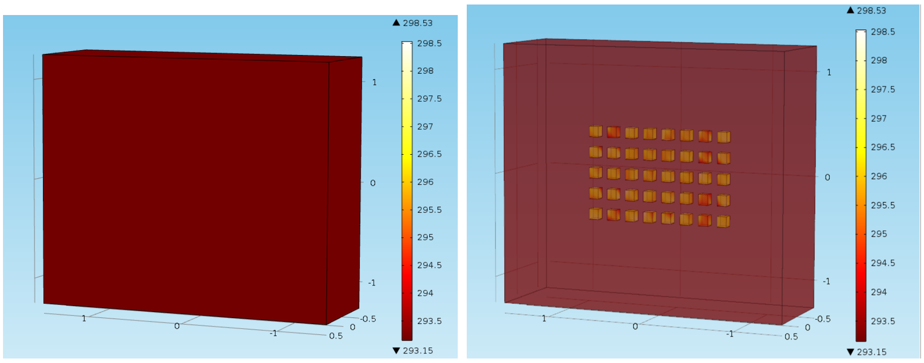


Figure 3: Simulation at time $t = 0s$.

Again, the behavior of the heat sources was studied in a time of 7200 seconds as shown in Figure 4.

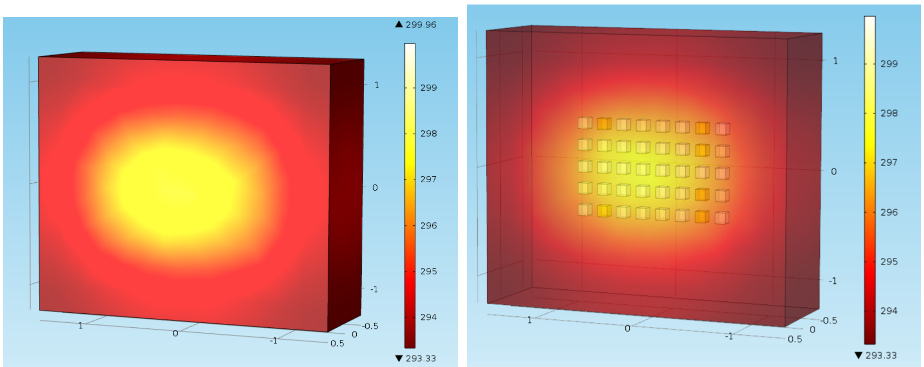


Figure 4: Simulation at time $t = 7200s$.

Based on these results, we proceeded to design the structure of the stabilization bank (Figure 5), considering both the results obtained in the heat transfer study and also the space requirements of the laboratory.

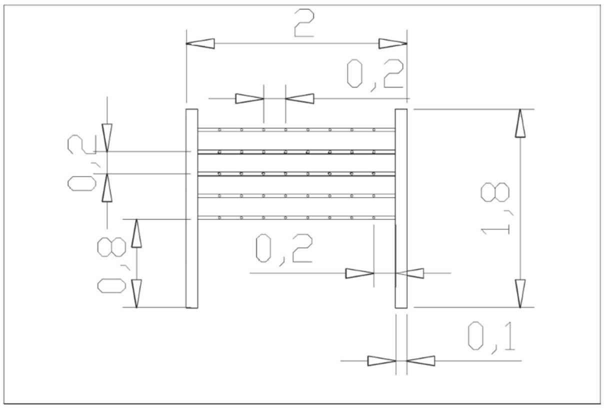


Figure 5: Stabilization bench construction diagram

Thermal insulation also performs an important function. The level of lighting of 40 lamps turned on simultaneously inside the laboratory was studied. The result was that at the operator's distance from the integrating sphere he/she would receive more than 4000 lx (Figure 6). For this reason, isolation not only prevents the thermal disturbance of the controlled environment, but also ensures that visible radiation does not affect the performance of people inside the laboratory.

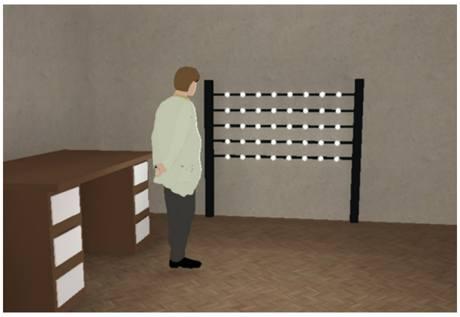


Figure 6: Lighting of 40 lamps without a physical light blockage

4 Conclusions

Market surveillance requires a significant capacity of testing laboratories. Generally, the test of a DUT according to LM 79 requires a stabilization prior to the test, which represents a loss of effective testing time. It is possible to stabilize the LED lamps outside the integrating sphere to perform many more measurements, meet the measurement standard and perform proper market surveillance. Modeling heat transfer for an array of LED lamps resulted in the selection of a matrix of positions in which the lamps must be located 30cm apart, considering thermal insulation at the edges.

Future work includes analysing the behaviour of lamps that are not necessarily for domestic use and understand the variations in distance required in higher powers. The electrical component is also important, designing the electrical diagram for the synchronization of power supplies to allow entering the lamp into the integrating sphere without going through OFF. Finally, the bank must be practically built and its results contrasted with the simulations.

References

[1] Bhagatsingh Amarnath Biradar, Subhasisa Rath, and Sukanta Kumar Dash. Orientation effects on conjugate natural convection heat transfer from an led bulb: A numerical study. *International Journal of Thermal Sciences*, 159:106640, 2021.

[2] European Commission. DocsRoom - European Commission Report on the 12th joint crossborder EMC market surveillance campaign on LED lighting equipment - Revision 1 (2019), 2019.

[3] International Organization for Standardization. Requisitos generales para la competencia de los laboratorios de ensayo y calibración ISO/IEC 17025:2017, 2017.

[4] Bruno Gayral. Leds for lighting: Basic physics and prospects for energy savings. *Comptes Rendus Physique*, 18(7):453–461, 2017. Demain l'énergie.

[5] Illuminating Engineering Society of North America. Approved method: Optical and electrical measurements of solid state lighting products. ANSI/IES lm-79-19, 2019.

[6] International Commission on Illumination. Test method for led lamps, led luminaires, and led modules. CIE S 025/e:2015, 2015.

[7] David Borge-Diez Santiago Pulla Galindo and Daniel Icaza. Energy sector in ecuador for public lighting: Current status. *Energy Policy*, 160:112684, 2022.

[8] Carlos Velásquez, M. Angeles Castro, Francisco Rodríguez, Francisco Espin, and Nathaly Falconi. Optimization of the calibration interval of a luminous flux measurement system in hid and ssl lamps using a gray model approximation. In *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*, pages 1–7, 2021.

[9] Hao Wengang, Wei Lulu, Zhang Zongmin, Lai Yanhua, and Lv Mingxin. Research on simulation and experimental of thermal performance of led array heat sink. Procedia Engineering, 205:2084–2091, 2017. *10th International Symposium on Heating, Ventilation and Air Conditioning*, ISHVAC2017, 19-22 October 2017, Jinan, China.

[10] Kai-Shing Yang, Chi-Hung Chung, Ming-Tsang Lee, Song-Bor Chiang, Cheng-Chou Wong, and Chi-Chuan Wang. An experimental study on the heat dissipation of led lighting module using metal/carbon foam. *International Communications in Heat and Mass Transfer*, 48:73–79, 2013.

49

A Seventh Order Family of Jarratt Type Iterative Methods for Solving Systems of Nonlinear Equations and Applications

Saima Yaseen$^{\flat,1}$ and Fiza Zafar$^{\flat}$

($\flat$) Centre for Advanced Studies in Pure and Applied Mathematics,
Bahauddin Zakariya University,
Multan 60800, Pakistan.

1 Introduction

Nonlinear phenomena occur in various fields of science, business, and engineering. Research in the area of computational science is constantly growing, with the development of new numerical schemes or with the modification of existing ones. One of the major issues of numerical analysis is the construction of efficient iterative methods for solving nonlinear equations and systems of equations. Applicability to several real-life problems is the reason behind its importance that can be seen in the available literature. Moré [9] presented several nonlinear model problems and majority of them are stated in the form of $G(X) = 0$. Also Grosan and Abraham [5] analyzed the relevancy of the nonlinear systems in neurophysiology, kinematics syntheses, chemical equilibrium, combustion and economics modeling problems. Furthermore, Lin et al. [8] also expressed the applications of the nonlinear system in transportation theory. Usually, nonlinear phenomena in physical and medical sciences are modelled as nonlinear equations. So, researchers many are actively engaged to give effective and widely applicable iterative methods. Recently, researchers have proposed sixth-order iterative methods using weight functions and parameter approaches (we can see in [2, 3, 6, 7]). Abad et al. [1] proposed an efficient Steffensen family of three-step iterative schemes with seventh-order convergence using four function evaluations. The proposed methods are obtained by using the weight functions procedure. However, such numerical schemes, objectively need to be computationally inexpensive with higher order of convergence. Taking into account these demanding features, this article attempts to develop a new seventh-order scheme based on weight functions using four function evaluations. Some applications are also given to check the numerical performance of the scheme.

2 Development of Seventh Order Method

We use the weight function approach in the development of our scheme. For the multivariate vector valued function $G : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^n$, we have

$$v^{(n)} \;=\; r^{(n)} - \left(G'\left(r^{(n)}\right)\right)^{-1} G\left(r^{(n)}\right),$$

[1] saimayasin08@gmail.com

$$\begin{aligned} z^{(n)} &= v^{(n)} - P\left(s^{(n)}\right).\left(G^{'}\left(r^{(n)}\right)\right)^{-1} G\left(v^{(n)}\right), \\ r^{(n+1)} &= z^{(n)} - \left(R\left(s^{(n)}\right) + Q\left(t^{(n)}\right)\right).\left(G^{'}\left(r^{(n)}\right)\right)^{-1} G\left(z^{(n)}\right). \end{aligned} \tag{1}$$

$$s^{(n)} = I - \left(G^{'}\left(r^{(n)}\right)\right)^{-1} G[r^{(n)}; v^{(n)}].$$

and

$$t^{(n)} = I - \left(G^{'}\left(r^{(n)}\right)\right)^{-1} G[v^{(n)}; z^{(n)}] P\left(s^{(n)}\right).$$

where $P : A_{n\times n}(\mathbb{R}) \to \Gamma(\mathbb{R}^n), Q : B_{n\times n}(\mathbb{R}) \to \Gamma(\mathbb{R}^n)$ $R : C_{n\times n}(\mathbb{R}) \to \Gamma(\mathbb{R}^n)$ *with* $A_{n\times n}$, $B_{n\times n}$ and $C_{n\times n}$ *be the set of* $n \times n$ *matrices and* $\Gamma(\mathbb{R}^n)$ *be the set of linear operators from* $\mathbb{R}^n$ *to* $\mathbb{R}^n$. We define the seventh-order scheme for nonlinear system of equations by using some weight functions that are described from the following theorem.

Theorem 22. *Let us suppose that* $G : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^n$ *be a sufficiently differentiable function in a closed neighborhood* $\mathbb{D}$ *. We consider that* $G'(X)$ *is continuous and nonsingular at* Υ *. In addition, that convergence is guaranteed if we consider that initial guess* $X^{(0)}$ *is close to root* Υ*. Then, the numerical scheme* (1) *has seventh order convergence under the following conditions:*

$$\begin{aligned} P(0) &= I, P'(0) = 2I, P''(0) = 0, \|P'''(0)\| < \infty, \\ R(0) &= I, R'(0) = 2, R''(0) = 2, \|R'''(0)\| < \infty, \\ Q(0) &= 0, Q'(0) = I, \|Q''(0)\| < \infty. \end{aligned}$$

Some particular cases of the proposed scheme are given as :
FS1: If we take the weight functions $P(s^{(n)})$, $R(s^{(n)})$ and $Q(t^{(n)})$ of the following form:

$$P(s^{(n)}) = (a_0 I + a_1 s^{(n)} + a_2 (s^{(n)})^2)^{-1},$$

$$R(s^{(n)}) = (b_0 I + b_1 s^{(n)} + b_2 (s^{(n)})^2)^{-1}$$

$$Q(t^{(n)}) = c_0 I + c_1 t^{(n)} + c_2 (t^{(n)})^2,$$

and

$$a_0 = 1, a_1 = -2, a_2 = 4,$$

$$b_0 = -1, b_1 = -2, b_2 = -5,$$

$$c_0 = 2, c_1 = 1, c_2 = c_2$$

We take $c_2 = 9$ in particular.
FS2: If the weight functions $P(s^{(n)})$, $R(s^{(n)})$ and $Q(t^{(n)})$ are taken of the following form:

$$P(s^{(n)}) = (a_0 + a_1 s^{(n)} + a_2 (s^{(n)})^2)^{-1},$$

$$R(s^{(n)}) = b_0 I + b_1 s^{(n)} + b_2 (s^{(n)})^2$$

$$Q(t^{(n)}) = c_0 I + c_1 t^{(n)} + c_2 (t^{(n)})^2,$$

with

$$a_0 = 1, a_1 = -2, a_2 = 4,$$

$$b_0 = -1, b_1 = 2, b_2 = 1,$$

$$c_0 = 2, c_1 = 1, c_2 = c_2.$$

In particular, we take $c_2 = 15$.
FS3:Let us take the weight functions$P(s^{(n)})$, $R(s^{(n)})$ and $Q(t^{(n)})$ of the following form:

$$P(s^{(n)}) = a_0 + a_1 s^{(n)} + a_2 (s^{(n)})^2,$$

$$R(s^{(n)}) = (b_0 + b_1 s^{(n)} + b_2 (s^{(n)})^2)^{-1}$$

and

$$Q(t^{(n)}) = c_0 I + c_1 t^{(n)} + c_2 (t^{(n)})^2,$$

such that

$$a_0 = 1, a_1 = 2, a_2 = 0,$$

$$b_0 = -1, b_1 = -2, b_2 = -5,$$

$$c_0 = 2, c_1 = 1, c_2 = c_2.$$

Particularly, we choose $c_2 = 10$.

3 Numerical Results

We want to verify the numerical results of our iterative method so, compare the results of our schemes namely $FS1$, $FS2$, and $FS3$ with respect to number of iterations n, absolute residual error of the corresponding function $\left\|G(r^{(n)})\right\|$ and absolute error in two consecutive iterations $\left\|r^{(n)} - r^{(n-1)}\right\|$. Our method is compared with seventh-order method presented by Abad et al. [1] that is given below

$$\begin{aligned} v^{(n)} &= r^{(n)} - (G'(r^{(n)}))^{-1} G(r^{(n)}), \\ z^{(n)} &= v^{(n)} - K(s^{(n)})[r^{(n)}, v^{(n)}; G]^{-1} G(v^{(n)}), \\ r^{(n+1)} &= z^{(n)} - H(s^{(n)})[v^{(n)}, z^{(n)}; G]^{-1} G(z^{(n)}), \end{aligned} \tag{2}$$

where

$$K(s^{(n)}) = I + s^{(n)},$$
$$H(s^{(n)}) = I + (s^{(n)})^2.$$

Example 1. *Consider a problem [10] in which we have to determine the steady state concentration of naphthalene in the lung and liver. In steady-state process, concentration of components remains the same i.e. no changes takes place with time. Hence, no material accumulates anywhere in the system. In engineering, a material balance is conducted over each compartment to account for every material that is entering and leaving a system. From this material balancing process, we obtain a system of equations to find unknowns such as concentrations of some specific species from each compartment. Now, consider a material balancing of naphthalene over two chambers such as lung and liver. Therefore, a* 2×2 *system of equations is obtained,* $G_4(X) = 0$*, and is described as:*

$$780(1 - R) + R(0.5C_{liver} + 1.5C_{lung}) - (\frac{8.75C_{lung}}{2.1 + C_{lung}})0.08 - 2C_{lung} = 0,$$

$$0.5C_{lung} - (\frac{118C_{liver}}{7 + C_{liver}})0.322 - 0.5C_{liver} = 0. \tag{3}$$

For given value $R = 0.6$*, the above equations can be written as:*

$$0.3C_{liver} - 1.1C_{lung} - \frac{0.7C_{lung}}{2.1 + C_{lung}} + 312 \quad = \quad 0,$$

Table 1: Comparison of methods for Example 1.

$Cases$	m	$\parallel r^{(n)} - r^{(n-1)} \parallel$	$\parallel G\left(r^{(n)}\right) \parallel$
$FS1$	1	79.31126	$1.03007e(-11)$
	2	$2.81033e(-11)$	$1.07399e(-91)$
	3	$9.75868e(-92)$	$1.36080e(-576)$
$FS2$	1	79.31126	$4.56475e(-12)$
	2	$1.24447e(-11)$	$4.47472e(-94)$
	3	$4.06887e(-94)$	$7.11030e(-591)$
$FS3$	1	79.31126	$9.10306e(-12)$
	2	$2.48351e(-11)$	$4.88632e(-92)$
	3	$4.43997e(-92)$	$1.2066e(-578)$
AC	1	79.31126	$9.99659e(-12)$
	2	$2.73194e(-11)$	$4.20853(-91)$
	3	$3.81759e(-91)$	$5.24784e(-573)$

$$0.5C_{lung} - (\frac{37.996C_{liver}}{7 + C_{liver}}) - 0.5C_{liver} \quad = \quad 0. \tag{4}$$

We will solve this system for the two unknowns i.e. C_{lung} *and* C_{liver}*. The root of this system is:*

$$(361.3112614, 287.1278085)^t.$$

The initial guess is taken as: $(C^{(0)}_{lung}, C^{(0)}_{liver}) = (358, 282)^t$.

Example 2. *We consider Combustion Problem taken from [5] for a temperature of 3000* ^0C*. The problem is described by the following sparse system of equations:*

$$\begin{aligned}
x_2 + 2x_6 + x_9 + 2x_{10} &= 10^{-5} \\
x_3 + x_8 &= 3 \cdot 10^{-5} \\
x_1 + x_3 + 2x_5 + 2x_8 + x_9 + x_{10} &= 5 \cdot 10^{-5} \\
x_4 + 2x_7 &= 10^{-5} \\
0.5140437 \cdot 10^{-7} x_5 &= x_1^2 \\
0.1006932 \cdot 10^{-6} x_6 &= 2x_2^2 \\
0.7816278 \cdot 10^{-15} x_7 &= x_4^2 \\
0.1496236 \cdot 10^{-6} x_8 &= x_1 x_3 \\
0.6194411 \cdot 10^{-7} x_9 &= x_1 x_2 \\
0.2089296 \cdot 10^{-14} x_{10} &= x_1 x_2^2.
\end{aligned}$$

and we take initial guess as (0.000018,0.000012,0.0013,0.0000000062,0.00006,0.00003, 0.00049,0.0016,0.000038,0.000038)^t.

Example 3. *We take nonlinear system of three equations, taken from [4], is:*

$$\begin{aligned}
3x_1 - \cos(x_2 x_3) - \frac{1}{2} &= 0, \\
x_1^2 - 81(x_2 + 0.1)^2 + \sin(x_3) + 1.06 &= 0, \\
e^{-x_1 x_2} + 20x_3 + (10\frac{\pi}{3} - 1) &= 0,
\end{aligned}$$

where we take initial guess as (0.6,1,-0.6)^t.

Table 2: Comparison of methods for Example 2.

$Cases$	m	$\parallel r^{(n)} - r^{(n-1)} \parallel$	$\parallel G\left(r^{(n)}\right) \parallel$
$FS1$	1	$1.61533e(-3)$	$3.15559e(-10)$
	2	$1.02184e(-3)$	$1.15147e(-11)$
	3	$1.06685e(-4)$	$3.60548e(-13)$
$FS2$	1	$1.57932e(-3)$	$4.45799e(-11)$
	2	$5.05179e(-4)$	$1.40737e(-12)$
	3	$1.53520e(-5)$	$1.28762e(-13)$
$FS3$	1	$1.51060e(-3)$	$4.33644e(-10)$
	2	$1.25310e(-3)$	$1.23174e(-12)$
	3	$1.87269e(-4)$	$9.49965e(-13)$
AC	1	$2.47379e(-3)$	$6.51052e(-10)$
	2	$2.14719e(-3)$	$1.94073e(-11)$
	3	$2.589545(-4)$	$3.57629(-13)$

Table 3: Comparison of methods for Example 3.

$Cases$	m	$\parallel r^{(n)} - r^{(n-1)} \parallel$	$\parallel G\left(r^{(n)}\right) \parallel$
$FS1$	1	$9.38605e(-1)$	1.30180
	2	$6.15009e(-2)$	$1.72088e(-3)$
	3	$1.06482e(-4)$	$4.02249e(-21)$
$FS2$	1	1.01300	$1.90013e(-1)$
	2	$1.30033e(-2)$	$6.10908e(-6)$
	3	$3.77576e(-7)$	$1.01367e(-38)$
$FS3$	1	$9.18735e(-1)$	1.85315
	2	$8.13347e(-2)$	$1.13912e(-3)$
	3	$7.04488e(-5)$	$3.10923e(-22)$
AC	1	$8.88511e(-1)$	2.80943
	2	$1.10954e(-1)$	$8.65758e(-3)$
	3	$5.34043e(-4)$	$3.68782e(-17)$

4 Conclusion

A new seventh-order convergent scheme with four function evaluations per iteration is presented to solve nonlinear models in the multivariate scenario. The proposed method is evaluated to be adequate in terms of accuracy and reliability. This was further supported when we compare the numerical results of non linear problems to our scheme with these of existing families of Steffensen type methods [1].

References

[1] Abad, M., Codero, A., Torregrosa, J.R., A family of seventh order schemes for solving non linear systems, Bull. Math. Soc. Sci. Math. Roumanie Tome, 57(105): 133-145, 2014.

[2] Behl, R., Sarria, I., Gonzalez, R., Magrenan, A.A., Highly efficient family of iterative methods for solving nonlinear models, J. Comput. Appl. Math. 346: 110-132, 2019.

[3] Behl, R., Argyros, I.K., A new higher order iterative scheme for the solutions of nonlinear systems, Mathematics 8(2): Art. 271, 2020

[4] Burden, R.L., Faires, J.D., Numerical Analysis, PWS Publishing Company, Bostan, MA, USA, 2001.

[5] Grosan, C., Abraham, A., A new approach for solving nonlinear equations systems, IEEE Trans. Syst. Man Cybernet Part A: Syst. Humans 38: 698-714, 2008.

[6] 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, Numer. Algorithms 87: 1017-1060, 2021.

[7] Lee, Kim, Y.I., Development of a family of Jarratt-like sixth-order iterative methods for solving nonlinear systems with their basins of attraction, Algorithms, 13(11): Art. 303, 2020.

[8] Lin, Y., Bao, L., Jia, X., Convergence analysis of a variant of the newton method for solving nonlinear equations, Comput. Math. Appl. 59: 2121-2127, 2020.

[9] Moré, J.J., A collection of nonlinear model problems, in: E.L. Allgower, K. Georg (Eds.), Computational Solution of Nonlinear Systems of Equations, in: Lectures in Applied Mathematics, Amer. Math. Soc. Providence, RI 26: 723-762, 1990.

[10] Michael R. King, Nipa A. Mody, Numerical and Statistical Methods for Bioengineering: Applications in Matlab, Cambridge University Press, New York, 2010.

A Seventh Order Jarratt type Iterative Method for Solving Systems of Nonlinear Equations and Applications

Fiza Zafar[♭,1] Alicia Cordero[♮], Husna Maryam[♭] and Juan R. Torregrosa[♮]

(♭) Centre for Advanced Studies in Pure and Applied Mathematics,
Bahauddin Zakariya University
Multan 60800, Pakistan.
(♮) I. U. de Matemática Multidisciplinar,
Universitat Politècnica de València,
Camí de Vera s/n, València 46022, Spain.

1 Introduction

Numerical analysis refers to the process in which we develop, analyze, and investigates numerous methods and algorithms for numerically solving the problems in the diverse domains such as electrical engineering, chemical engineering, mathematics, physics, and applied sciences. The most significant and difficult task is to find an efficient and accurate solutions to systems of non-linear equations $G(X) = 0$ with $G : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^n$. For this purpose, many higher order iterative methods are developed that are efficient and have less computational cost such that Ostrowski [8], Traub [9]. The simplest and most powerful root finding method to solve non-linear systems is Newton-Raphson method, defined as:

$$X^{(n+1)} = X^{(n)} - G'(X^{(n)})^{-1}G(X^{(n)}),$$

where $G'(X^{(n)})$ is the Jacobian matrix. Many researchers work to improve the efficiency of iterative methods and developed some sixth-order derivative based methods [3,4,6], a seventh-order method [1] that was not Jarratt type, and a non-optimal seventh-order Jarratt type method [2].

To achieve higher convergence order, we develop three-step seventh-order Jarratt type method for solving systems of non-linear equations. The suggested multidimensional method requires only three function evaluations and one Jacobian at each iteration. It also involves the divided differences in weight functions, defined as:

$$\begin{aligned}[y,x;G]_{ji} = (&G_j[y_1, y_2, ..., y_{i-1}, y_i, x_{i+1}, ..., x_n]- \\ &G_j[y_1, y_2, ..., y_{i-1}, x_i, x_{i+1}, ..., x_n])/(y_i - x_i), \\ &1 \le j, i \le n,\end{aligned}$$

[1] fizazafar@bzu.edu.pk

This research was partially supported by Ministerio de Ciencia, Innovación y Universidades, Spain PGC2018-095896-B-C22.

where the index j denotes the jth function and the index i represents the nodes. Furthermore, comparison of our method for some specific cases of weight functions to the existing sixth-order method [3] and a non-optimal seventh-order Jarratt type method [2] shows that the proposed method is more or equally efficient than already existing methods by taking some nonlinear systems from mechanical, civil and environmental, and bioengineering.

2 Development of the Seventh Order Method

We developed a three-step seventh order method to solve systems of non-linear equations, which is described as:

$$\begin{aligned} Y^{(n)} &= X^{(n)} - G'(X^{(n)})^{-1}G(X^{(n)}), \\ Z^{(n)} &= Y^{(n)} - H(U^{(n)})G'(X^{(n)})^{-1}G(Y^{(n)}), \\ X^{(n+1)} &= Z^{(n)} - L(U^{(n)})K(V^{(n)})G'(X^{(n)})^{-1}G(Z^{(n)}), \end{aligned} \tag{1}$$

where for multidimensional case, $U^{(n)}$ and $V^{(n)}$ are formulated as:

$$\begin{aligned} U^{(n)} &= I - G'(X^{(n)})^{-1}[X^{(n)}, Y^{(n)}; G], \\ V^{(n)} &= I - H(U^{(n)})G'(X^{(n)})^{-1}[Z^{(n)}, Y^{(n)}; G]. \end{aligned}$$

$H, L : A_{n\times n}(\mathbb{R}) \to \Gamma(\mathbb{R}^n), K : B_{n\times n}(\mathbb{R}) \to \Gamma(\mathbb{R}^n)$ with $A_{n\times n}$, $B_{n\times n}$ are the set of $n \times n$ matrices and $\Gamma(\mathbb{R}^n)$ is the set of linear operators from $\mathbb{R}^n$ to $\mathbb{R}^n$.

Theorem 23. *Let us suppose that $G : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a Frechet differentiable function in $\mathbb{D}$ containing simple root Q. We consider that G' be continuous and non-singular at a point Q. The described scheme has seventh-order convergence if we consider initial guess $X^{(0)}$ sufficiently close to the root Q and the proposed seventh-order scheme satisfies the following conditions:*

$$\begin{aligned} &H_0 = H(0) = I, H_1 = H'(0) = 2I, H_2 = H''(0) = -2I, \\ &H_3 = H'''(0) = 36I, \|H^{(iv)}(0)\| < \infty, \\ &L_0 = L(0) = I, L_1 = L'(0) = 2I, L_2 = L''(0) = 0, \|L'''(0)\| < \infty, \\ &K_0 = K(0) = I, K_1 = K'(0) = I, \|K''(0)\| < \infty. \end{aligned}$$

The final error equation is: $E^{(n+1)} = -\frac{1}{6}(-C_3+6C_2^2)C_2^2(-72C_2^2+C_2^2L'''(0)+12C_3){E^{(n)}}^7+O({E^{(n)}}^8)$.

From Theorem 1, several cases of our proposed scheme (1) are obtained for various choices of the weight functions. We consider some specific cases of weight functions for the proposed multidimensional scheme.

HM_1 : We get a seventh-order scheme for multidimensional case by taking the weight functions as:

$$\begin{aligned} H(U^{(n)}) &= I + 2U^{(n)} - (U^{(n)})^2 + 6(U^{(n)})^3, \\ L(U^{(n)}) &= (I - 2U^{(n)} + 4(U^{(n)})^2)^{-1}, \\ K(V^{(n)}) &= (I - V^{(n)})^{-1}. \end{aligned}$$

HM_2 : A seventh-order scheme is obtained for multidimensional system when we take $H(U^{(n)}), L(U^{(n)})$ and $K(V^{(n)})$ as rational functions such that:

$$\begin{aligned} H(U^{(n)}) &= (I + 6U^{(n)})^{-1}(I + 8U^{(n)} + 11(U^{(n)})^2), \\ L(U^{(n)}) &= (I + U^{(n)} - 2(U^{(n)})^2)^{-1}(I + 3U^{(n)}), \\ K(V^{(n)}) &= (I - V^{(n)})^{-1}. \end{aligned}$$

$HM_3:$ We get another multivariate seventh-order scheme by taking the weight functions as:

$$H(U^{(n)}) = (I + \frac{8}{5}U^{(n)} - \frac{11}{5}\left(U^{(n)}\right)^2)^{-1}\left(I + \frac{18}{5}U^{(n)}\right),$$

$$L(U^{(n)}) = I + 2U^{(n)},$$
$$K(V^{(n)}) = I + V^{(n)}.$$

3 Numerical Results

Here, we consider some examples for system of nonlinear equations and obtain numerical results for different cases of our proposed scheme (1). These cases are named as HM_1, HM_2 and HM_3. Now, to check the computational performance and effectiveness of our methods, we compare these results with existing sixth and seventh order methods. First comparison method that we take is a two-point sixth order multivariate method suggested by Behl et al. [3] for the case where $\alpha = \frac{7}{4}$ named as OM. Moreover, we compare our schemes with seventh-order scheme proposed by Behl and Arora [2] denoted as PM. To find the numerical solutions, various systems are taken from different fields of science such as mechanical, civil and environmental engineering, [5], and a bioengineering problem [7]. We take first three iterations in each case in the comparison Tables 1,2 and 3 which includes the number of iterations, n, residual error $\parallel G(X^{(n)}) \parallel_\infty$ and error between two consecutive terms $\parallel X^{(n)} - X^{(n-1)} \parallel_\infty$. Software Matlab R2014a is used to perform all the computations for the multivariate case.

Example 1. ***(Mechanical Engineering)****Let us consider a cylindrical object at room temperature $T_3 = 25\ ^0C$. The innermost section of the cylinder is separated from middle section by vacuum. The outer shell around the cylinder is separated from the middle container by air. The outermost layer of cylinder is in contact with room air.The heat flux between these layers of cylinder is equal i.e. $q = q_1 = q_2 = q_3$. We have to find temperatures T_1, T_2 at steady state and the heat flux q. We take $T_0 = 450\ ^0C$. The 3×3 system of nonlinear equations, $G_1(X) = 0$, is given as:*

$$\begin{aligned} q - 273.245607 + 10^{-9}(T_1 + 273)^4 &= 0, \\ q - 4(T_1 - T_2) &= 0, \\ q - 1.3(T_2 - 25)^{\frac{4}{3}} &= 0. \end{aligned} \tag{2}$$

The desired root of this system is,

$$(137.0988374, 75.85865250, 244.9607397)^t.$$

The initial guess is $X^{(0)} = (138, 74, 245)$.

Table 1: Comparison of Seventh-Order Schemes for Vector-Valued Function $G_1(X)$.

$Cases$	n	$\parallel X^{(n)} - X^{(n-1)} \parallel_\infty$	$\parallel G(X^{(n)}) \parallel_\infty$
HM_1	1	1.8586524981	7.630716552(−11)
	2	4.434213497(−11)	5.327559676(−47)
	3	7.845399845(−59)	5.327559676(−47)
HM_2	1	1.8586524981	4.500340268(−11)
	2	4.146245999(−11)	7.492873218(−47)
	3	1.769517727(−47)	5.327559676(−47)
HM_3	1	1.8586524981	4.532685546(−11)
	2	4.234684718(−11)	4.908886205(−47)
	3	9.879220128(−49)	5.327559676(−47)
PM	1	1.8586514560	6.036597396(−11)
	2	5.567449222(−11)	4.334055473(−47)
	3	1.413897451(−58)	4.334055473(−47)
OM	1	1.85865149236	1.701474025(−7)
	2	1.4980561404(−7)	1.375887236(−30)
	3	9.173626483(−35)	1.376434519(−30)

Example 2. ***(Civil and Environmental Engineering)*** *Let us take five nonlinear system of equations, $G_2(X) = 0$, that describes chemistry of rainwater:*

$$\begin{aligned} k_{h1} &= \frac{10^6[H^+][HCO_3^-]}{K_H P_{CO_2}},\ k_{h2} = \frac{[H^+][CO_3^{2-}]}{[HCO_3^-]},\ k_w = [H^+][OH^-], \\ C_T &= \frac{K_H P_{CO_2}}{10^6} + [HCO_3^-] + [CO_3^{2-}], \\ 0 &= [HCO_3^-] + 2[CO_3^{2-}] + [OH^-] - [H^+]. \end{aligned} \tag{3}$$

Here, k_H =Henry's constant, k_{h1}, k_{h2}, k_w are coefficients of equilibrium.

In order to find five unknowns, C_T= total inorganic carbon, $[HCO_3^-]$ = bicarbonate, $[CO_3^{2-}]$ = carbonate, $[H^+]$ = hydrogen ion, and $[OH^-]$ = hydroxyl ion such that

$$[HCO_3^-] = a, [CO_3^{2-}] = b, [H^+] = c, C_T = d, [OH^-] = e,$$

equations in (3) can be rearranged as:

$$\begin{aligned} g_1 &= k_{h1} k_H P_{CO_2} - 10^6 ac, \\ g_2 &= k_{h2} a - cb, \\ g_3 &= k_w - ce, \\ g_4 &= d - a - b - \frac{k_H P_{CO_2}}{10^6}, \\ g_5 &= a + 2b + e - c. \end{aligned} \tag{4}$$

Given that

$$k_H = 10^{-1.46}, k_{h1} = 10^{-6.3}, k_{h2} = 10^{-10.3}, k_w = 10^{-14}, P_{CO_2} = 315.$$

The required solution is:

$$(0.2337 \times 10^{-5}, 5.0025 \times 10^{-11}, 0.2341 \times 10^{-5}, 0.1325 \times 10^{-4}, 4.2701 \times 10^{-9})^t.$$

The initial approximation is taken as $X^{(0)} = (1, 1, 1, 1, 1)$.

Table 2: Comparison of Seventh-Order Schemes for Vector-Valued Function $G_2(X)$.

$Cases$	n	$\parallel X^{(n)} - X^{(n-1)} \parallel_\infty$	$\parallel G(X^{(n)}) \parallel_\infty$
HM_1	1	8.180985819(−1)	7.210728392(4)
	2	3.147311532(−1)	3.061238662(3)
	3	6.484832440(−2)	1.299616559(2)
HM_2	1	9.914107904(−1)	1.864216660(3)
	2	1.764948121(−1)	6.506165653(1)
	3	3.297207774(−2)	2.2706670513
HM_3	1	9.350523772(−1)	2.257986376(4)
	2	2.744642063(−1)	1.000942430(3)
	3	5.778692776(−2)	4.437075942(1)
PM	1	8.943697089(−1)	4.463103353(4)
	2	3.291662682(−1)	2.178784144(3)
	3	7.272838683(−2)	1.063632162(2)
OM	1	9.600176000(−1)	2.674045563(4)
	2	2.374449999(−1)	1.279572582(3)
	3	5.194109365(−2)	6.122954564(1)

Example 3. ***(Bioengineering)*** *Consider a problem in which we have to determine the steady state concentration of naphthalene in the lung and liver. In steady-state process, concentration of components remains the same i.e. no changes takes place with time. Hence, no material accumulates anywhere in the system. In engineering, a material balance is conducted over each compartment to account for every material that is entering and leaving a system. From this material balancing process, we obtain a system of equations to find unknowns such as concentrations of some specific species from each compartment. Consider a material balancing of naphthalene over two chambers such as lung and liver. Therefore, a* 2×2 *system of equations is obtained,* $G_3(X)$*, and is described as:*

$$780(1-R) + R(0.5C_{liver} + 1.5C_{lung}) - (\frac{8.75C_{lung}}{2.1 + C_{lung}})0.08 - 2C_{lung} = 0,$$

$$0.5C_{lung} - (\frac{118C_{liver}}{7 + C_{liver}})0.322 - 0.5C_{liver} = 0.$$

For given value $R = 0.6$*, the above equations can be written as:*

$$\begin{aligned} 0.3C_{liver} - 1.1C_{lung} - \frac{0.7C_{lung}}{2.1 + C_{lung}} + 312 = 0, \\ 0.5C_{lung} - (\frac{37.996C_{liver}}{7 + C_{liver}}) - 0.5C_{liver} = 0. \end{aligned} \tag{5}$$

We will solve this system for the two unknowns i.e. C_{lung} *and* C_{liver} *given as* $(361.3112614, 287.1278085)^t$. *The initial guess is taken as:* $(C_{lung}^{(0)}, C_{liver}^{(0)}) = (300, 200)^t$.

Table 3: Comparison of Seventh-Order Schemes for Vector-valued Function $G_3(X)$.

$Cases$	n	$\parallel X^{(n)} - X^{(n-1)} \parallel_\infty$	$\parallel G(X^{(n)}) \parallel_\infty$
HM_1	1	8.712780853(1)	2.185470307(−9)
	2	5.961335590(−9)	3.488610168(−46)
	3	3.739677655(−46)	4.023806166(−47)
HM_2	1	8.712780853(1)	2.499859628(−9)
	2	6.818811960(−9)	3.488610168(−46)
	3	3.7396776561(−46)	4.023806166(−47)
HM_3	1	8.712780853(1)	2.009863852(−9)
	2	5.482470512(−9)	3.488610168(−46)
	3	3.739677656(−46)	4.023806166(−47)
PM	1	8.712780854(1)	1.332283585(−10)
	2	3.634956392(−10)	4.023806166(−47)
	3	1.254281853(−58)	4.023806166(−47)
OM	1	8.712780837(1)	5.956254525(−8)
	2	1.619277035(−7)	7.881694568(−31)
	3	2.690016233(−37)	7.8816945677(−31)

The numerical results shows that our newly developed seventh-order method is better or equally efficient as that of already existing methods.

References

[1] Abad, M., Cordero, A., Torregrosa, J. R., A Family of Seventh-Order Schemes for Solving Nonlinear Systems, Bull. Math. Soc. Sci. Math., 57:, 133-145, 2014.

[2] Behl, R., Arora, H., CMMSE: A Novel Scheme having Seventh-Order Convergence for Nonlinear Systems, J. Comput. Appl. Math., 404: 113301, 2022.

[3] Behl, R., Cordero, A., Torregrosa, J. R., High Order Family of Multivariate Iterative Methods: Convergence and Stability, J. Comput. Appl. Math., 405: 113053, 2022.

[4] Behl, R., Kanwar, V., Highly Efficient Classes of Chebyshev-Halley Type Methods Free From Second Order Derivative, IEEE, 1-6, 2014.

[5] Chapra, S. C., Canale, R. P., Numerical Methods for Engineers, The McGraw-Hill Companies, Inc., 1221 Avenue of the Americas, New York, 2010.

[6] 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, Numer. Algor., 87: 1017–1060, 2021.

[7] King, M. R., Mody, N. A., Numerical and Statistical Methods for Bioengineering: Applications in Matlab, Cambridge University Press, New York, 2010.

[8] Ostrowski, A. M., Solution of Equations and Systems of Equations, Academic Press, New York, 1960.

[9] Traub, J. F., Iterative Methods for Solution of Equations, Prentice-Hall, New Jersey, 1964.

A Multiplicative calculus approach to solve nonlinear equations

S. Bhalla ♭,[1] and R. Behl♮

(♭) Chandigarh University,
Department of Mathematics, Chandigarh University, Gharuan, Mohali, India.
(♮) King Abdulaziz University,
Department of Mathematics Sciences, King Abdulaziz University, Jeddah 21589, Saudi Arabia.

1 Introduction

In seventeenth century, Newton and Leibnitz created the differential and integral calculus concept based on subtraction and addition operation. Later on 1970's, Grossman and Katz [1] developed a different definition of differential and integral calculus that utilize the multiplication and division operation instead of addition and subtraction. This definition of differential and integral calculus named as Multiplicative Calculus. In 2008, Bashirov et al. [2], contributed on multiplicative calculus and its applications. After this, some authors worked on some applications of multiplicative calculus in different areas like in biology [3], in science, and finance [4], in the biomedical sciences [5], in economic growth [6] etc.

Motivated from the relevance of multiplicative calculus, some authors [7, 8] apply it to solve multiplicative nonlinear model $g_1(x) = 1$ instead of solving nonlinear equation $g(x) = g_1(x) - 1 = 0$. Basically, these authors worked on simple root finding iterative methods with multiplicative derivatives. In literature, there is no work related to evaluate the multiple root of nonlinear models. So, focusing on same multiplicative derivative concept, we proposed the validity of multiple root finding Schröder Method [9] in multiplicative calculus. We also proposed the joint Schröder Method to enhance the convergence order.

2 Proposed Schemes

Here, we developed the following iterative methods to solve the nonlinear equation in context to multiplicative derivative.

Multiplicative Schröder Method (MSM)

$$x_{k+1} = x_k - \frac{ln\left(g(x_k)\right) ln\left(g^*(x_k)\right)}{\left(ln\left(g^*(x_k)\right)\right)^2 - ln\left(g(x_k)\right) ln\left(g^{**}(x_k)\right)}, \quad \forall k = 0, 1, 2... \tag{1}$$

.

Joint Multiplicative Schröder Method (JMSM)

[1] soniamaths5@gmail.com

$$
\begin{aligned}
y_k =& x_k - \frac{ln\left(g(x_k)\right) ln\left(g^{*}(x_k)\right)}{\left(ln\left(g^{*}(x_k)\right)\right)^2 - ln\left(g(x_k)\right) ln\left(g^{**}(x_k)\right)},\\
x_{k+1} =& y_k - \frac{g(y_k)g'(y_k)}{\left(g'(y_k)\right)^2 - g'(y_k)g''(y_k)}, \quad \forall k = 0, 1, 2...
\end{aligned}
\tag{2}
$$

3 Convergence analysis

Theorem 1: Assume the sufficiently multiplicative differential function $g : \Omega \subseteq \mathbb{R} \to \mathbb{R}^+$ with r_1 multiplicative zero in an open interval Ω. Whenever x_0 is sufficiently close to r_1, then multiplicative Schröder scheme (1) has quadratic convergence.

Theorem 2: Let the function $g : \Omega \subset \mathbb{R} \to \mathbb{R}+$ has a root r in an open interval I. Assume that $g(x)$ is sufficiently ordinary, and multiplicative differentiable about the point r. Then the method JMSM given in (2) has fourth order of convergence.

4 Results

In this section, some experiments have been performed on suggested iterative methods for solving nonlinear equations. All the numerical work has been done on *Mathematica* 11 Software with the stopping criterion $|g(x_k)| < 10^{-50}$ in ordinary derivative case and in multiplicative derivative case $|g_1(x_k) - 1| < 10^{-50}$. The experimental work of proposed multiplicative Schröder method(MSM), and joint multiplicative Schröder method($JMSM$) are compared with ordinary Newton methods(NM), ordinary Schröder method(SM), modified Newton's method($MONM$) [9], and multiplicative Newton method(MNM) [7]. In every numerical work the approximate computational order of convergence(ACOC) ρ, iteration index k, and consecutive iteration error $|x_{k+1} - x_k|$ presented in [10] are evaluated and showed in Table 1.

Example 1. *We consider the population growth model that formulate the following nonlinear equation*

$$g(x) = \frac{1000}{1564}e^x + \frac{435}{1564}(e^x - 1) - 1.$$

In this model we evaluate the birth rate denoted as x, if in a specific local area has 1000 *thousand people at first and* 435 *thousand move into the local area in the first year. Likewise, assume* 1564 *thousand individuals toward the finish of one year. The computed results towards the root* $x_r = 0.1009979\ldots$ *are displayed in Table 1. Clearly, the methods* MSM*, and* $JMSM$ *shows better results in terms of consecutive error and number of iteration in comparison of existing ones.*

5 Conclusions

This paper presents a Schröder iterative method and joint Schröder iterative method with use of multiplicative derivative. The convergence analysis of the constructed schemes are studied. In order to measures its efficiency the proposed techniques are tested to find the roots of nonlinear equation with multiplicity $m \geq 1$. The numerical results depict that the new Schröder methods provides satisfactory outcomes as compared with other techniques on the basis of errors, and iteration index.

Table 1: Convergence behavior of the methods NM, MNM, SM, MSM, $JMSM$ at approximation $x_0 = 1$.

$Schemes$	k	$\lvert x_{(k+1)} - x_{(k)}\rvert$	ρ	$Total\ number\ of\ iterations$	$CPU\ time(Seconds)$
NM	2	$3.7(-2)$	2.000	5	0.281
	3	$6.7(-4)$			
	4	$2.1(-7)$			
MNM	2	$2.6(-5)$	1.999	4	0.359
	3	$2.9(-11)$			
	4	$3.6(-23)$			
SM	2	$4.2(-1)$	2.006	6	0.329
	3	$1.1(-1)$			
	4	$5.8(-3)$			
MSM	2	$1.7(-2)$	2.001	4	0.313
	3	$1.2(-5)$			
	4	$6.4(-12)$			
$JMSM$	2	$1.3(-4)$	4.000	4	0.469
	3	$2.5(-19)$			
	4	$3.3(-78)$			

References

[1] Grossman M, Katz R. Non-Newtonian Calculus. Pigeon Cove, MA, USA: Lee Press, 1972.

[2] Bashirov, Agamirza E and Kurpınar, Emine Mısırlı and Özyapıcı, Ali, Multiplicative calculus and its applications. J Math Anal Appl 2008; 337:36-48.

[3] Englehardt, J., Swartout, J., Loewenstine, C.: A new theoretical discrete growth distribution with verification for microbial counts in water. Risk Anal. 29, 841–856 (2009).

[4] Bashirov, A.E., Misirli, E., Tandogdu, Y., Özyapıcı, A.: On modeling with multiplicative differential equations. Appl. Math. J. Chin. Univ 26, 425—428 (2011).

[5] Florack, L., Assen, H.: Multiplicative calculus in biomedical image analysis. J. Math. Imaging Vis. 42, 64–75 (2012).

[6] Filip DA, Piatecki C. A non-Newtonian examination of the theory of exogenous economic growth. HAL 2014: hal-00945781.

[7] Özyapıcı, Ali and Sensoy, Zehra B and Karanfiller, Tolgay, Effective Root-Finding Methods for Nonlinear Equations Based on Multiplicative Calculi, Journal of Mathematics, Hindwai Publications, 2016.

[8] Özyapıcı, Ali, Effective numerical methods for non-linear equations, International Journal of Applied and Computational Mathematics, 6(2), 1–8, 2020.

[9] Traub, J. F.: Iterative Method for the Solution of Equations. Prentice-Hall. Englewood cliffs (1964).

[10] Behl, R.; Bhalla, S.; Magreñán, ÁA.; Moysi, A. An Optimal Derivative Free Family of Chebyshev–Halley's Method for Multiple Zeros. Mathematics,(2021) 9, 546.

[11] J.M. Douglas, *Process Dynamics and Control.* vol. 2 Prentice Hall, Englewood Cliffs **1972** .

[12] Chicharro, F.I., Cordero, A., Torregrosa, J.R.:Drawing dynamical and parameters planes of iterative families and methods. Sci. World J. **11**, Article ID 780153 (2013).

Basins of Convergence in a Modified CR3BP

D. Villalibre$^{\flat 1}$, A. Herrero$^{\natural}$, J.A. Moraño$^{\natural}$ and S. Moll$^{\diamondsuit}$

($\flat$) Escuela Técnica Superior de Ingeniería del Diseño, Universitat Politècnica de València.
($\natural$) Instituto de Matemática Multidisciplinar, Universitat Politècnica de València.
($\diamondsuit$) Departamento de Matemática Aplicada, Universitat Politècnica de València.

1 Introduction

The 3-body problem is a well-known problem in classical mechanics [1]. An approximate mathematical model is the circular restricted 3-body problem (CR3BP) [2,3], based on differential equations under specific hypotheses that allow its simplification and numerical simulation. In [4], the CR3BP is studied adding additional factors to the mathematical model corresponding to the gravitational potential and the resulting basins of convergence. On the other hand, the angular velocity has also been considered in [5] similarly.

The study of the stationary points (also called Lagrange points) carries a significant difficulty, so several simplifying hypotheses are usually incorporated in the literature. The CR3BP [1] reduces the complexity by considering only the masses of the two primary bodies, which orbit circularly around their center of masses. The mass of the third body is considered negligible for the problem. In addition, the CR3BP considers that all the bodies' orbits are on the same plane.

To better analyse the orbit of the third mass, different CR3BP-modified models have been used in the literature, considering some perturbing accelerations by modifying the potential function [2–4]. In these papers, the effects of irregularities on the shapes of the primary masses, the radiation pressure, or relativistic effects, are considered.

A basin of convergence (BC) is a set of points where the dynamics of an infinitesimal mass with null velocity will evolve to a Lagrange point. These regions and the equilibrium points are calculated using numerical methods, such as Newton–Raphson's (NR) [6] or Broyden's [7] methods. In fact, the NR method has been used in many works to study the BC in different restricted problems of 3-bodies with modifications in the potential function, to consider radiation pressure and oblateness effects, see for example [4, 5, 8].

The goal of this work is to analyze how to introduce simultaneously several additional perturbations to the CR3BP model to increase its precision. These additional factors affect the dynamics of the model by considering the radiation forces, the flattening of the primary bodies, the presence of natural satellites, and some relativistic effects. Once modified, the resulting model is integrated by using the NR method to determine the position of the Lagrange points and the basins of convergence of the system.

This paper is structured as follows: in Section 2 the dynamical model description is given by adding different perturbations to the classical formulation. In addition, the developed algorithm to calculate the Lagrange points and the convergence basins is presented in Section 3. Next, in

[1] daniel.villalibre.vilarino@gmail.com

Section 4, the algorithm is applied to study the Sun - Mars system considering different effects. Finally, in Section 5, the conclusions can be found.

2 Model Description

The CR3BP considers the masses of two primary bodies, m_1 and m_2, orbiting circularly around their center of masses and on the same plane together with the third infinitesimal body. The corresponding equations are usually written in a rotating coordinate system, whose origin is the center of masses of the primaries system. It has an angular velocity equal to the mean motion of the orbits of the primaries which is given by:

$$n_D = \sqrt{\tfrac{G\,(m_1+m_2)}{R^3}} \tag{1}$$

where R is the distance between the primary bodies and G is the gravitational constant. Besides, the equations of the CR3BP are expressed in dimensionless terms, taking the appropriate unit measuring system. Then, the equations describing the dynamics of the infinitesimal mass from classical mechanics are:

$$\ddot{x} - 2\dot{y} = \frac{\partial \Omega}{\partial x} \qquad \ddot{y} + 2\dot{x} = \frac{\partial \Omega}{\partial y} \tag{2}$$

where x and y are the coordinates of the infinitesimal mass and Ω the potential function defined as

$$\Omega\,(x, y) = \tfrac{1}{2}\left(x^2 + y^2\right) + \tfrac{1-\mu}{r_1} + \tfrac{\mu}{r_2} \tag{3}$$

with r_1 and r_2 the distances between the infinitesimal body and each one of the primary masses, and $\mu = m_2/(m_1 + m_2)$ the dimensionless mass parameter. This parameter gives the positions of the primary bodies as:

$$x_1 = -\mu \quad x_2 = 1 - \mu \quad y_1 = 0 \quad y_2 = 0 \tag{4}$$

Some additional effects will be separately considered in the equations following several studies [3,8–11]. These effects will be added to the effective potential modifying the effective potential in a cumulative way. The final potential (Ω^*) would incorporate all these effects. Since the physical and geometrical properties of the primary masses also affect the mean motion of the orbit, a modified mean motion n_D^* must be used. Then, the equations of motion are:

$$\ddot{x} - 2\,n\,\dot{y} = \frac{\partial \Omega^*}{\partial x} \qquad \ddot{y} + 2\,n\,\dot{x} = \frac{\partial \Omega^*}{\partial y} \tag{5}$$

where $n = n_D^*/n_D$ is the dimensionless corrected mean motion.

The first effect to be considered will be the radiation forces produced by one or both primary masses. Considering this radiation as isotropic, from [2,5] its effects are introduced in the dynamic system by means of a multiplicative radiation factor q_i in the terms associated with the primary masses. This factor is given by a relation between the gravitational effects F_{g_i} and the effects of the forces associated with radiation F_{r_i}

$$q_i = 1 - \tfrac{F_{r_i}}{F_{g_i}}. \tag{6}$$

Next, the flattening or oblateness of the primary masses is considered. Several authors have modified the effective potential in order to introduce these new geometries [3, 9, 10]. Then, 3 semi-axes, a_j, b_j and c_j, have to be considered in each primary mass j, measured on a right-hand trihedral axis. Thus, the triaxiality parameters σ_{i_j} are defined by:

$$\sigma_{1_j} = \frac{a_j^2 - c_j^2}{5R^2} \qquad \sigma_{2_j} = \frac{b_j^2 - c_j^2}{5R^2} \tag{7}$$

and, each primary mass will have two triaxiality functions $f_{i_j}, i = 1, 2$ given by:

$$f_{1_j} = 2\sigma_{1_j} - \sigma_{2_j} \qquad\qquad f_{2_j} = \sigma_{2_j} - \sigma_{1_j} \tag{8}$$

which will appear in the modified effective potential.

In order to introduce the relativistic effects, the effective Schwarzschild potential is considered, which introduces a term of the form ϵ/r^3. According to the development shown in [4], the parameter $\epsilon \in [0, 1]$ is used to incorporate these relativistic effects in the term of the second primary mass potential function.

Finally, the effect of the presence of natural satellites in the CR3BP model is studied in a similar way as in [12]. It will be assumed that the satellites are spherical bodies with a significantly lower mass than the primaries and they are not a source of radiation.

For each satellite $i = [1, n_s]$ its mass is denoted by m_{si} and it modifies the effective potential with an additive factor of the form α_i/r_{s_i}, where and $r_{s_i}\alpha_i = m_{si}/(m_1 + m_2)$ and r_{s_i} is the position of the satellite.

Therefore, the effective potential considering all the aforementioned potential effects is:

$$\Omega^*(x, y, \tau) = \frac{n^2}{2}(x^2 + y^2) + \frac{1-\mu}{r_1}\left(q_1 + \frac{f_{1_1}}{2r_1^2} + \frac{3y^2 f_{2_1}}{2r_1^4}\right) + \frac{\mu}{r_2}\left(q_2 + \frac{f_{1_2}}{2r_2^2} + \frac{3y^2 f_{2_2}}{2r_2^4} + \frac{\epsilon}{r_2^2}\right) + \sum_{i=1}^{n_s} \frac{\alpha_i}{r_{s_i}} \tag{9}$$

On the other hand, the influence of the considered effects on the mean motion n has to be analyzed too. This analysis leads to that the radiation and the natural satellites effects do not affect to the mean motion of the primaries while the oblateness and the relativistic effects modify the mean movement resulting in

$$n = \sqrt{\left[1 + \frac{3}{2}(2\sigma_{1_1} - \sigma_{2_1}) + \frac{3}{2}(2\sigma_{1_2} - \sigma_{2_2})\right](1 + 3\epsilon)} \tag{10}$$

Then, the equations of motion are (5) with n and Ω^* given by (10) and (9), respectively.

3 Lagrange Points and Basins of Convergence

The Lagrange points of the system are the extrema of the resulting effective potential (9). At these points the position of the third mass can remain stationary since the gravitational field exerted by the primary bodies is compensated by the centrifugal acceleration of the infinitesimal mass. The basins of convergence determine towards which Lagrange point an infinitesimal mass evolves, given its initial conditions (x_0, y_0) and assuming a null initial velocity $(\dot{x}_0 = \dot{y}_0 = 0)$.

In general, in classical mechanics, 5 Lagrange points appear in the study of the effective potential. However, when perturbations are added, the number of Lagrange points can increase [9]. Since, the equations that will define them are:

$${\Omega^*}_x = 0, \quad {\Omega^*}_y = 0 \tag{11}$$

a numerical method can be used to solve them. In order to apply an iterative method, a portion of the planed will be divided into $n_x \times n_y$ elements uniformly distributed. The centroid of each of these cells will be taken as initial conditions.

One of the most common numerical methods used to find zeros of functions is the NR method. In fact, many works employing it to solve the equations (11) can be found in the literature [4–6]. This method, based on a first-order linear approximation of the function, can be formulated as:

$$\mathbf{F}(\mathbf{x}_{n+1}) = \mathbf{F}(\mathbf{x}_n) + \mathbf{F}'(\mathbf{x}_n)(\mathbf{x}_{n+1} - \mathbf{x}_n) \tag{12}$$

where $\mathbf{x} = \{x, y\}^T$, $\mathbf{F}(\mathbf{x}) = \{\Omega^*{}_x, \Omega^*{}_y\}^T$, and $\mathbf{F}'(\mathbf{x}_n)$ represents the jacobian matrix of the function $\mathbf{F}(\mathbf{x})$ evaluated at the point $\mathbf{x}_n$. Then, the zeros of the function $\mathbf{F}(\mathbf{x}_{n+1})$ are iteratively obtained:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - (\mathbf{F}'(\mathbf{x}_n))^{-1}\mathbf{F}(\mathbf{x}_n). \tag{13}$$

A solution is considered to converge when the distance between two consecutive iterations is below a given precision threshold. Moreover, two obtained solutions will be considered equal if they are close enough, that is, if the distance between them is less than a given threshold. Once all the possible solutions have been determined, they define the different Lagrange points of the problem. Afterwards, the solution reached in each cell is compared with the Lagrange points to assign the cell to the corresponding Lagrange point. The convergence basins of the system are the graphic representations of these identifications.

Then, a modular code has been developed, which allows to compare the obtained solutions for the different effects one by one with the existing in the literature [4, 5, 9]. In these articles, each one of the effects appear separately. Using the same parameters values for each situation the same results have been obtained, which validates the code developed in this paper.

All the effects considered are validated in the literature, except for the natural satellites perturbations. This effect has been studied in [12] for a single satellite with a different formulation. Moreover, it is worth mentioning that our algorithm allows considering more than one satellite.

4 Application: Sun - Mars System

In this section, the algorithm is applied to the Sun - Mars system including the effects of the triaxiality of the primary masses, the radiation forces, and the presence of the natural satellites of Mars, Phobos, and Deimos. The solution is compared to the classical CR3BP one. The masses of the Sun and Mars and the distance between them have been taken from [13,14], where the distance is considered equal to the semi-major axis of the elliptical orbit of Mars. These values give the mass parameter, $\mu = 3.22710 * 10^{-7}$ [$-$]. Under these considerations, the convergence basins of the problem in the classical CR3BP formulation are shown in Figures 1, 2 and 3 for different sizes in the domain. The estimated coordinates of the Lagrange points in this case are shown in Table 1.

Clearly, a larger convergence region has been obtained than in the validation process (Sun-Earth system) since the dimensionless parameter of masses is smaller in this case. It can also be observed that the Lagrange points are located very close to the primary masses, specially to Mars in the case of L_1 and L_2 (see Figure 3). On the other hand, L_4 and L_5 remain at the midpoint between the primary masses, as it can be seen in Table 1.

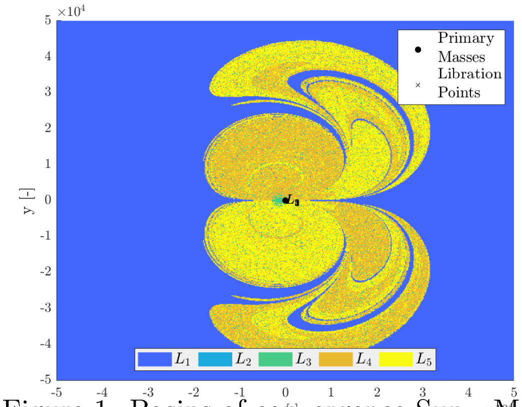


Figure 1: Basins of convergence Sun - Mars (classical formulation) [broad domain]

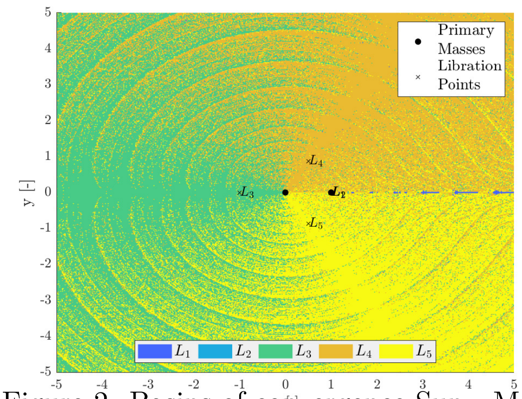


Figure 2: Basins of convergence Sun - Mars (classical formulation) [reduced domain]

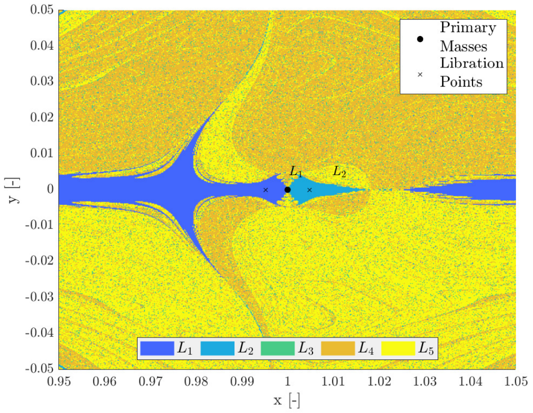


Figure 3: Bases of convergence Sun - Mars (classical formulation) [zoom Mars]

Libration Points	Obtained Coordinates $(x_L, y_L)\,[-]$
L_1	(0.99525, 0)
L_2	(1.00476, 0)
L_3	(-1.00000, 0)
L_4	(0.50000, 0.86603)
L_5	(0.50000, -0.86603)

Table 1: Coordinates of the Sun - Mars libration points (classical formulation)

4.1 Triaxiality and Satellites Effects

Now, in order to introduce the triaxiality and satellites effects together, the necessary Sun and Mars data are taken from [13] and [14], respectively. And the corresponding to Phobos and Deimos are obtained from [15] and [16]. These data allow to calculate the oblateness parameters appearing in the model described in Section 2. Note that the terms associated with the Sun hardly change due to its very low ellipticity, while, the corresponding to Mars subtancially change because of a greater ellipticity. In this case the location of the Lagrange points has hardly changed by the inclusion of both effects with respect to the classical formulation. This can be explained because the terms associated with the Sun have more weight than those associated with Mars and its satellites, and because the star is practically spherical. However, a slight change can be seen in far-field solutions.

4.2 Triaxiality and Radiation Effects

In this subsection the effects of the Sun radiation forces are included along with triaxiality effects. The satellites are not included in the simulation since their effect has been proven to be negligible. For the simulation an arbitrary value of radiation factor $q_1 = 0.4\,[-]$ for the Sun and a null radiation factor for Mars ($q_2 = 0\,[-]$) have been chosen. The results are shown in Figures 4, 5 and 6 and Table 2. It can be seen that the radiation does have an important effect on the convergence basins and on the Lagrange points.

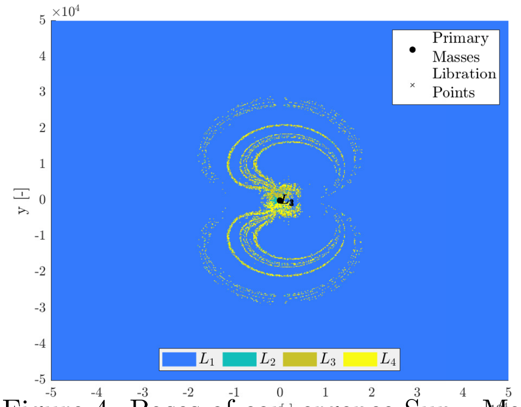


Figure 4: Bases of convergence Sun - Mars (triaxiality + radiation effects) [broad domain]

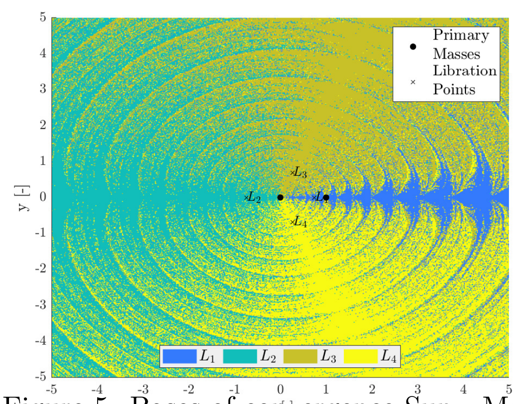


Figure 5: Bases of convergence Sun - Mars (triaxiality + radiation effects) [reduced domain]

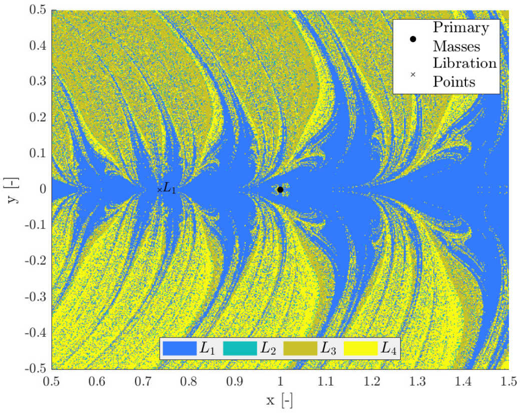


Figure 6: Bases of convergence Sun - Mars (effects of triaxiality + radiation) [zoom Mars]

Libration Points	Obtained Coordinates (x_L, y_L) $[-]$
L_1	(0.73680, 0)
L_2	(-0.73681, 0)
L_3	(0.27144, 0.68498)
L_4	(0.27144, -0.68498)

Table 2: Coordinates of the Sun - Mars libration points (triaxiality + radiation effects)

In this case, the number of Lagrange points has been reduced to four and the convergence regions also change with respect to those in the classical formulation. Moreover, it can be observed that the triangular points are now closer to the Sun and the collinear points are located almost symmetrically with respect to the Sun.

5 Conclusions

In this paper, the classical CR3BP has been modify to increase the precision of the model by adding to the effective potential the effects of the radiation forces, the oblateness of the primary masses, some relativistic effects and satellites contribution.

An algorithm has been developed in a modular way to calculate the Lagrange points and the convergence basins of the model. It has been seen that the results obtained in the literature are recovered when working in the same conditions.

The code has been applied to the Sun - Mars system including triaxiality, radiation effects and Phobos and Deimos contribution. In this case, it is observed that the convergence basins and the Lagrange points are mainly affected by the radiation effects since the terms in the effective potential associated to the star are more important than those of Mars.

6 Acknowledgment

This research has been partially supported by the Spanish Ministerio de Ciencia, Innovación y Universidades and the Fondo Europeo de Desarrollo Regional, Project PID2019-109753GB-C22.

References

[1] Krishnaswami, G.S.; Senapati, H. An Introduction to the Classical Three-Body Problem. *Resonance* **2019**, *24*, 87–114. DOI: https://doi.org/10.1007/s12045-019-0760-1

[2] Chernikov, Yu. A. The photogravitational restricted three-body problem. *Soviet Astronomy* **1970**, *14*, 176. DOI: http://adsabs.harvard.edu/full/1970SvA....14..176C (Accessed on February, 9 2022).

[3] Elshaboury, S.M.; Abouelmagd, E.I.; Kalantonis, V.S.; Perdios E.A. The planar restricted three-body problem when both primaries are triaxial rigid bodies: Equilibrium points and periodic orbits. *Astrophysics Space Science* **2016**, *361*, 315. DOI: https://doi.org/10.1007/s10509-016-2894-x

[4] Zotos, E.E.; Chen, W.; Abouelmagd, E.I.; Han, H. Basins of convergence of equilibrium points in the restricted three-body problem with modified gravitational potential. *Chaos, Solitons & Fractals* **2020**, *134*, 109704. DOI: https://doi.org/10.1016/j.chaos.2020.109704

[5] Suraj, M. S.; Aggarwal, R.; Mittal, A.; Asique, M.C. The perturbed restricted three-body problem with angular velocity: Analysis of basins of convergence linked to the libration points. *International Journal of Non-Linear Mechanics* **2020**, *123*, 103494. DOI: https://doi.org/10.1016/j.ijnonlinmec.2020.103494

[6] Zotos, E.E. On the Newton–Raphson basins of convergence of the out-of-plane equilibrium points in the Copenhagen problem with oblate primaries. *International Journal of Non-Linear Mechanics* **2018**, *103*, 93–103. DOI: .

[7] Broyden, C.G. A class of methods for solving nonlinear simultaneous equations, mathematics of computation. *Am. Math. Soc.* **1965**, *19*, 577—593.

[8] Zotos, E.E.; Suraj, M.S.; Aggarwal, R.; Satya, S.K. Investigating the basins of convergence in the circular Sitnikov three-body problem with non–spherical primaries. *Few-Body Syst.* **2018**, *59*, 69. DOI: https://doi.org/10.1007/s00601-018-1393-8

[9] Saeed, T.; Zotos, E.E. On the equilibria of the restricted three-body problem with a triaxial rigid body - I. Oblate primary. *Results in Physics* **2021**, *23*, 103990. DOI: https://doi.org/10.1016/j.rinp.2021.103990

[10] Duggad, P.; Dewangan, S.; Narayan, A. Effects of triaxiality of primaries on oblate infinitesimal in elliptical restricted three body problem. *New Astronomy* **2021**, *85*, 101538. DOI: https://doi.org/10.1016/j.newast.2020.101538

[11] Shao-wu, S.; Da-zhu, M. Research on the Stability of the Circular Restricted Three-body Problem with Radiation and Oblateness. *Chinese Astronomy and Astrophysics* **2019**, *43(4)*, 549–562. DOI: https://doi.org/10.1016/j.chinastron.2019.11.006

[12] Huang, S.S. Very Restricted Four-Body Problem. *The Astronomical Journal* **1960**, *70*, 347. DOI: https://doi.org/10.1086/108151 (Accessed on February 9, 2022)

[13] NASA, Sun fact sheet 2018. Available online: https://cutt.ly/eWrGKBK (Accessed on February 9, 2022).

[14] NASA, Mars fact sheet. Available online: https://cutt.ly/aWrG7u2 (Accessed on February 9, 2022).

[15] N.T. Redd, Phobos: Facts about the Doomed Martian Moon. Available online: https://cutt.ly/VWrJwuI, Space.com, (2017) (Accessed on February 9, 2022).

[16] N.T. Redd, Deimos: Facts about the smaller martian moon. Available online: https://cutt.ly/5WrHXds, Space.com, Space (2017) (Accessed on February 9, 2022).

SPECIAL SESSION: STUDENT'S PROJECTS

53

Predicting PDA battery health using machine learning methods

Alberto Bono Monreal[♭], Irene Cánovas Vidal[♭], Eduardo Gómez Fernández[♭], Rubén Marco Cabanes[♭], Andrea Pérez López[♭], Alejandra Sánchez Torres[♭] and Marta Valero Buj[♭] (♭)

Data Science Degree,
Universitat Politècnica de València,
Camì de Vera s/n, València, Spain.

Abstract

In this project, we have developed an application through which the Spanish multinational company Inditex can obtain objective information about the health of the batteries of their PDAs. To implement this application, we developed an executable where we generated a Classification Trees model. Through this method, the PDA is classified depending on the range of its battery health using the SMOTE technique due to the unbalanced data found in the sample.

1 Background

Batteries currently used in electronic devices are made of lithium due to the advantages of high energy density and low discharge rate, which makes lithium batteries a suitable energy storage solution for many applications [1]. In fact, in addition to dominating the electronic device market, it also dominates the automotive market. However, the main problem is that these batteries become less effective over time [3].

Because lithium batteries have many types of applications, the development of the technology is increasing, especially in the automotive market, where the demand for more efficient batteries is continuously growing to provide competitive technology in terms of range and durability. In addition, governments are taking measures to stimulate this type of vehicle, contributing to sustainability and reducing polluting gases [2]. All this means that the use of lithium and its value is steadily increasing, while lithium reserves are being depleted. That is why we must try to maximise the life of these batteries, studying their behaviour and what habits we should follow to take care of the batteries' health. In our case, when we want to predict when it is necessary to change a device due to the poor state of the battery, we have to try to be as accurate as possible to avoid this waste of lithium.

The battery durability and health of the world of bateries is still very complex and challenging to understand. In addition, the environmental impact of lithium-ion batteries depends significantly on the battery's lifetime, which is limited by the degradation of the battery. Because of this, there are a multitude of studies that discuss what factors contribute to an increase or decrease in battery life. In fact, N. Omar says that the impact of operating temperature is a determining factor in slowing down battery ageing [3], whereas voltage changes under different charging currents can reduce the electrical performance of batteries [4].

Our research simply classified where the battery of an android device currently stands in its life, based on information extracted from the device itself and information on the use of the device provided by third parties. To do this, we have relied on the methodology used by Apple to provide an estimated battery health value. Determining the state of battery health is a pending task for Android, as its main competitor in the mobile market (Apple) is ahead in this area. As already mentioned, the latter company estimates the percentage of useful life remaining in the device's battery, and depending on the value obtained, it is advisable or not to replace the device (the advisable value for replacing the device in iOS is usually around 40%)[1].

The main drawback is when an Android user wants to know when it is the right time to replace their old device and buy a new one because of the battery condition. This problem is intensified in the case of department stores that work with Android devices constantly and rely heavily on them. These devices are often referred to as Personal Digital Assistant (PDA) and are widely used in companies for functions such as labelling and relabelling, distribution, shipping, etc.

The research work we have carried out focuses on the study and behaviour of the batteries of Android devices. This work has been carried out with the help of INDITEX, a Spanish multinational textile manufacturing, and distribution company, whose objective with the collaboration in this work is to be able to analyse their PDA devices used in warehouses and other places and to be able to carry out a diagnosis, in order to quantify the number of devices in poor condition. The aim is to anticipate events and be able to have the necessary stock of PDAs, as well as to reduce costs and contribute to sustainability by avoiding unnecessary purchases of material.

2 Data availability and software

The data has been provided by Inditex using an Android device to extract information for the PDAs. The tool used to extract information from our Android devices is dumpsys [2]. You can call dumpsys to get a diagnostic result for all system services running on a connected device. We will focus on battery usage in our case so that the dumpsys command batterystats will be used[3].

Figure 1 shows the process we have carried out to obtain the final database. When a PDA is connected, the information is extracted by dumpsys, and we get as many TXTs as PDAs. It must be cleaned and processed to turn it into a proper database. We first carried out a process of investigating the meaning of the variables. After eliminating those unrelated to our objective, we eliminated missing values, as we understand that the type of device 'PDA' does not create information of this type, and we carry out the necessary transformations.

In addition to the TXT with the information of each PDA extracted through battery stats, our Inditex contact also shared with us an Excel with information like the function performed by each PDA, whether the PDAs have an RFID device, the time of use... and the state of the battery, our response variable. These variables are also added to the final database (Figure 1).

One of the problems we encountered in making these transformations is that the difference in the number of values for each PDA is abysmal for each variable. Therefore, variables with few values or numerical variables have a simpler transformation. But in the case of variables with around 400 values and categorical variables, the transformation becomes more complicated (Figure 2).

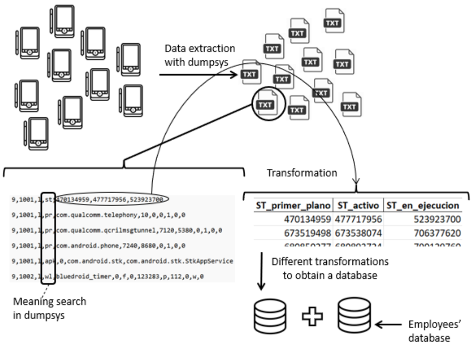
We performed two different modifications, one for the numerical data and one for the qualitative data, this can be seen in Table 1.

In addition, we perform transformations to move from qualitative to numerical data (Table 2).

[1] https://appletoolbox.com/at-what-percentage-replace- iphone-battery/

[2] https://developer.android.com/studio/command-line/dumpsys

[3] https://developer.android.com/studio/command-line/dumpsys#battery



```
9,1001,1,st,470134959,477717956,523923700
9,1001,1,pr,com.qualcomm.telephony,10,0,0,1,0,0
9,1001,1,pr,com.qualcomm.qcrilmsgtunnel,7120,5380,0,1,0,0
9,1001,1,pr,com.android.phone,7240,8680,0,1,0,0
9,1001,1,apk,0,com.android.stk,com.android.stk.StkAppService
9,1002,1,wl,bluedroid_timer,0,f,0,123283,p,112,0,w,0
```

ST_primer_plano	ST_activo	ST_en_ejecucion
470134959	477717956	523923700
673519498	673538074	706377620

Figure 1: Process carried out to obtain the database

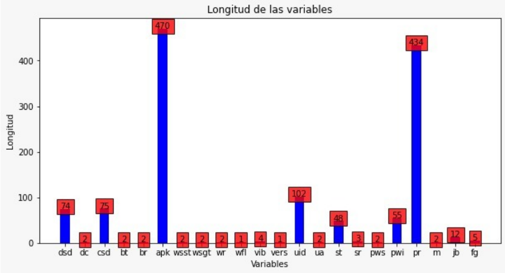
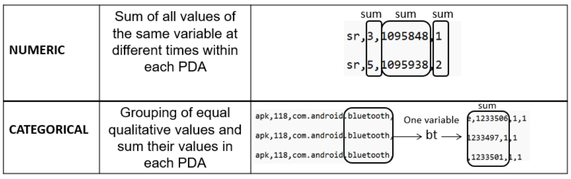
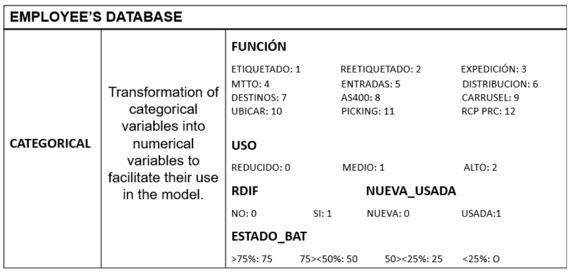


Figure 2: Number of times each variable is repeated



NUMERIC	Sum of all values of the same variable at different times within each PDA	sum sum sum sr,3,1095848,1 sr,5,1095938,2
CATEGORICAL	Grouping of equal qualitative values and sum their values in each PDA	apk,118,com.android.bluetooth, apk,118,com.android.bluetooth, apk,118,com.android.bluetooth One variable → bt → sum e,1233506,1,1 1233497,1,1 ,1233501,1,1

Table 1: Transformations made on numerical and categorical variables



EMPLOYEE'S DATABASE		
CATEGORICAL	Transformation of categorical variables into numerical variables to facilitate their use in the model.	**FUNCIÓN** ETIQUETADO: 1 REETIQUETADO: 2 EXPEDICIÓN: 3 MTTO: 4 ENTRADAS: 5 DISTRIBUCION: 6 DESTINOS: 7 AS400: 8 CARRUSEL: 9 UBICAR: 10 PICKING: 11 RCP PRC: 12 **USO** REDUCIDO: 0 MEDIO: 1 ALTO: 2 **RDIF** **NUEVA_USADA** NO: 0 SI: 1 NUEVA: 0 USADA:1 **ESTADO_BAT** >75%: 75 75><50%: 50 50><25%: 25 <25%: O

Table 2: Transformations of categorical variables into numerical

As can be seen in the Figure 3, the battery's state can be less than 25, greater than 25 but less than 50, greater than 50 but less than 75, or greater than 75. We observed that most PDAs have a high battery status. We did find scant values of the battery status group below 25 and even less between 50 and 25.

Thus, in our final database, we have obtained a good number of rows with 83 variables, combining manually extracted data and data extracted from dumpsys.

3 App

All the preprocessing done is combined in an application. By connecting a PDA and running the app, the data will be downloaded, and the PDA diagnostics will be performed, giving you back the value of the battery health status.

This can be seen in the following video: **SaludPDAs**

4 Methods

A classification problem is characterized by having a qualitative variable Y as the answer. This dependent variable is a standard quality of the observations in the data set that we want to predict. The prediction of the category or class to which an individual belongs within this variable is to assign a class to that observation. In our case, we are faced with a clear classification problem in which the qualitative variable to be predicted the state of the PDA's battery with four different classes. Our main goal is to build a simple, intelligent, and robust classification model in an efficient and scalable way to perform a correct classification of PDA battery status. In order to carry out

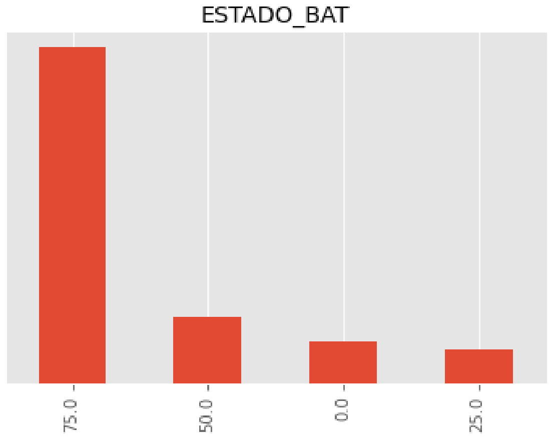
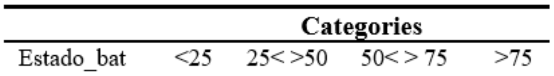


Figure 3: Bar chart of the number of samples of each class.



	Categories			
Estado_bat	<25	25< >50	50< > 75	>75

Table 3: Manually provided battery status categories.

the modeling we have to deal with some limitations of the models used, for example, we cannot work with incomplete information. That is when unknown values appear in some attributes of the training cases. For this purpose, once the database has been loaded, a cleaning function has been applied to eliminate the NaN and Inf values from the database.

Due to the data collection format, we can only have missing data in any input variable if it, in turn, contains null data in the response variable "STAT_BAT".

Futhermore, the variables of non-numerical character will be eliminated because they cannot be treated and are not fundamental at the time of classification. In our case, we eliminate PDA and battery identifiers and the date of manufacture of the battery.

One of our main problems is that we face a classification problem with unbalanced classes, so we decided to try several oversampling techniques. There are a few reasons why undersampling is not possible in this Project: First, when you undersample, you are essentially throwing away data, and we have a small database. This can lead to the sample of the majority class chosen could be biased; Second, undersampling can lead to problems with your model's ability to generalize to new data; Third, during undersampling, you are removing data from the majority class. Because of this, you can lose potentially important information. And fourth, it may not be possible to accurately identify which instances are most important to keep. Mainly we opted for SMOTE, since, in our situation, adding data is preferable to removing data.

SMOTE first selects a minority class instance at random and finds its k nearest minority class neighbors. Then an instance is created by choosing one of the k nearest neighbors at random and connecting them to create a new observation in the feature space. The synthetic instances are generated as a convex combination of the two chosen neighbors' instances[4]. This technique is only applied to the training set because if you modify the test set, you will predict non-real data.

Another limitation is the small number of observations we have. By creating new observations with the oversampling method, we increase the size of the database, but still, due to the scope of this project, the number of observations we have is limited.

To evaluate the model's predictive capacity, it is necessary to check how close its predictions are to the valid values of the response variable. To quantify it correctly, it is essential to have a set of observations of which the response variable is known but which the model has not "seen", that is to say, which have not participated in its adjustment. So, the step before applying a classification method is to partition the data set into two smaller data sets that will be used for the following

[4] https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

purposes: training and test.

The training data subset is used to estimate the model parameters, and the test data subset is used to check the behavior of the estimated model.

To carry out the validation, we apply two techniques, a k-fold division with $k = 5$ and an allocation of 70% of the data to the train set and 30% to the test set. In both techniques, we will do it in a stratified way; that is to say, the division is done by keeping the same proportion of classes in the two sets.

The K-fold method divides a data set into K blocks. Considering one of them as a validation set and the rest (K-1) as the training set. The first combination is the one that considers the first block as a validation set and the rest as a training set, and so on with the rest of the combinations.

Firstly, k-fold splitting was proposed to vary the train and test set to solve the problem of having too little data. Using this technique, we split the data[5] times successively into train and test sets, respectively. Once we have the candidate models with the selected optimal parameters, we evaluate them by performing cross-validation with these sets. This technique was discarded since we later received more data and were able to apply a single split.

Finally, the evaluation is carried out with a single split of the data into a train and a test set. This is done by training the model with 70% of the observations to which the SMOTE algorithm has been applied and validating it with cross-validation. The 30% set will be used exclusively to evaluate the predictive capability of the proposed model.

To find out which specific parameters were the most suitable for each of the possible estimators, the GridSearchCV tool was used. This technique performs an exhaustive search of the different parameters within a model to choose the combination of parameters that best scores the selected score. In this case, the scoring of f15 has been used because it is the classifier score that best suits the situation of our data since they are unbalanced[6].

$$\begin{aligned}
\text{precision} &= \frac{tp}{p + fp} = \frac{\text{retrieved and relevant documents}}{\text{all retrieved documents}} \\
\text{recall} &= \frac{tp}{p + fn} = \frac{\text{retrieved and relevant documents}}{\text{all relevant documents}} \\
F &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\end{aligned}$$

A measure such as an accuracy measure has problems with giving the same importance to a well-ranked individual from the majority class as one from the minority class. This is incorrect in a study of these characteristics, as it is required that the importance lies in a good classification of all classes and there is no bias in classifying only the majority class well, as the model would have a good accuracy but could be a poor predictor for the minority classes. In the case of the f1 score, accuracy and recall (which tells us how well the model can detect the types) are combined, which gives a much more objective value and is widely used in cases of unbalanced databases.

The models selected for this GridSearch were Logistic Regression, Nearest Neighbours, Neural Networks, Decision Trees, and Support Vector Machines. After searching the possible combinations of parameters in this GridSearch, we found the most suitable criteria to maximise the f1 score in each model.

Once the models and their parameters were obtained, we proceeded to test them by themselves, with Bagging, Boosting and Stacking. These techniques[7] combine several types of models with different strategies. In the case of Bagging, a trained model is created for one of the training

[5] https://deepai.org/machine-learning-glossary-and-terms/precision-and-recall

[6] https://medium.com/@nicolasarrioja/c%C3%B3mo-actuar-ante-el-desbalance-de-datos-a0d64f2b9619

[7] https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

samples, which produces the same number of models as training samples, and once all the predictions have been obtained, an average is made. On the other hand, Boosting consists of creating models from previous models, as the training data are the same but for each iteration, weights are assigned depending on the error, to improve their accuracy; Finally, the Stacking technique trains weak models in parallel to then create a metamodel that combines the predictions of different weak models.

After running all these models, we decided to choose based on the Brier Score[8], which focuses on magnifying how large the forecasting error was on a scale of 0 to 1. Another suggested option was to apply a cost matrix with subjective weights adapted to the problem. However, this way of the evaluation was discarded because it's an ordinal problem, so the matrix had to be symmetrical, something the team disagreed with due to the fact that some errors were more significant than others. Furthermore, it was impossible to apply the cost matrix to train the model. Therefore, it was decided to discard this form of evaluation between the difficulties encountered and the fact that it would be a subjective way of dealing with the problem without being experts in the field.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

The modes with the best Brier Score are Decision Tree, Stacking and K-Neighbours, all of them with a Brier Score of 0. We have chosen Decision Tree because it is a simpler model and can be interpreted more easily than the others. Furthermore, Decision Tree is capable of handling both numerical and categorical data and works well with large data sets. In addition, techniques such as Boosting can lead to over-fitting, especially when the amount of training data is so small.

Finally, two scores have been used to choose the parameters and models for this work (Table 4). First, the f1 score was used to calculate the parameters best adapted to this database. This score was chosen for this purpose because of its robustness to unbalanced data, as was the case. Next, the Brier Score was used for the final decision between models. It was chosen for its qualities against unbalanced and because we were dealing with an ordinal classification problem. To sum up, we chose these two metrics because they are appropriate for the problem we were dealing with and to have more information about the models and not base our decision on one metric alone as this can be risky.

5 Results

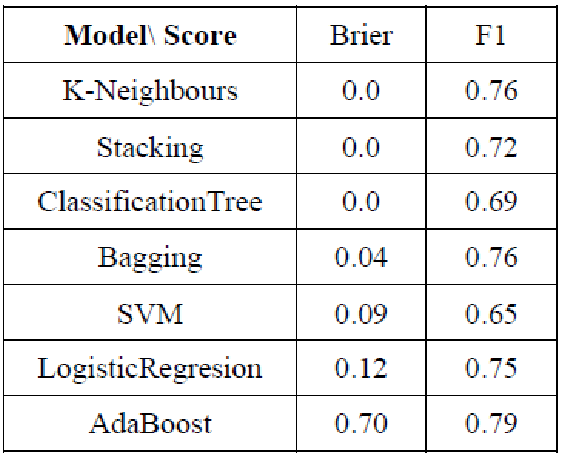
The classification tree model, the one finally selected for this work, has as its main virtue that they are accessible to understanding the classification model, which makes this method a very attractive alternative for the end-user of a knowledge extraction system: the knowledge obtained during the supervised learning process is represented by a tree in which each internal node contains a binary question about the different variables present in our database (with a child node for each possible answer) and in which each leaf refers to a decision (labelled with one of the classes of the problem).

Once the model was implemented on our data with the optimal parameters, the interpretation shown in Figure 4 was obtained.

When the PDAs present a value greater than 61.04 percent spent in the variable that indicates battery expenditure (GASTO_BAT), they will be classified with 0% battery status.

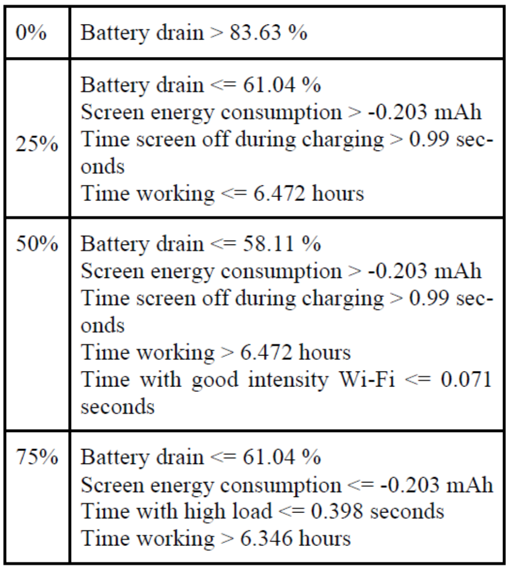
PDAs will be classified with a 25% battery status if they have a value less than or equal to 61.04 percentage spent on battery expenditure (GASTO_BAT), a value greater than -0.203 mAh in the energy consumed on-screen (PWI_mAh_scrn), a value greater than 0.99 seconds in the time that

[8] https://www.statisticshowto.com/brier-score/



Model\ Score	Brier	F1
K-Neighbours	0.0	0.76
Stacking	0.0	0.72
ClassificationTree	0.0	0.69
Bagging	0.04	0.76
SVM	0.09	0.65
LogisticRegresion	0.12	0.75
AdaBoost	0.70	0.79

Table 4: Results of the metrics used in the models.



0%	Battery drain > 83.63 %
25%	Battery drain <= 61.04 % Screen energy consumption > -0.203 mAh Time screen off during charging > 0.99 seconds Time working <= 6.472 hours
50%	Battery drain <= 58.11 % Screen energy consumption > -0.203 mAh Time screen off during charging > 0.99 seconds Time working > 6.472 hours Time with good intensity Wi-Fi <= 0.071 seconds
75%	Battery drain <= 61.04 % Screen energy consumption <= -0.203 mAh Time with high load <= 0.398 seconds Time working > 6.346 hours

Table 5: Manually provided battery status categories.

the PDA has had the screen off during the charging stage (CSD_duration_s_neg) and a value less than or equal to 6.472 hours in the hours that have worked with the PDA (HORAS_TRABAJO).

In the case of PDAs presenting a value less than or equal to 58.11 percent spent in battery expenditure (GASTO_BAT), a value greater than -0.203 mAh in the energy that has been invested in screen (PWI_mAh_scrn), a value greater than 0.99 seconds in the time that the PDA has had the screen off during the charging stage (CSD_duracion_s_neg), a value greater than 6. 472 hours in the hours they have worked with the PDA (HORAS_TRABAJO), a value less than or equal to 0.071 seconds when the PDA has had good Wi-Fi signal strength (WSGT_intensi_buena) will be rated at 50% battery status.

For a PDA to be rated at 75% battery, the PDA has to have a value less than or equal to 61.04 percent spent on battery expenditure (GASTO_BAT), a value less than or equal to -0. 203 mAh in the energy consumed on-screen (PWI_mAh_scrn), a value less than or equal to 0.398 seconds in the time the battery is low discharged (DC_baja), and a value greater than 6.34 hours in the hours worked with the PDA (HORAS_TRABAJO). A summary can be seen in Table 5.

It should be noted that when interpreting the model for each class, we have selected the leaves where the **highest number of instances were found.**

The importance of each predictor in the model is calculated as the total (normalized) reduction in the division criterion. In this case, we have based it on the entropy that the predictor achieves in the divisions in which it participates in. If a predictor has not been selected in any division, it has not been included in the model and will get a higher entropy value. A low entropy value indicates that the node contains observations of only one class.

Once the model has been interpreted, we can see that the most important variables have been the following: the variable GASTO-BAT with an entropy of 0.43, the variable HORAS_TRABAJO with an entropy of 0.15, the variable PWI_mAh_scrn with an entropy of 0.088, the variable DC_baja with an entropy of 0. 062, the variable NUEVA_USADA with an entropy of 0.06, the variable CSD_duration_s_neg with an entropy of 0.05, the variable BR_claro with an entropy of 0.045 and the variable CSD_duration_p_neg with an entropy of 0.04 and the variable WSGT_internal_buena with an entropy of 0.03.

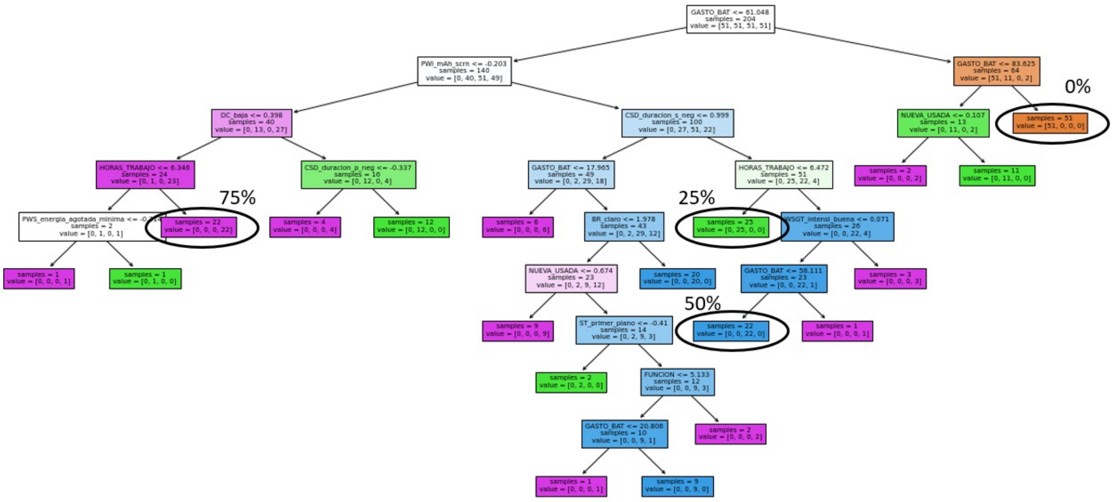


Figure 4: Decision Tree of the final mode.

6 Risk

Concerning the study of the impact of our decisions and the risks we had faced; we differentiate between internal and external threats.

6.1 Internal risks

On the one hand, internal risks were the lack of data at the beginning of the project and its subsequent long collection process, because the company did not have an historical record about the PDA's batteries. This situation limited us significantly at the time of concluding, testing different models, or making certain decisions about the disbalanced classes problem. To mitigate this risk, we collected data from personal devices trying to make progress in pre- analyzing the data in order to select the useful variables from de dumpys database and starting to change the data format from one text file per device to an csv file with all PDAs. In turn, this may lead us to the following internal risk we have gone through, this is that the model may not be the most appropriate due to the short time we have to decide an algorithm, try hyperparameters, etc. To mitigate this risk, as soon as we had a representative sample of the processed data, we started the corresponding tests to choose the model and then test it on a larger scale.

6.2 External risks

On the other hand, when discussing external risks that may arise when testing the model in Inditex warehouses. When the loading cycles, usage habits, or PDA model change and the data used to train the model become obsolete, the model will continue to predict based on old data and will not provide good results as the population would have changed. The confidence that users place in the results of our predictions may also influence, as this may lead them to overuse the devices. To solve this risk, we tried to investigate how to implement the model in the cloud and be able to keep it updated with new data. Unfortunately, due to the time limit we were not able to achieve this, but we have shared our idea with the Inditex team as a future work to get done.

Another external risk that we have detected in our project is that we had to anticipate possible changes in variables in the data. We started working with data obtained from Android mobiles, with different Android versions of the PDAs. Because of this, when we received the data from the

PDAs, the information had differences, as some variables did not appear and there were new ones. For this, we decided to anticipate and reserving a period of two weeks at the beginning assigning three people to analyse and compare it with the data we already had and understand the latest data.

7 Discussion

We have shown that the best oversampling data processing method for this project is SMOTE due to its results with a high number of f1-scores in the different models. Furthermore, the best model adapted to this project is decision trees, as demonstrated by the Brier score criterion. This sets a precedent as no studies treat PDA battery life in a classificatory way with the Android dumpsys tool. We hope this will help future work that wants to use this Android tool on new data.

8 Legacy

As for the lessons learned from this work, as a group, we have learned methodologies to put projects in order, such as CRISP-DM. Despite applying it, it has not been followed to the letter and has led to misunderstandings between the Modeling and Data Preparation processes. This is why, after this experience, we have acquired the necessary knowledge of this type of methodologies and how useful this knowledge is when carrying out data science projects that will be applied in future works.

In addition, the hard work done in the search for documentation and models has caused sklearn to limit us in testing different techniques as this library does not offer enough tools, such as the impossibility of implementing models with cost matrices as metrics or the use of these models with missing values. This will be considered in the future to discover more machine learning libraries and methods.

On the other hand, being a large group of 7 people, there were times when we were not very clear on how to divide tasks, and there were even people with no work to do. In addition, we did not have a leader figure to direct, assign tasks, and, most importantly, control where each group member was and make sure that the established deadlines were met. In our case, this responsibility was divided among three team members, which increased their workload.

Because of this, deadlines were delayed, and misunderstandings arose, so we believe we have learned our lesson and in future work, we will take into account the importance of the leader's figure. By assigning this role to one person, the work and organisation would be more effective. At the same time, this person would carry out fewer tasks related to the project, but would be in charge of organising the subdivisions of the group, controlling that everything goes according to plan and that all the members of the group know where they are in the project and what the problems are at any given moment.

In addition, we are used to receiving data that does not involve as much pre-processing work as on this occasion and what this involves. However, when carrying out a real project developed by a company, we have faced challenges and problems continuously, such as having to update some of the decisions taken due to the gradual collection of data, which has helped us to learn how to act in the future with this type of problems that are so common in the world of work. Regarding communication with the company, we believe that there are points that could be improved for future projects, but we believe that we have had good follow-up and that they have been involved in all the important decisions. We recommend that the deadlines for data delivery be defined from the beginning of the project so that the group of data scientists can better plan the tasks to be carried out and measure times.

9 Code

The code used in the realization of this project can be found in this GitHub repository: **Project-Batteries2022**.

10 Conclusions and future work

The lack of a tool that would provide the well-known Inditex company with objective information about the health of its PDA batteries was of vital importance to adjusting its budgets and material costs. Having a system to inform the company about this data would allow them to determine when the devices would no longer be valid and when they would have to invest in the material renewal. In addition, it would also allow them to avoid technical problems such as PDAs failing in the middle of a working day. We developed an application that returns the PDA's battery health percentage in response to this situation.

With this application, we can classify PDAs through a classification tree model. The device is classified depending on the range of its battery health. Thanks to this application, the company will be able to anticipate having PDAs available for replacement and maximize the use of the battery and the device.

Our model has an f1-score coefficient of 0.83. This statistic could be improved if historical data on PDA batteries were available, as it would allow us to know fundamental characteristics to enhance the model's prediction. Furthermore, we could obtain better results if the data collected included a more significant number of PDAs in poor condition. We have had access to very little data on devices with these characteristics during this project.

Thanks to our application, Inditex will be able to adjust its annual budgets. Incorporating our application into the company's operations will mean a significant reduction in the company's budget for purchasing PDAs or batteries. As for future work, the current model has been trained with a reduced sample of data due to the project's time constraints. With a massive collection of new data extracted from more PDAs, we could improve the model's effectiveness by using more representative samples with more observations. Finally, although the application created to run the extraction, cleaning, and diagnosis of the PDA works correctly, we still need to reduce the size of this software and improve its speed and efficiency when running it.

Aknowledgements

The authors are very grateful to Professor José Hernández Orallo for his valuable contribution to the project and the company Inditex, specifically to the employee Sergio Álvarez for providing the data and monitoring the study.

References

[1] Liu, K. et al., 2022. Interpretable machine learning for battery capacities prediction and coating parameters analysis,. Volume 124.

[2] Messagie, M. et al., 2015. Environmental performance of lithium batteries: life cycle analysis. pp. 303-318.

[3] Omar, N. et al., 2015. Aging and degradation of lithium-ion batteries. pp. 263-279.

[4] Zhang, Q. et al., 2022. A deep learning method for lithium-ion battery remaining useful life prediction based on sparse segment data via cloud computing system,. Volume 241.

Solving a Crime with Graph Theory

Elsa Blasco Novell$^{\flat 1}$ and Lucía López Ribera$^{\flat 2}$

($\flat$) Doble Grado en Matemáticas e Ingeniería en Tecnologías y Servicios de Telecomunicación
Escuela Técnica Superior de Ingeniería de Telecomunicación Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Could mathematicians solve a crime? This article is a proof that graphs are a useful tool for solving many problems in everyday life, apart from the field of mathematics itself. Next, we will present and develop a fictitious case of a murder and, modelling it as a graph theory problem, we will see how to solve it.

Graph theory is a branch of mathematics dedicated to the study of the properties and results related to objects known as graphs. But what is a graph? As the work is the result of an activity proposed in a first-year undergraduate level Discrete Mathematics class, only basic graph concepts and algorithms have been used (see [1], [2] and [3]).

A graph G is an ordered pair (V, E), where V is a finite (non-empty) set and E is a set of pairs of elements of V (called vertices). An undirected graph is one in which the elements of E, called edges, are no ordered pairs of V, that is, represent symmetric relationships. H is a subgraph of G if $V(H)$ is contained in $V(G)$ and $E(H)$ is contained in $E(G)$. A path is a finite alternating succession of vertices and arcs/edges in which neither vertices nor arcs/edges are repeated. If the firts and last vertex of a path are equal, we call it a cycle.

Throughout the development, we will want to see if the graphs obtained are connected or not. A graph G is called connected if any two vertices are connected by a path. One way to determine which vertices v reaches a given one u (that is, if there exists a path from u to v) would be to apply the DFS or BFS search algorithms (Depth First Reach and Breadth First Reach) algorithms. We represent this in the so-called access matrix. Having $G = (V, E)$ where $|V| = n$, the access matrix $A = [a_{ij}]$ is the matrix of size $n \times n$ whose elements are defined as:

$$a_{ij} = \begin{cases} 1 \text{ if } v_i \text{ reaches } v_j \\ 0 \text{ if } v_i \text{ does not reach } v_j \end{cases}$$

Such a matrix can be obtained from the repeated application of the DFS and BFS algorithms. Concretely, each of the rows, $i = 1, 2, ...n$ of the access matrix is obtained by applying one of the two aforementioned methods to the vertex v_i.

[1] eblanvov@etsid.upv.es
[2] llloprib@teleco.upv.es

On the other hand, we will also work with weighted graphs, that is to say graphs such that each edge/edge has an associated weight. Since the weights we will use are positive, we will apply Dijkstra's algorithm to obtain the minimum cost path from a given vertex to the rest of the vertices.

Finally, the tree concept will be used. A tree is an undirected, connected, acyclic graph. An spanning tree of an undirected connected graph G is a subgraph of G that is a tree which includes all of the vertices of G. A minimum spanning tree is a spanning tree with a weight less than or equal to the weight of every other spanning tree. After applying Kruskal's algorithm we will find such a tree.

2 Methods

As good mathematical researchers, at the beginning of the problem, we were presented with a case of murder at a party. We were given a list of clues as to which of the party guests were related to each other and a floor plan of the house showing where everyone was at the time of the event, as well as the distances between rooms.To solve the crime, we had to eliminate suspects by answering a series of questions.

We first modelled the problem by constructing two graphs.The first graph represented the relationships between the guests and matched a vertex to each of them according to their first initial. An edge existed between two vertices if the people to whom those vertices corresponded knew each other. The other graph represented the floor plan of the house, with the vertices as the different rooms.An edge existed between two vertices if access could be gained from each of the rooms corresponding to the vertices to the other and the edges were weighted with the distance in metres between the two rooms. Both were undirected graphs because friendship relationships are considered reciprocal and a corridor can be used both to go to and from a room.

We then discarded the guests who were not related to the murdered person, using the concepts of connection in graphs, connected components, access matrix and the DFS Algorithm. Applying the DFS algorithm to the friendship graph, we obtained the connected components of that graph. We discarded vertices that did not belong to the connected component of the murdered person's vertex (in a green circle in Figure 1), as this is evidence that they did not know each other and they could not commit the crime. We were still working with 7 suspects.

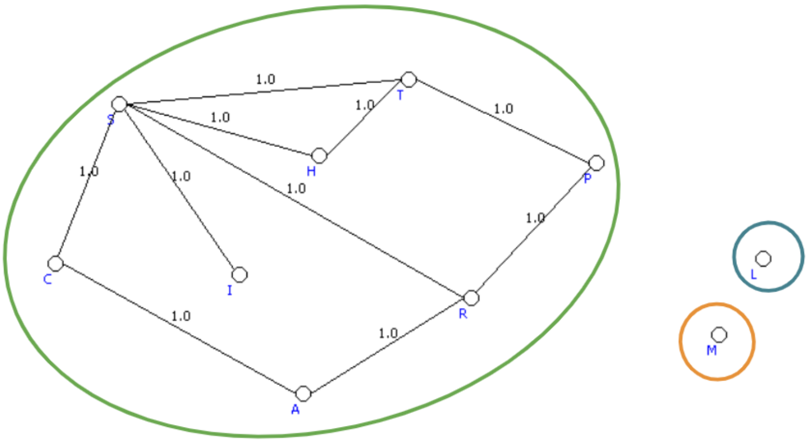


Figure 1: Result of the exercise.

Following a confession by one of those present at the crime scene, it was discovered that some room doors had been locked before the murder took place. As some rooms had been inaccessible after this closure, it is logical to try to rule out those people who had not been able to access the room where the crime was committed. Thus, we constructed a new graph by varying the initial one that modelled the rooms of the house, eliminating the edges corresponding to the closing of the doors.

We used the concepts of accessibility in graphs and the BFS Algorithm to locate the rooms that had been isolated and then eliminate the corresponding suspects. By applying the BFS algorithm we located the rooms that were accessible from the murder site (as we can see in Figure 2). We focused our attention on the rooms that were cut off from that access and ruled out the people who were in the blocked rooms. We were still working with 5 suspects.

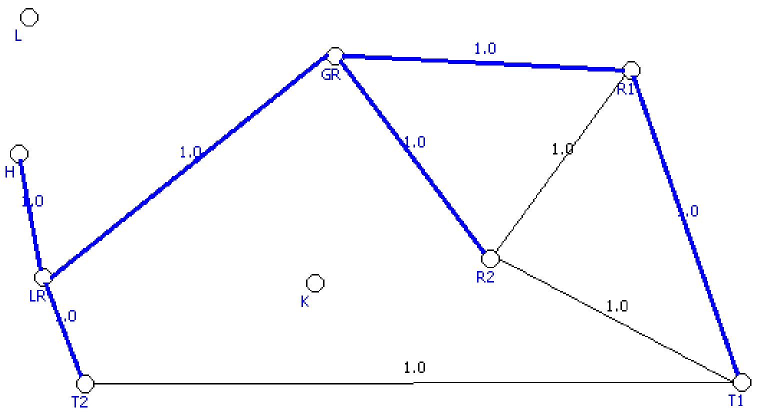
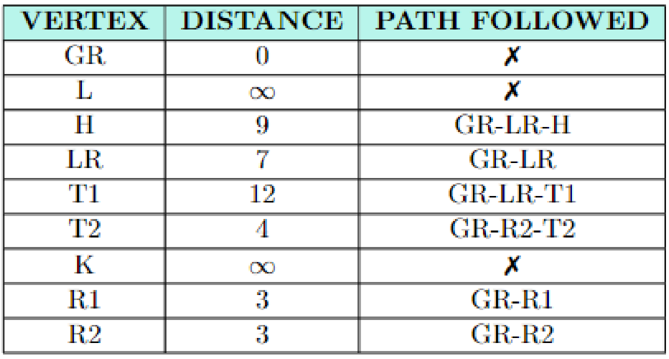


Figure 2: Result of the exercise. Source: SWGraphs.

As a third and final question for solving the crime, we posed that the culprits had to have done it in the shortest possible time and assumed that the time is taken to move from one place to another depends on the distance travelled. Thus, we looked for the shortest path from the murder room to all the other rooms and from there we took the smallest value.

To do this, we used Dijkstra's Algorithm on the graph corresponding to the house with the doors closed. By applying the algorithm we obtained the results shown in the table of Figure 3, from which we deduced that the culprits were in room 1 or room 2. By checking the floor plan we found that room 2 was empty, so we established the two guests in room 1 as the culprits.



VERTEX	DISTANCE	PATH FOLLOWED
GR	0	✗
L	∞	✗
H	9	GR-LR-H
LR	7	GR-LR
T1	12	GR-LR-T1
T2	4	GR-R2-T2
K	∞	✗
R1	3	GR-R1
R2	3	GR-R2

Figure 3: Table compiled from data provided by SWGraphs.

Finally, we were asked to provide the optimal installation of security cameras in the house to prevent a similar situation from occurring again, taking into account that, as conditions of the owners and the installation company: at least one door in each room should be covered by the cameras, that the installation should be uninterrupted, and that it was intended to be as cost-effective as possible (the cost depended on the metres of corridor covered by the installation itself).

Thus, we modelled the problem using the graph corresponding to the initial house plan (without closing doors) and weighted with the distances between the different rooms.

The conditions provided corresponded to a graph that, starting from the aforementioned, has all the vertices, does not contain cycles (this would be an unnecessary route for the installation), is connected and whose sum of the weights of its edges is the minimum possible. These characteristics correspond to those of a minimum spanning tree, which is obtained by applying Kruskal's algorithm to the graph.

We used this lgorithm, which gave us the optimal route to follow for the installation (showed in the Figure 4, marked in blue).

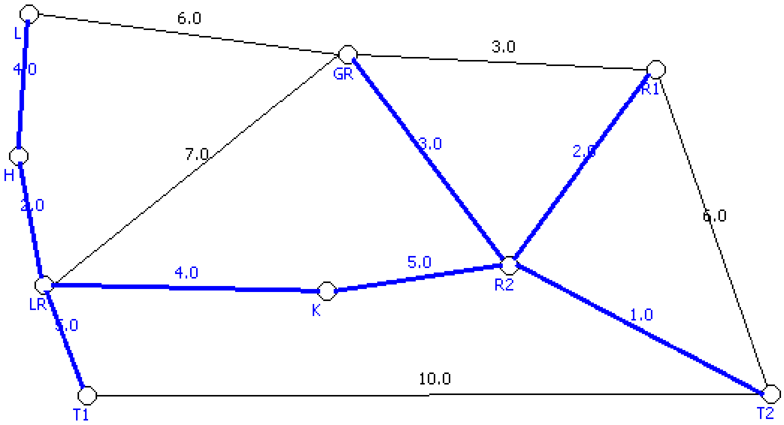


Figure 4: Result of the exercise. Source: SWGraphs.

3 Results

As a result of the problem we obtained that the murder had been committed by two people, who knew the victim, were not affected by the closing of doors, and were able to do it in less time than the rest of the guests to avoid being seen, so they were not eliminated in any of the discarding processes based on the knowledge of graph theory.

In addition, we obtained a way to solve a problem such as the need to install a security camera installation and want to spend as little money as possible by using concepts such as spanning trees or Kruskal's Algorithm.

4 Conclusions

The main purpose of the elaboration and presentation of this project is to show a non-mathematical a priori context and to show how it can be solved through Graph Theory. In this way, it has been demonstrated once again how mathematics is present in our everyday life, and the conception that it only deals with abstract terms is put aside. It has been shown that graph theory is an area of mathematics that is particularly well suited to modelling real problems, as it is a useful tool for tackling a wide range of problems in diverse areas.

Although we have presented a very particular and fairly simple case of what could have been a crime, we have shown that within the field of criminology and criminalistics we can make use of graphs to mathematically model a data structure. Thus, we have been able to see the relationships that exist between these two disciplines, normally considered to be very distant both in content and in their working techniques.

References

[1] J. Gross, J. Yellen, Graph theory and its applications *CRC, 1999*,

[2] C. Jordán, J.R. Torregrosa. Introducción a la teoría de grafos y sus algoritmos. Servicio publicaciones UPV, Valencia, 1996

[3] C. Jordán, OCW Estructuras Matemáticas para la Informática 2, 2010.

55

Harry Potter and the relics of the graphs

D. Vañó Fernández[1] and R. Fornas Sáez[2]

Students of Mathematics Degree
Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

In this project we have created different problems related to graph theory and the Harry Potter saga of JK. Rowling.

The objective is to link fantasy and mathematics through a problem and to be able to bring the students closer to graph theory. For that reason we have decided to create an immersive exercise where we will send letters to the pupils so that they can answer them and solve the different statements.

To solve them, it is necessary to model and apply algorithms that we have learned in the first year of the Double Degree in Mathematics and Civil Engineering. The structure of this summary will be divided into two parts: The first one will consist of a different statements where the student must apply their knowledge in order to find the solution. The second part deals with the resolution where the algorithm or model to be used will be explained in detail.

2 Methods

2.1 Statements

1.- Problem

We urgently need you to send us the school plan in order to draw strategic attacks. It is necessary for you to translate the map into the language of the SWGraphs spell, an ancient language known as graphs. We know that you are one of the greatest professionals in this area so, you will not have any problems.You should take into account that for now on each road we can go both ways. Send us the map in .xml format. There we have the original map.

[1] dovaofer@cam.upv.es

[2] rforsae@etsiamn.upv.es

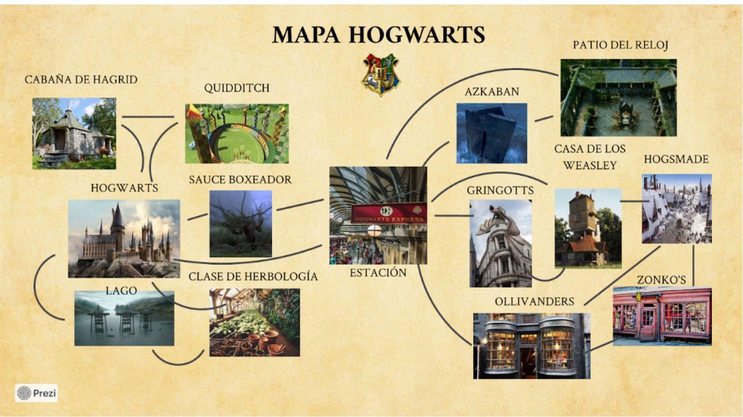
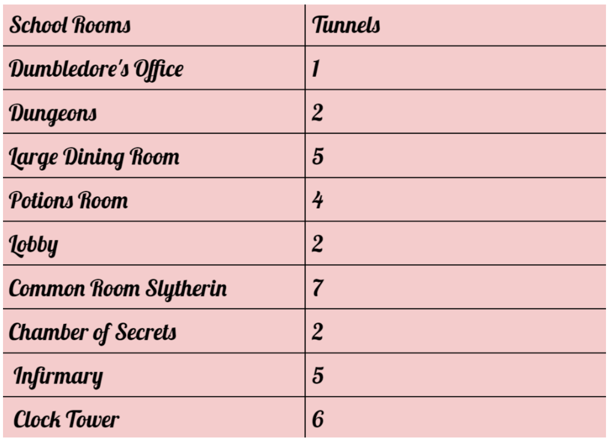


Figure 1: Map of the School made with Prezi

2.- Problem

We have to protect the students so we're going to take them to the shelters. They are very old shelters and the only information we have is the number of tunnels that are connected to each room and they are round-trip.Besides, each tunnel begins in one room and ends in a different one (see Figure 2). We send you the information on another parchment. We need to know if the tunnel network is applicable to SWGraphs.



School Rooms	Tunnels
Dumbledore's Office	1
Dungeons	2
Large Dining Room	5
Potions Room	4
Lobby	2
Common Room Slytherin	7
Chamber of Secrets	2
Infirmary	5
Clock Tower	6

Figure 2: Chart with the rooms and the number tunnels

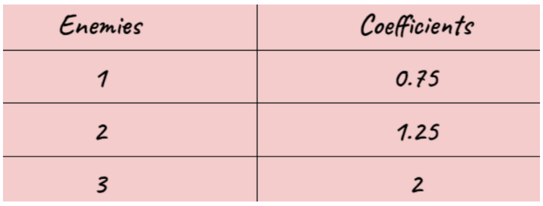
3.- Problem

a) We have observed that the enemy is preparing to launch some attack. We want a map of the school where all the buildings are connected to each other and all the original sections of the map appear, but you can pass through them with a single direction due to the fact that Voldemort has cast a spell which forces people to walk in one direction along the roads. b) There are some injured or endangered students so the objective is to send a squad of animagus to help them.We need you to give us a graph where you can reach all places from the station. We don't need them to go back to the station as the principal is planning a plan to pick up the students.

4.- Problem

We are going to evacuate the school so that all the students can return home safely. Dumbledore's plan is:A train pass through all areas of Hogwarts and return to it with all the students to depart from Hogwarts to the city. Using a magical spell, the principal has ordered all students and staff to take to the streets to speed up the collection process. Our objective is to carry out the principal's plan by making the path with the least possible weight. We have assigned a coefficient to each section and we have multiplied it by the distances of the paths (see it on the Figure 3). Voldemort's army has cast two spells of Petrificus Totalus on the Station-Ollivanders and Weasley's house-Hogsmade sections.

The coefficients that we have applied based on the number of enemies have been the following:



Enemies	Coefficients
1	0.75
2	1.25
3	2

Figure 3: Chart with the coefficients

5.- Problem

Our objective is to rebuild the school after the enemy attack. They want to reform only the necessary roads, thus being able to spend as little as possible. All buildings must be connected to each other. What is the best way to act?

2.2 Results

1.- Problem

First we model the problem:

G is the set formed by the sets of vertex (V) and edges (E)

This is an undirected graph (UG)

V represents the different places in the school.

E=((u,v)/u,v ∈ V and u is connected to v by a path with no intermediate vertex.)

Graphical representation with SWGraphs:

2.- Problem

To see if it is applicable or not, we have to know if our graph is a graphical sequence.

We model the problem:

G is the set formed by the sets of vertex (V) and edges (E). In this case, we have an undirected graph.

V represents the different places in the school.

E= (u,v) / u,v ∈ V u and v are connected by a tunnel

HAKIMI

A → Slytherin Common Room

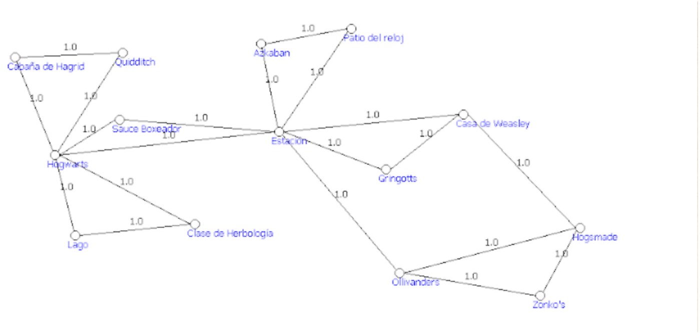
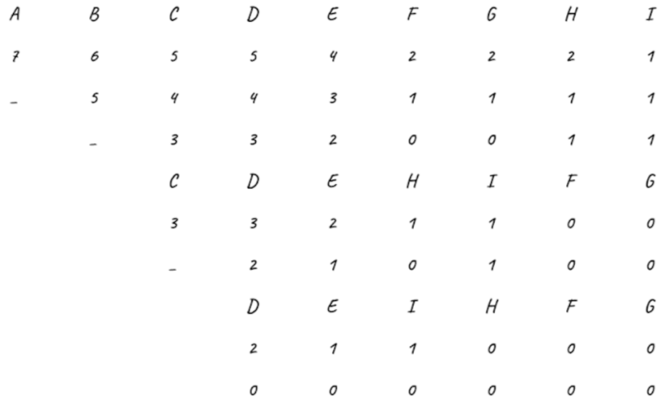


Figure 4: Graphical representation of the graph

B → Clock Tower
C → infirmary
D → Large dining room
E → Potions room
F → Dungeons
G → Chamber of Secrets
H → Lobby
I → Dumbledore's Office



A	B	C	D	E	F	G	H	I
7	6	5	5	4	2	2	2	1
-	5	4	4	3	1	1	1	1
	-	3	3	2	0	0	1	1
		C	D	E	H	I	F	G
		3	3	2	1	1	0	0
		-	2	1	0	1	0	0
			D	E	I	H	F	G
			2	1	1	0	0	0
			0	0	0	0	0	0

Figure 5: Resolution with Hakimi's algorithm

The number of vertex of odd degree is even. There are no vertex of degree greater than or equal to the number of vertex.
It is a graphical sequence. Therefore, it can be represented as a simple undirected G. The solution provided by Hakimi is not unique. As a consequence, we are going to apply a spell to the school so it adapts (moving stairs and rooms) to our new plane. Then, the network of tunnels exists.

3.- Problem

a) First we see if our graph is directed. We apply the Robbin's Theorem.We note that my graph is connected and has no bridges, since every edge belongs to a cycle. As it is orientable we will apply the Hopcroft-Tarjan algorithm.

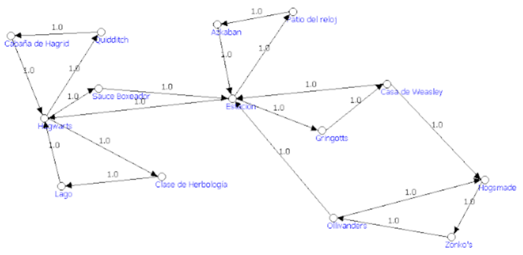


Figure 6: Representation of the oriented graph

b) We want a directed tree with root where the root is the Station so we apply DFS and BFS algorithm. Each algorithm will give a directed tree with a different root.

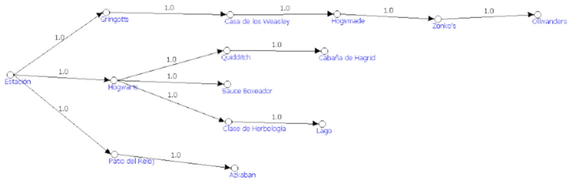


Figure 7: Resolution with BFS algorithm

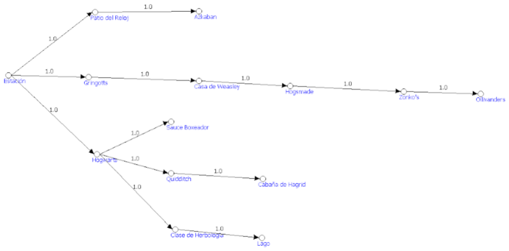


Figure 8: Resolution with DFS algorithm

4.- Problem

G is the set formed by the sets of vertex (V) and edges (E), This is a directed graph (DG) V=Hagrid's Hut, Quidditch, Hogwarts, Boxer Willow, Lake, Herbology Class, Station, Azkaban, Clock Yard, Gringotts, Weasley's House, Hogsmade, Zonko's, Ollivanders E=(u,v)/u,v ∈ V and u is connected to v by a path. p(x,y)=degree of enemies by distance. We attach the map with the modified weights:

Next, we are told that there are paths that are cut off by the spells. If we eliminate these paths, that is, the indicated edges, we obtain a subgraph G'=(V',E'),G' is the set formed by the sets of vertex (V') and edges (E'), to which we can apply the desired algorithm. G' is the subgraph induced by the following vertices: V'=Hagrid's Hut, Quidditch, Hogwarts, Boxer Willow, Lake, Herbology Class, Station, Azkaban, Clock Yard, Gringotts, Weasley's House

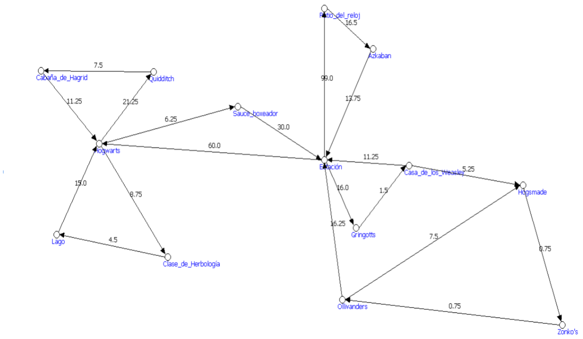


Figure 9: New subgraph with eliminated paths

The algorithm to apply is Hierholzer's algorithm and the weights constitute irrelevant information

RESULT: Station → Hogwarts→ Herbology Class → Lake → Hogsmade → Quidditch → Hagrid's House → Hogsmade → Boxer Willow → Station → Gringotts → Weasley's House → Station → Azkaban → Clock Yard → Station

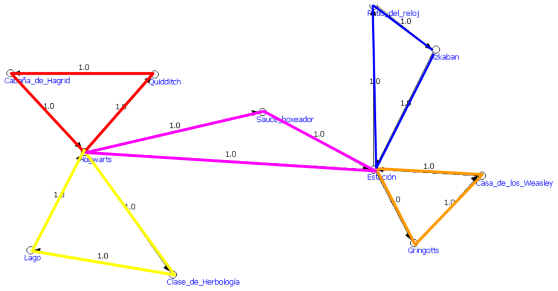


Figure 10: Result of applying Hierholzer's algorithm

5.- Problem

The aim is to reduce the expenses of the work, we seek to rebuild part of the roads so that you can go from anywhere to any other.

To do this, we apply Kruskal's algorithm. We apply this algorithm to the initial plan of the school, since, as we have been informed, everything has been destroyed. We have obtained the minimum cost spanning tree.

3 Conclusions

To sum up, mathematics is linked to all areas. The project we have presented is a clear example of this. We can solve different types of problems using Graph Theory. However, developing statements to solve with this method is not as simple as it seems, since normally a series of requirements have to be met in the graphs for the algorithms to be applicable. The fact of inventing a context that gives meaning to the exercise is important because that can help you modify the problem and give

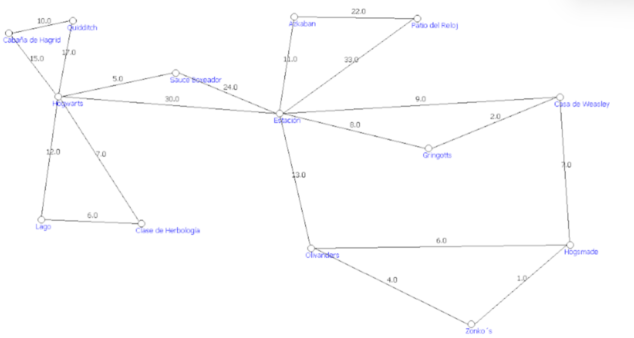


Figure 11: Initial plan of the school

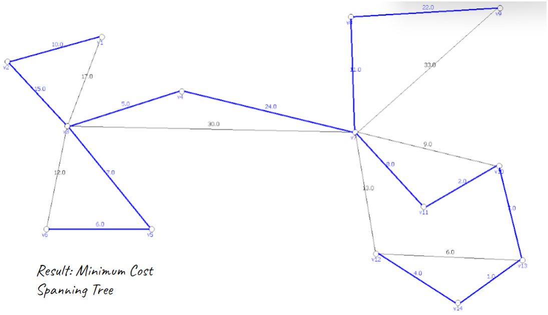


Figure 12: Result of applying Kruskal's algorithm

it meaning. Finally, we can say that in every corner of Hogwarts there is a little bit of magic, a little bit of mathematics.

References

[1] J.A Conejero, C. Jordan., Curso MOOC de aplicaciones de la teoría de grafos a la vida real.

[2] C. Jordán, A.,Catálogo de vídeos de matemática discreta.

[3] K. H. Rosen.,Discrete Mathematics and Its Applications Seventh Editions.

[4] Susanna Epp. Matemáticas discretas con aplicaciones

[5] Félix García Merayo. Matemática discreta

[6] F. García Merayo, G. Hernández Peñalver, A. Nevot Luna.,Problemas resueltos de matemática discreta. 2ª edición ampliada

[7] SWGraphs program, UPV software.

[8] J.K.Rowling, Harry Potter novels. Barcelona, Ediciones Salamandra, 1997-2022.

[9] J.K.Rowling, Heyday Films, Warner Bros,Harry Potter films.

[10] Ádám Somlai-Fischer, Péter Halácsy y Peter Arvai, Images from Prezi.

56

Evaluating the sudden change in flight cancellations in Paris during November 2015

D. Romero[♭], C. Gallego[♭], R. Gironés[♭], J. Grau[♭], L. Lillo[♭] and A. Losa[♭],[1]

(♭) Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

The air travel industry is a really volatile market, where sudden events can lead to unexpected results in flight transactions. Companies of the sector must be ahead of the curve in predicting and understanding the impact of these events.

ForwardKeys, a data collection and analysis company, has provided us data about flight transactions in Paris during November 2014 and 2015. In this paper we aim to study the data in depth so that we can obtain detailed information on cancellations, using association rules to discover the main repeating patterns among the data, to better explain and outline the clients profile of the market.

Furthermore, to better understand what defines the profile of a client, a linear discriminant analysis and a decision tree are performed to determine importance to attributes of the airline tickets

Our first action was to look for what happened in Paris in November 2015. As expected, we found an event that considerably changed air traffic in Paris. On November 13, 2015, the terrorists attacks that shocked the city of Paris took place. The impact of such events on the air transport sector is not a widely studied topic, although some studies show that most of the costs in these situations are cancelled and delayed flights. [4]

2 Methods

In order to achieve our main objectives (finding patterns in the data and better understanding our data-set to study cancellations), a variety of machine learning models can be used. In this case, three approaches will be utilised: an exploratory analysis, an unsupervised model and a supervised model.

2.1 Finding patterns among the clients: association rules [2]

One of the most useful things that an exploratory analysis can show is which combinations of conditions lead to more cancellations or what are the characteristics of people that cancel the most

[1] dromalv@inf.upv.es; cgaland@inf.upv.es; rgirsan@inf.upv.es; jvgragil@inf.upv.es; llilcol@inf.upv.es; alosbri@inf.upv.es

(given a terrorist attack has happened). This could be extrapolated to other situations of fear and uncertainty. Moreover, the information is going to be extremely helpful to travel retailers, hotels, and tourism-related companies.

Given our dataset, we will be focusing on the following variables: bookingsign (Type of transaction: new booking, modification of the booking or a cancellation), paxprofile (Estimated client profile by ForwardKeys' classification algorithm), losname (Type of stay or transfer from the flight), distchannel (Type of travel agency) and cabinclass (Code of the cabin class type).

2.2 Understanding the client profile: linear discriminant analysis for paxprofile

Linear discriminant analysis is a generalisation of Fisher's linear discriminant, a statistical method that tries to find a linear combination of predictor variables to describe the differences between the different classes of a categorical target variable. This method can be later used to reduce the dimensionality of the data or as a linear classifier [3].

As previously mentioned, the client profile is a really interesting and important variable to explore and fully comprehend. According to ForwardKeys, this variable is "the purpose of the trip according to the company's classification algorithm". This classification decision is unknown to us, and thus, considered as a black box model. An interpretable model that can explain the decisions the classification algorithm makes when categorising was trained in order to help companies reach their target and fully comprehend passenger preferences. We opted for a linear discriminant analysis model to predict the variable paxprofile.

Taking these limitations into consideration, our model uses eight variables, continuous and categorical that have been converted into dummy variables: leadtime (Difference in days between the date of booking and the date of the flight), cabinclass, distchannel, pax (difference in passengers from last modification of the reservation), numpss (number of previous steps in the itinerary), numpss (number of next steps in the itinerary) and losname.

2.3 Other interpretable model: recursive partitioning trees

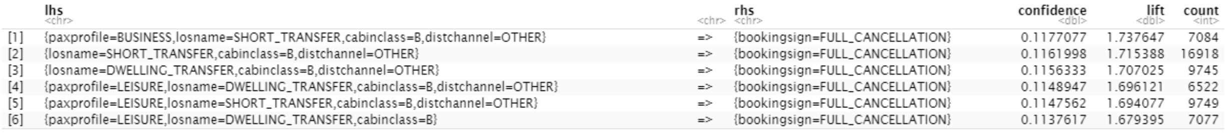
Decision trees are another easy-to-interpret model. These models divide the data according to splits in the input features so that they maximise the decrement in the impurity of the resulting children nodes [6]. The purity is the homogeneity of the target variable in the node.

In order to compare this model with the LDA one, we will be using the same dataset (with undersampling to take into account class imbalance) and the same seed for the train-test partition. Since the data is the same, and it is still unbalanced, the metrics used to evaluate the effectiveness of the model are the same as in LDA: the Kappa Coefficient and the weighted balanced accuracy.

3 Results

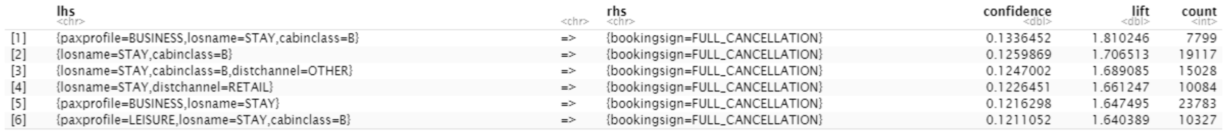
3.1 Association Rules

Rules are valued by lift as it is the most important metric for our analysis. A value 1 means that the condition (left hand side) helps the rhs happen, the higher, the better.
In 2015 people that went to Paris to stay had the highest chance of cancelling as it appears in every rule, and because this condition is nowhere to be seen in 2014, we could argue that these people are the first to cancel for fear.



	lhs		rhs	confidence	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<int>
[1]	{paxprofile=BUSINESS,losname=SHORT_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1177077	1.737647	7084
[2]	{losname=SHORT_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1161998	1.715388	16918
[3]	{losname=DWELLING_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1156333	1.707025	9745
[4]	{paxprofile=LEISURE,losname=DWELLING_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1148947	1.696121	6522
[5]	{paxprofile=LEISURE,losname=SHORT_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1147562	1.694077	9749
[6]	{paxprofile=LEISURE,losname=DWELLING_TRANSFER,cabinclass=B}	=>	{bookingsign=FULL_CANCELLATION}	0.1137617	1.679395	7077

Figure 1: Top 6 rules which consequent is a full cancellation sorted by lift (2014). Notice how common cabinclass B and distchannel OTHER are.



	lhs		rhs	confidence	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<int>
[1]	{paxprofile=BUSINESS,losname=STAY,cabinclass=B}	=>	{bookingsign=FULL_CANCELLATION}	0.1336452	1.810246	7799
[2]	{losname=STAY,cabinclass=B}	=>	{bookingsign=FULL_CANCELLATION}	0.1259869	1.706513	19117
[3]	{losname=STAY,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1247002	1.689085	15028
[4]	{losname=STAY,distchannel=RETAIL}	=>	{bookingsign=FULL_CANCELLATION}	0.1226451	1.661247	10084
[5]	{paxprofile=BUSINESS,losname=STAY}	=>	{bookingsign=FULL_CANCELLATION}	0.1216298	1.647495	23783
[6]	{paxprofile=LEISURE,losname=STAY,cabinclass=B}	=>	{bookingsign=FULL_CANCELLATION}	0.1211052	1.640389	10327

Figure 2: Top 6 rules which consequent is a full cancellation sorted by lift (2015). losname STAY is one of the strongest conditions, in addition, it is nowhere to be seen in 2014. This means these passengers are really sensitive to sudden events of fearful nature.

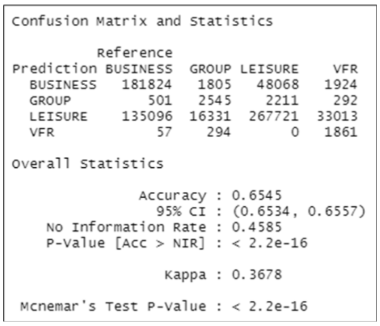
Cabinclass B, “Business”, is another condition that people who cancel usually have. But because it appears on both years we dont know how impactful the terrorist attack was for them. Perhaps they just cancel more often.

The distchannel = “other” condition (travel agencies different from online, retail or corporate) appears in more rules in 2014 than in 2015, thus arguing that people that take these flights are not really fearful about the attack, as its importance on cancellations during 2015 has at least not kept up with the other conditions.

A “long transfer” length of stay is a strong condition for not cancelling on both years, same goes for distchannel = “other” and cabinclass “T” and “E” (Tourist/Economy).

3.2 Linear Discriminant Analysis

Figure 3 shows the confusion matrix and some performance metrics on the test set for the final model, that uses numerical and categorical variables and on which undersampling was performed to reduce class imbalance:



```
Confusion Matrix and Statistics

          Reference
Prediction BUSINESS  GROUP LEISURE    VFR
  BUSINESS   181824   1805   48068   1924
  GROUP         501   2545    2211    292
  LEISURE    135096  16331  267721  33013
  VFR            57    294       0   1861

Overall Statistics

               Accuracy : 0.6545
                 95% CI : (0.6534, 0.6557)
    No Information Rate : 0.4585
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3678

 Mcnemar's Test P-Value : < 2.2e-16
```

Figure 3: Confusion matrix in the set dataset of linear discriminant model. Since the target variable still presents imbalance, the accuracy is not relevant to evaluate the model. The Kappa Coefficient is pretty low.

The model still has a tendency to classify most samples as LEISURE. The Kappa Coefficient is not good either: sitting below the 0.4 mark, the model does not distinguish well enough between classes.

Regarding which variables are more important when building the discriminant dimensions, Figures 4 and 5 show that some variables such as the number of passengers of having a cabin class code F or

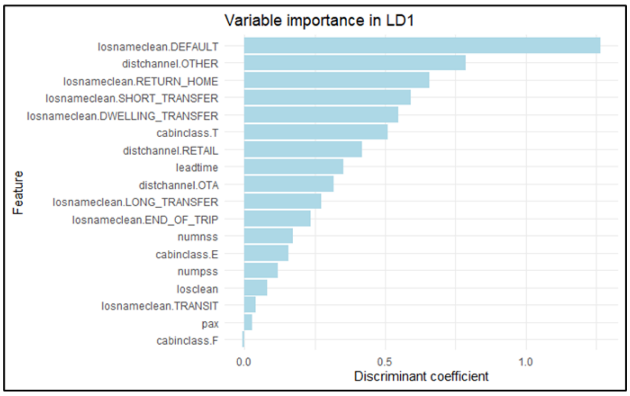


Figure 4: Feature importance in the first discriminatory dimensions (that explains 73% of the data variability). Different categories of the type of stay are very important

E do not contribute that much to any dimension, whereas staying at the destination (losnameclean "default") has a great contribution in the two main dimensions.

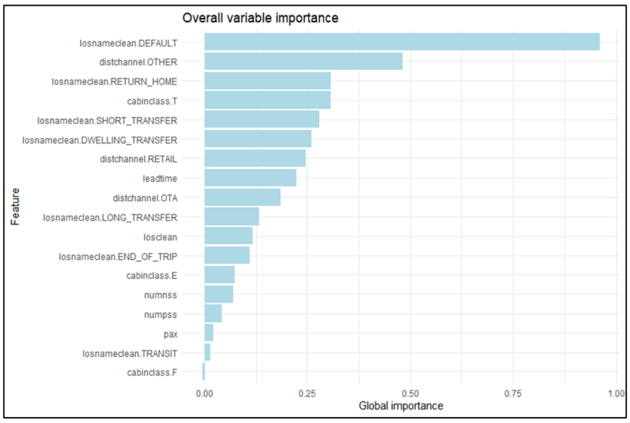


Figure 5: Overall feature importance weighted by the % of variability explained per dimension. The same tendencies as in the first discriminant dimension (Figure 1) can be seen in the global behaviour

If we take a look class-wise, Figure 6 shows similar results: in the groups VFR and GROUP, the length of stay is the most important factor, while in the most common groups (LEISURE and BUSINESS), this attribute is not so important.

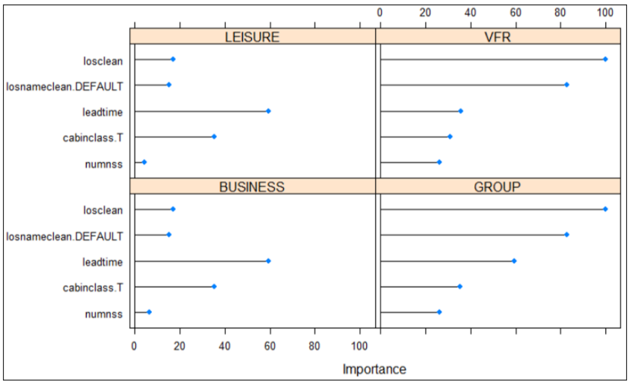
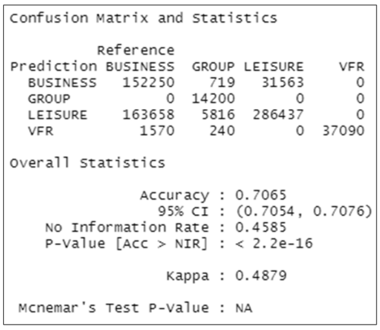


Figure 6: Top 5 most important variables for each class in paxprofile. As Table 4, losnameclean DEFAULT is a very decisive attribute to decide the profile of a passenger

3.3 Recursive partitioning tree

Figure 7 shows the confusion matrix and some performance metrics on the test set (which is the same as in the LDA model, and hence, with undersampling) for the final tree model:
The use of a recursive partitioning tree shows a small enhancement on the performance, improving the Kappa Coefficient by 0.12 and the weighted balanced accuracy by 0.05. Still, these values are



```
Confusion Matrix and Statistics

          Reference
Prediction BUSINESS  GROUP LEISURE    VFR
  BUSINESS   152250    719   31563      0
  GROUP           0  14200       0      0
  LEISURE    163658   5816  286437      0
  VFR          1570    240       0  37090

Overall Statistics

               Accuracy : 0.7065
                 95% CI : (0.7054, 0.7076)
    No Information Rate : 0.4585
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4879

 Mcnemar's Test P-Value : NA
```

Figure 7: Confusion matrix in the set dataset of rtree model. Although slightly better than in the LDA model, the Kappa coefficient is still below 0.5. The influence of LEISURE is clearly shown in its prevalence.

pretty low, and the prevalence of the class LEISURE is still significantly greater than the prevalence of the other classes.

Class-wise, the classifier performs greatly in distinguishing the minor classes from the most common ones. However, it continues to struggle differentiating between the main two ones BUSINESS and LEISURE.

When comparing ROC curves of the two models, Figure 8 shows that, in fact, LDA performs better in the BUSINESS/LEISURE portion. The tree excels at every other pairing (Figure 9).

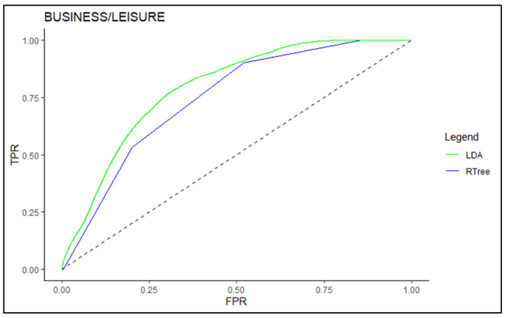


Figure 8: ROC curves comparison between BUSINESS vs LEISURE, The LDA model performs slightly better

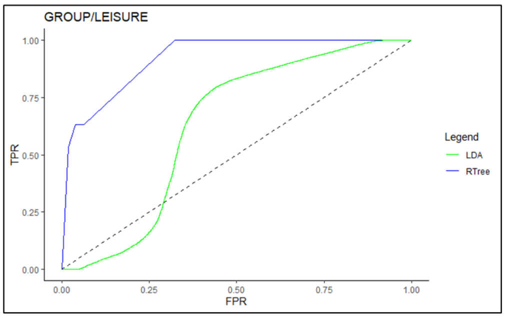


Figure 9: ROC curves comparison between GROUP vs LEISURE, The tree model is much better in distinguishing these classes

4 Conclusions

4.1 Association rules

After comparing the rule-set for each year, the team has decided that clients with a business profile and staying in the city are the strongest conditions for a cancellation in November of 2015 (which are not for November of 2014). Figure 10 shows the evolution of relative cancellations throughout the month. The cabin class code B is not taken into account because it was also a common condition in 2014. When it comes to the day of the attack, we can see a huge peak. About 35% of that day's transactions were cancellations, when usually it stays around 10%.
Also because we remove the cabinclass condition, there are more samples and the conclusions are more robust. People willing to STAY with a BUSINESS paxprofile are the most likely to cancel due to a surprising/fearful event.

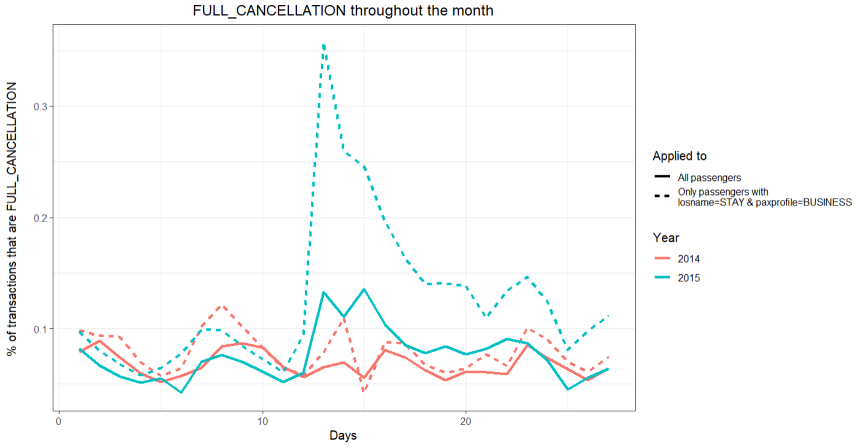


Figure 10: Relative % of cancellations per day for both 2014 and 2015 with and without the losname = STAY and paxprofile = BUSINESS. Notice the strength at the peak and the following days, whereas 2014 is not altered as much.

4.2 Linear discriminant analysis and partitioning trees

Despite not fully following all the assumptions and conditions to perform an ideal linear discriminant analysis, and possibly due to the amount of the data provided by Forwardkeys, our LDA model with 3 discriminant dimensions gives a weighted balanced accuracy of around 68% of the clients' profile from the test set, whereas a recursive partitioning tree improves this metric by 5%. Despite the improvements, both models fail to reach a good Kappa coefficient (both below 0.5), not providing a good model to understand ForwardKeys' black box model.

In both models the most important attributes to determine the client's profile are type of stay, the cabin class code and the type of travel agency.

Since the main objective of these models was trying to provide interpretable and contrastive explanations to a profile classification done by an unknown model, rather than precisely predicting said profile (therefore giving really good results in metrics such as the Kappa Coefficient and the weighted balanced accuracy score), we can conclude that we may be able to have a loose intuition of what attributes are more determinant to assign a client a specific profile. One thing we can be sure of is that more unprovided variables could help construct a more robust and complete model.

References

[1] B. Jiang, C. Leng, C. Wang, Z. Yang., Linear Discriminant Analysis with High-dimensional Mixed Variables. *arXiv:2112.07145*, Vol.2, 2022.

[2] S. Kotsiantis, D. Kanellopoulos., Association Rules Mining: A Recent Overview. *GESTS International Transactions on Computer Science and Engineering*, Vol.32 (1), pp. 71-82, 2006.

[3] A. J. Izenman., Linear Discriminant Analysis. In: Modern Multivariate Statistical Techniques. *Springer Texts in Statistics*, Springer, New York, NY, pp 237–280, 2013.

[4] M. Janić., Reprint of Modelling the resilience, friability and cost of an air transport network affected by a large-scale disruptive event. *ScienceDirect*, Vol.81, pp 77-92, 2015.

[5] W.J. Krzanowski, Mixtures of continuous and categorical variables in Discriminant Analysis. *JSTOR*, pp 493-499, 1980.

[6] C. Strobl, J. Malley, G. Tutz. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, Vol.14(4), pp 323–348, 2009.