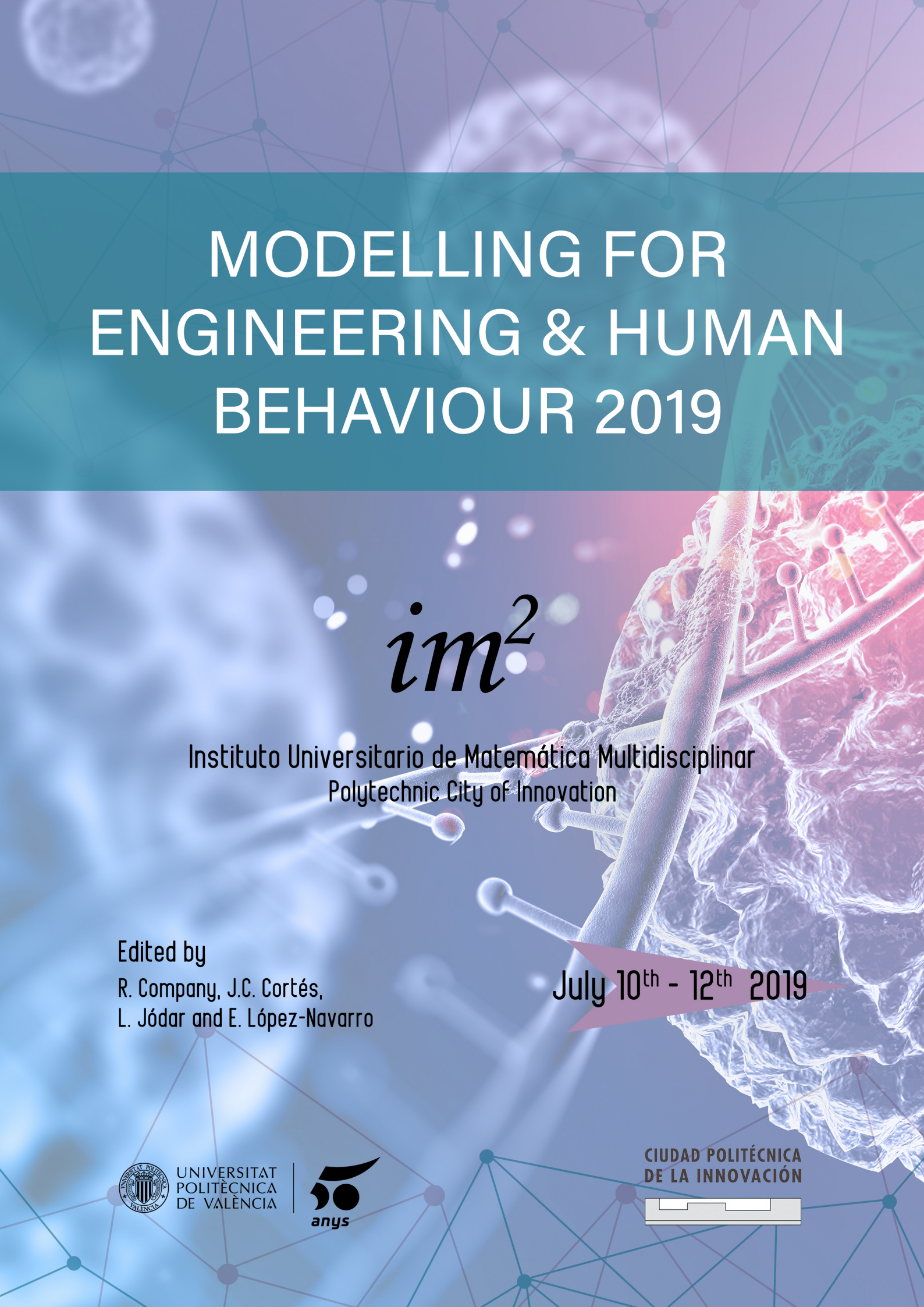
MODELLING FOR ENGINEERING & HUMAN BEHAVIOUR 2019



im^2

Instituto Universitario de Matemática Multidisciplinar
Polytechnic City of Innovation

Edited by
R. Company, J.C. Cortés,
L. Jódar and E. López-Navarro

July 10th - 12th 2019

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
50 anys
CIUDAD POLITÉCNICA DE LA INNOVACIÓN

Modelling for Engineering & Human Behaviour 2019

València, 10 – 12 July 2019

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

I.S.B.N.: 978-84-09-16428-8
Version: 15/11/19
Report any problems with this document to ellona1@upvnet.upv.es.

Edited by: R. Company, J. C. Cortés, L. Jódar and E. López-Navarro.
Credits: The cover has been designed using images from kjpargeter/freepik.

Instituto Universitario de Matemática Multidisciplinar

This book has been supported by the European Union through the Operational Program of the [European Regional Development Fund (ERDF) / European Social Fund (ESF)] of the Valencian Community 2014-2020. [Record: GJIDI/2018/A/010].

Contents

A personality mathematical model of placebo with or without deception: an application of the Self-Regulation Therapy . . . 1

The role of police deterrence in urban burglary prevention: a new mathematical approach . . . 9

A Heuristic optimization approach to solve berth allocation problem . . . 14

Improving the efficiency of orbit determination processes . . . 18

A new three-steps iterative method for solving nonlinear systems . . . 22

Adaptive modal methods to integrate the neutron diffusion equation . . . 26

Numerical integral transform methods for random hyperbolic models . . . 32

Nonstandard finite difference schemes for coupled delay differential models . . . 37

Semilocal convergence for new Chebyshev-type iterative methods . . . 42

Mathematical modeling of Myocardial Infarction . . . 46

Symmetry relations between dynamical planes . . . 51

Econometric methodology applied to financial systems . . . 56

New matrix series expansions for the matrix cosine approximation . . . 64

Modeling the political corruption in Spain . . . 70

Exponential time differencing schemes for pricing American option under the Heston model . . . 75

Chromium layer thickness forecast in hard chromium plating process using gradient boosted regression trees: a case study . . . 79

Design and convergence of new iterative methods with memory for solving nonlinear problems . . . 83

Study of the influence falling friction on the wheel/rail contact in railway dynamics . . . 88

Extension of the modal superposition method for general damping applied in railway dynamics . . . 94

Predicting healthcare cost of diabetes using machine learning models . . . 99

Sampling of pairwise comparisons in decision-making . . . 105

A multi-objective and multi-criteria approach for district metered area design: water operation and quality analysis . . . 110

Updating the OSPF routing protocol for communication networks by optimal decision-making over the k-shortest path algorithm . . . 118

Optimal placement of quality sensors in water distribution systems . . . 124

Mapping musical notes to socio-political events . . . 131

Comparison between DKGA optimization algorithm and Grammar Swarm surrogated model applied to CEC2005 optimization benchmark . . . 136

The quantum brain model . . . 142

Probabilistic solution of a randomized first order differential equation with discrete delay . . . 151

A predictive method for bridge health monitoring under operational conditions . . . 155

Comparison of a new maximum power point tracking based on neural network with conventional methodologies . . . 160

Influence of different pathologies on the dynamic behaviour and against fatigue of railway steel bridges . . . 166

Statistical-vibratory analysis of wind turbine multipliers under different working conditions . . . 171

Analysis of finite dimensional linear control systems subject to uncertainties via probabilistic densities . . . 176

Topographic representation of cancer data using Boolean Networks . . . 180

Trying to stabilize the population and mean temperature of the World . . . 185

Optimizing the demographic rates to control the dependency ratio in Spain . . . 193

An integer linear programming approach to check the embodied CO_2 emissions of the opaque part of a façade . . . 199

Acoustics on the Poincaré Disk . . . 206

Network computational model to estimate the effectiveness of the influenza vaccine *a posteriori* . . . 211

1

A personality mathematical model of placebo with or without deception: an application of the Self-Regulation Therapy

S. Amigó [♭1], Joan C. Micó[♮] and Antonio Caselles[♯]

(♭) Departament de Personalitat, Avaluació i Tractaments Psicològics,
Universitat de València,
(♮) Institut Universitari de Matemàtica Multidisciplinar,
Universitat Politècnica de València,
(♯) IASCYS member (retired), Departament de Matemàtica Aplicada,
Universitat de València.

1 Introduction

A series of studies show that placebo has an important impact on the improvement in various disease symptoms [1]. However, placebo has important ethical limitations since it is based on deception [2]. Besides, a series of pioneering studies show that placebo is effective even without deception. In fact, the placebo without deception improves the symptoms of irritable bowel syndrome, allergic rhinitis, headache and back pain, major depression, attention-deficit hyperactivity disorder (ADHD) and cancer-related fatigue [3]. Two mechanisms that have been proposed to explain the placebo effect: Expectations, and Classical Conditioning [4]. However, it is possible that the same mechanisms act to both deceptive placebo and without deception. For instance, a study on the placebo without deception to treat pain proves that the conditioned patients experienced a therapeutic effect for longer periods (4 days) even though when they know they are receiving placebo. Thus, the placebo without deception can be independent of expectations [5]. In other studies, increasing positive expectations improves the placebo without deception result [6]. It is very likely that the two mechanisms, the classical conditioning and the increase in expectations, contribute to the placebo effect of no deception [7]. On the other hand, the Self-Regulation Therapy (SRT) is a therapeutic procedure based on suggestion. It combines positive expectations and classical conditioning to reproduce the effects of drugs [4]. Therefore the SRT can be considered a placebo without deception procedure. This study compares the effectiveness of the SRT and deceptive placebo for the reproduction of the effect of a stimulant drug, methylphenidate (MPH), as well as the dynamical response to both SRT and deceptive placebo can be reproduced with a personality mathematical model presented.

[1]e-mail: salvador.amigo@uv.es

2 Methodology

A within-subject, crossover, double-blind, placebo-control design was employed in this study. Two healthy male volunteers participated in this study, with ages of 56 and 57 years old. A single-case experimental ABC design was used. In each phase, one of three conditions was administered: placebo, 5 mg or 10 mg of MPH. The order of administration (MPH or placebo) was determined by random assignment, unknown both for the participants and the research assistant. In a previous study [8], Participant 2 used the SRT to reproduce the effect of MPH, whose result will be considered in this study. In all phases the participants filled in a sheet of adjectives every 10 minutes over a 3-hour period. These adjectives measure the General Factor of Personality (GFP), which represents the organism's general activation. It is a Five-Adjective Scale of the General Factor of Personality, and the five adjectives are adventurous, daring, enthusiastic, merry and bored [9]. The participants had already experienced the effects of MPH on previous occasions. At the end of the experiment they tried to guess what had been taken in each phase.

3 Results

Fig. 1 depicts the GFP scores during the 3-hour period of Phase Placebo and 10 mg condition for Participant 1. He had correctly guessed that he took 10 mg of MPH (the most intense effect of the two MPH conditions) but thought that the placebo was 5 mg. This was a deceptive placebo. The shape of both curves is very similar (inverted U) as well as the peak, but the effect is clearly faster (the slope is more pronounced) in the 10 mg of MPH condition. Fig. 2 depicts the GFP scores during the 3-hour period of Phase Placebo, 5 mg (the most intense effect of the two MPH conditions) and SRT conditions for Participant 2. This participant had correctly guessed the placebo condition and thought that the 5 mg was 10 mg MPH. The three curves are very similar (inverted U). It only fits to stand out that SRT effect is slightly more intense than placebo. By comparing the placebo response of the two participants, we observe that the peak of the effect is similar for both of them, but that the shape of the curve is different. Thus, in the case of Participant 2, the shape of the placebo curve was very similar to that of the SRT, with a very fast onset, indicating that the placebo effect in Participant 2 is a more accurate reproduction of the effect of the drug.

4 The personality mathematical model

The model presented is a stimulus-response model, where the stimulus can be MPH, SRT or Placebo. It si written as:

$$\left.\begin{array}{rcl} \frac{dm(t)}{dt} & = & -\alpha \cdot m(t) \\ m(0) & = & M \\ \frac{ds(t)}{dt} & = & \alpha \cdot m(t) - \beta \cdot \int_0^t e^{\frac{x-t}{\tau}} \cdot s(x) \cdot y(x) dx \\ s(0) & = & 0 \\ \frac{dy(t)}{dt} & = & a(b - y(t)) + p \cdot s(t) \cdot y(t) \\ y(0) & = & y_0 \end{array}\right\} \quad (1)$$

Note that (1) is a coupled a system of two differential equations and one integro-differential equation. The $m(t)$ variable is evolution of the stimulus before entering in the metabolizing organism system, being M the stimulus initial amount and α is the stimulus assimilation rate. The $s(t)$ variable represents the stimulus, i.e., the amount in organism of the stimulus, assuming that the its initial value is zero, due to the experimental conditions, being β is the stimulus metabolizing rate. The $y(t)$ variable represents the GFP dynamics; and b and y_0 are respectively its tonic level and its initial value. Its dynamics is a balance of three terms, which provide the time derivative of the GFP: the homeostatic control ($a(b - y(t))$), i.e., the cause of the fast recovering of the tonic level b, the excitation effect ($p \cdot s(t) \cdot y(t)$), which tends to increase the GFP, and the inhibitor effect ($\beta \cdot \int_0^t e^{\frac{x-t}{\tau}} \cdot s(x) \cdot y(x)dx$), which tends to decrease the GFP and is the cause of a continuously delayed recovering. Parameters a, p, q and τ are named respectively the homeostatic control power, the excitation effect power, the inhibitor effect power and the inhibitor effect delay. Unlike the model presented in [10], the excitation effect is non-linear and the inhibitor effect appears in the differential equation of the $s(t)$ dynamics. This last feature permits to reduce from eight to seven the number of parameters to be calibrated, which represents a reduction in the model calibration complexity. In addition, the hypothesis that underlies the model is different to the presented in [10]: the inhibitor effect is the delayed organism's reaction to the effects produced by the stimulus in order to decrease the amount of the stimulus in the organism.

The calibration of (1) provides two different kinds of results, depending on the participant and on the kind of stimulus. On a hand, the model does not calibrate correctly for the scores of Participant 1 when the stimulus is placebo: the determination coefficient for the GFP response is $R^2 = 0.49$ and the residuals are not random. This case corresponds with a non-clear subjective sensation of the response. However, the model does calibrate correctly when the stimulus is 10 mg of MPH, with determination coefficient for the GFP response R^2=0.94 and random residuals. It does correspond with a clear subjective sensation of the response. The result is provided in Fig. 3. On the other hand, the model does calibrate correctly for the three kinds of stimulus in Participant 2. For placebo stimulus $R^2 = 0.98$, for the 5 mg of MPH stimulus R^2=0.97, for the SRT stimulus $R^2 = 0.85$, and for the three cases the residuals are random. They also correspond with clear subjective sensations of the response. The results are provided, respectively, in Fig. 4, Fig. 5 and Fig. 6. We show the model's parameters for the most important conditions for both participants in the Tables 1-4. We can see in case 1 that parameter M in table 1, which represents the equivalent of 10 mg of MPH in the placebo condition, is low (6.13), while in case 2, we observe that parameter M is very close to 10 mg, being 9.19 and 9.44 for placebo and SRT conditions, respectively. All this confirms what has been said above, and it is that in the SRT-trained participant a more similar effect to the drug is observed, whether it is a placebo with deception or without deception.

5 Discussion

In this study we compared the effect of a single dose of methylphenidate with placebo over 3 hours on a personality scale. The participants were volunteers who had already experienced the effects of MPH previously. The effect curve (inverted U) of the placebo was very similar to that of MPH. This supports the thesis of classical conditioning as a basic mechanism of the placebo

effect. Participant 1 thought he was taking 5 mg of MPH when he actually took a placebo. This can support the idea that novelty and uncertainty can be positive factors in placebo response, especially in placebo without deception, whose basic mechanism is the prediction and error processing. More dopamine is only release with the uncertainty [11]. The peak of the curve of the placebo effect was very similar for the two participants and although the shape of the curve was similar (U Inverted), the slope was much more pronounced in Participant 2, which makes the effect more similar to that produced by the MPH. This can be interpreted as the deceptive placebo in a person trained with SRT is very similar to the reproduction of the effects of the drug (placebo without deception). In addition, Participant 2 guessed the placebo since he was able to distinguish it from the effect of the drug, sometimes by the fact of being trained to discriminate the situations in which he takes the drug and in which he reproduces the drug. In summary, the experience with the drug increases the placebo effect, and the training in SRT causes the placebo to be easily detected but, even so, its effect is even more similar to that of the drug. The therapeutic application of these results is possible, using deceptive and placebos without deception as "dose extenders" based on classical conditioning. In a study with children with ADHD, pairing open-label placebo pills with amphetamines allowed children to be treated effectively with a lower dose of stimulant medication [12]. Besides, training patients with SRT can allow them to learn and use coping strategies to overcome anxiety and depression disorders and other kinds of diseases [13]. Finally, the personality mathematical model presented, which has a stimulus-response model, is an advance respect to the presented in [10]: it has one less parameter to be calibrated and links non-linearly the stimulus dynamics to the GFP dynamics. It can become in a future a good mathematical tool to predict the FGP responses to the placebo and SRT stimulus, after being calibrated the GFP dynamics as a consequence of MPH consumption, as well as to classify typologies of personality that help us to solve the personality disorders mentioned above.

Figures

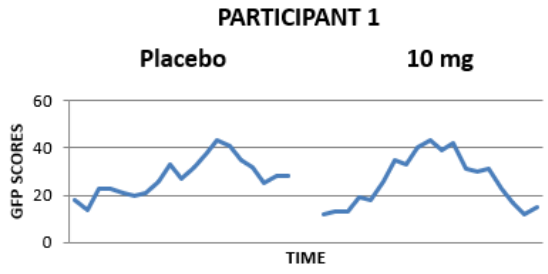


Figure 1: GFP dynamics of Placebo and 10 mg conditions for Participant 1.

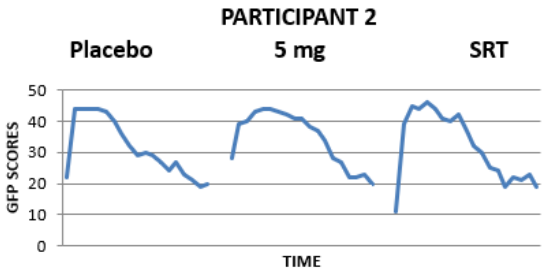


Figure 2: GFP dynamics of Placebo, 5 mg and SRT conditions for Participant 2.

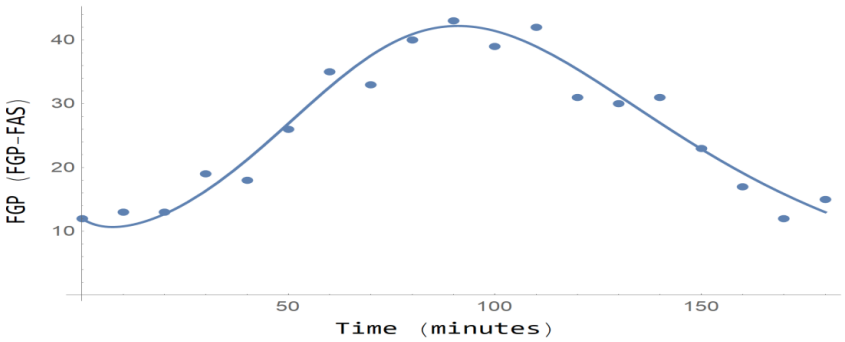


Figure 3: GFP dynamics for 10 mg of MPH (Participant 1). $R^2 = 0.94$.

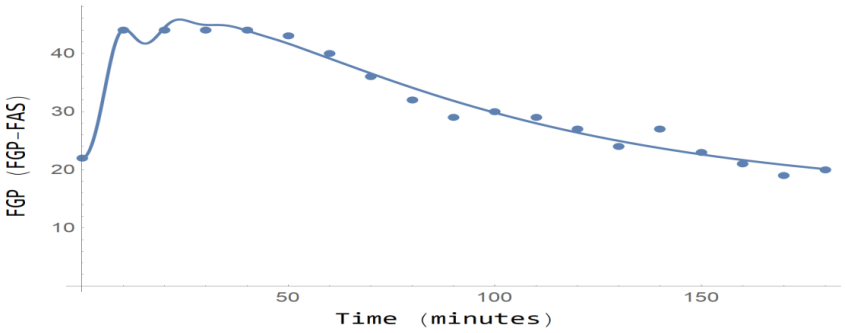


Figure 4: GFP dynamics for placebo (Participant 2). $R^2 = 0.98$.

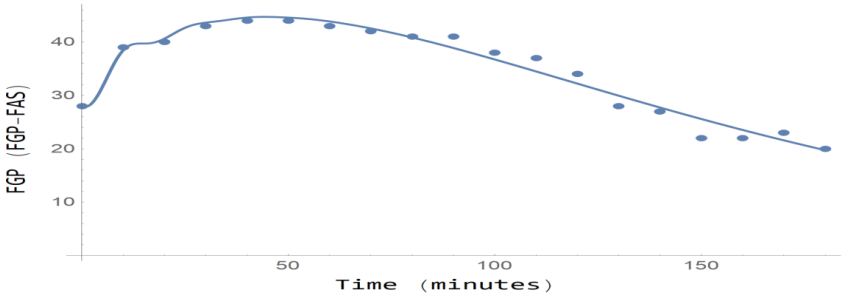


Figure 5: GFP dynamics for 5 mg of MPH (Participant 2). $R^2 = 0.97$.

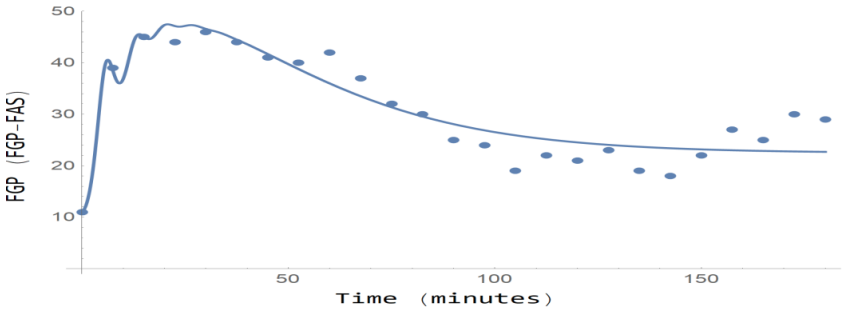


Figure 6: GFP dynamics for SRT (Participant 2). $R^2 = 0.85$.

Tables

Parameter symbol	Name	Optimalvalue
M	PLACEBO-MPH	6.13
τ	Inhibitor effect delay	12.0136952188179790
α	Assimilation rate	0.097804
β	Distribution rate	0.111893
a	Homeostatic control power	0.0437382
b	Tonic level	33.5433959960937500
p	Excitation effect power	0
q	Inhibitor effect power	0.00196436

Table 1: Optimal values of the model parameters, **Placebo Phase**, corresponding to the GFP dynamics (Y). Participant 1.

Parameter symbol	Name	Optimalvalue
M	Methylphenidate	10
τ	Inhibitor effect delay	230.8067913884371800
α	Assimilation rate	0.0240989
β	Distribution rate	0.0365474
a	Homeostatic control power	0.00703561
b	Tonic level	1.04588×10^{-7}
p	Excitation effect power	0.0113137
q	Inhibitor effect power	0.0000795275

Table 2: Optimal values of the model parameters, **10 mg Phase**, corresponding to the GFP dynamics (Y). Participant 1.

Parameter symbol	Name	Optimalvalue
M	PLACEBO-MPH	9.19
τ	Inhibitor effect delay	214.018949817644650
α	Assimilation rate	0.0356595
β	Distribution rate	0.000724157
a	Homeostatic control power	0.538046
b	Tonic level	40.788269042968750
p	Excitation effect power	0.0160338
q	Inhibitor effect power	0.000504011

Table 3: Optimal values of the model parameters, **PLACEBO PHASE**, corresponding to the GFP dynamics (Y). Participant 2.

Parameter symbol	Name	Optimalvalue
M	PLACEBO-MPH	9,44
τ	Inhibitor effect delay	125.9781022862648600
α	Assimilation rate	0.173822
β	Distribution rate	0.559143
a	Homeostatic control power	0.0136305
b	Tonic level	36.866760253906250
p	Excitation effect power	0.0966999
q	Inhibitor effect power	0.00110927

Table 4: Optimal values of the model parameters, **SRT PHASE**, corresponding to the GFP dynamics (Y). Participant 2.

References

[1] Finniss, D.G., Kaptchuk, T.J., Miller, F. and Benedetti, F., Biological, clinical, and ethical advances of placebo effects. *Lancet*, 686-695, 2010.

[2] Blease, C.R., Bishop, F.L. and Kaptchuk, T.J., Informed consent and clinical trials: where is the placebo effect? *BMJ Clinical Research* , 356:j463, 2017.

[3] Charlesworth, J.E.G., Petkovic, G., Kelley, J.M., Hunter, M., Onakpoya, I., Roberts, N., Miller, F.G. and Howick, J., Effects of placebos without deception compared with no treatment: a systematic review and meta-analysis. *J Evid Based Med*, 10: 97-107, 2017.

[4] Amigó, S., Sugar Pills To Experience Cocaine and Other Drug Effects: The Self-Regulation Therapy as a Placebo Without Deception. *EC Neurology*, 3.1: 320-331, 2016.

[5] Schafer, S.M, Colloca, L. and Wager, T.D., Conditioned placebo analgesia persists when subjects know they are receiving a placebo. *The journal of pain : official journal of the American Pain Society*, 16: 412–420, 2015.

[6] Wei, H., Zhou, L., Zhang, H., Chen, J., Lu, X. and Hu, L., The influence of expectation on nondeceptive placebo and nocebo effects. *Pain Res Manag*, 2018.

[7] Petkovic, G., Charlesworth, J.E.G. and Kelley, J. et al., Effects of placebos without deception compared with no treatment: protocol for a systematic review and meta-analysis. *BMJ Open*, 5:e009428, 2015.

[8] Amigó, S., Micó, J.C. and Caselles, A., Methylphenidate and Self-Regulation Therapy: A systemic mathematical model. *Modelling for Engineering & Human Behaviour 2017*, 13-18. Valencia (Spain).

[9] Amigó, S., Micó, J.C. and Caselles, A., Five adjectives to explain the whole personality: a brief scale of personality. *Revista Internacional de Sistemas*, 16: 41-43, 2009.

[10] Micó, J.C., Amigó, S., Caselles, A., From the Big Five to the General Factor of Personality: a Dynamic Approach, *Span. J. Psychol.* 17:E74: 1-18, 2014.

[11] Kaptchuk, T.J., Open-label placebos: reflections on a research agenda. *Perspect Biol Med*, 61: 311–34, 2018.

[12] Sandler, A.D. and Bodfish, J.W., Open-label use of placebos in the treatment of ADHD: A pilot study. *Child: Care, Health and Development*. 34: 104–110, 2008.

[13] Amigó, S., Uso potencial de metilfenidato y la sugestión en el tratamiento psicológico y en el aumento de las potencialidades humanas: Un estudio de caso [Potential use of methylphenidate and suggestion in the psychological treatment and in the increase of human potentialities: A case study]. *Análisis y Modificación de Conducta*, 23: 863–890, 1997.

2

The role of police deterrence in urban burglary prevention: a new mathematical approach

J. Saldaña [♭], M. Aguareles[♭], A. Avinyó[♭1], M. Pellicer[♭] and J. Ripoll[♭]

(♭) Universitat de Girona.

1 Introduction

We introduce a model for the dynamics of burglars and victimized houses which takes dynamic police deterrence into account [5]. In contrast to previous works on urban crime modeling [3,6], mainly based on the spatio-temporal description of criminal activities, here we have adopted a new approach focusing on the timing of burglary activity itself and it is inspired by age-structured population models [2]. In fact, the model is capable to tackle several scenarios on the criminal activity by assuming different forms of vulnerability and recurrence rates. Moreover, it permits to obtain some interesting analytic results on the expected times between two consecutives offenses.

2 The model

From now on, $N(\tau_1, t)$ will denote the density of burglars at time t that have offended τ_1 units of time ago, and $H(\tau_2, t)$ will denote the density of houses at time t that have been burgled τ_2 units of time ago. So, we can think of τ_1 as the burglary age of a burglar, and of τ_2 as the victimization age of a house. We consider that the total number of burglars remains constant in time and the same is assumed for the total number of houses:

$$N^0 := \int_0^\infty N(\tau, t)\, d\tau, \qquad H^0 := \int_0^\infty H(\tau, t)\, d\tau, \ \ \forall\, t \geq 0. \tag{1}$$

To derive the model equations, we need to establish how offenders act and how houses are burgled. First, we define the propensity function of a burglar of burglary age τ_1, $0 \leq f_0(\tau_1) \leq 1$, that represents its natural predisposition for crime, the vulnerability function, $0 \leq \alpha_0(\tau_2) \leq 1$, that reflects the ease to break into a house of age τ_2 and the mean vulnerability of houses at time t

$$\overline{\alpha}_0(t) := \int_0^\infty \alpha_0(\tau) \frac{H(\tau, t)}{H^0}\, d\tau. \tag{2}$$

Following the classic notion in criminology that for a burglary to occur, a motivated burglar must find a suitable house [1], we define the per-capita deterred recurrence rate $f_{D_T}(\tau_1, t)$ of

[1]e-mail: albert.avinyo@udg.edu

burglars of age τ_1 at time t as

$$f_{_{D_T}}(\tau_1, t) := D_T\, f_0(\tau_1)\, f_1(\overline{\alpha}_0(t)), \tag{3}$$

where $f_1(\overline{\alpha}_0)$ is a strictly increasing function tending to 1 as $\overline{\alpha}_0 \to 1$. So, the propensity level of a burglar is assumed to be a function of his age τ_1, the accessibility to target sites is represented by the mean vulnerability $\overline{\alpha}_0$ of a house at time t, and the police deterrence by the extra factor $0 < D_T \leq 1$:

$$D_T\left(H\left(0,\cdot\right), t\right) := F\left(\int_{t-T}^{t} H(0,s)e^{-\xi(t-s)}\, ds\right). \tag{4}$$

Here, T is a certain observation period of time, F is a strictly decreasing differentiable function of the weighted number of burglaries occurred during the last T units of time with $F(0) = 1$ and $F(x) \to 0$ as $x \to \infty$. The decreasing behavior of F implies that, if the number of burglaries within the observation period experiences a dramatic increase, then the deterred recurrence rate of burglars will be significantly reduced due to the dissuasive action of police. On the other hand, if the length of the observation period T is set to zero, the police does not take into account any past event and, hence, the recurrence rate does not change. The exponential weight is a discount factor with ξ being the discount rate. This term reflects the idea that recently committed burglaries have a higher impact on the current police response than those that have been perpetraded long time ago.

Finally, the victimization rate $\alpha_{_{D_T}}$ of a house of age τ_2 will be proportional to the number of active burglars per house and per unit of time, that is,

$$\alpha_{_{D_T}}(\tau_2, t) := \frac{\alpha_0(\tau_2)}{H^0} \int_0^\infty f_{D_T}(\tau, t) N(\tau, t)\, d\tau, \tag{5}$$

where the proportionality constant is the vulnerability function $\alpha_0(\tau_2)$. The burglary rate at time t of houses of age τ_2 is then given by $\alpha_{_{D_T}}(\tau_2, t)H(\tau_2, t)$.

The dynamics for the densities $N(\tau_1, t)$ and $H(\tau_2, t)$ will be described in terms of a predator-prey type of interaction between houses and burglars, where the former are the preys and the latter the predators. Since when a burglar strikes or when a house is burgled, their age resets to $\tau_1 = 0$ and $\tau_2 = 0$, respectively, the burglary model is given by the following nonlinear system of first-order partial differential equations with nonlocal boundary conditions:

$$\begin{cases} \partial_t N(\tau_1, t) + \partial_{\tau_1} N(\tau_1, t) &= -f_{_{D_T}}(\tau_1, t) N(\tau_1, t), \\ \partial_t H(\tau_2, t) + \partial_{\tau_2} H(\tau_2, t) &= -\alpha_{_{D_T}}(\tau_2, t) H(\tau_2, t), \\ N(0, t) &= \displaystyle\int_0^\infty f_{_{D_T}}(\tau, t) N(\tau, t)\, d\tau, \\ H(0, t) &= \displaystyle\int_0^\infty \alpha_{_{D_T}}(\tau, t) H(\tau, t)\, d\tau, \end{cases} \tag{6}$$

endowed with the initial conditions $N(\tau_1, 0) = N_0(\tau_1)$ and $H(\tau_2, 0) = H_0(\tau_2)$, both nonnegative functions in $L^1(0, \infty)$, and being the initial history $H(0, t) = h_0(t)$ for $-T \leq t < 0$, a continuous bounded function in $[-T, 0)$.

Under suitable hypotheses on f_0, f_1 and α_0, the well-posedness of (6) and the existence of a nonnegative global solution follows in a similar way as the one for n-species model presented in [4].

3 The equilibrium

By a suitable rescaling of the time variable, one can see that the solution of (6) tends to a unique equilibrium $(N^*(\tau_1), H^*(\tau_2))$:

$$N^*(\tau_1) = N^0 \frac{\Pi_b^D(\tau_1)}{\int_0^\infty \Pi_b^D(\tau)\, d\tau}, \quad H^*(\tau_2) = H^0 \frac{\Pi_h^D(\tau_2)}{\int_0^\infty \Pi_h^D(\tau)\, d\tau}, \tag{7}$$

where the probability at equilibrium that a burglar remains inactive up to time τ_1 under the dissuasive action of police provided by D_T is

$$\Pi_b^D(\tau_1) := \exp\left(-\int_0^{\tau_1} f^*_{D_T^*}(\tau)\, d\tau\right), \tag{8}$$

and the probability at equilibrium of a house not being burgled up to time τ_2 also under deterrence is

$$\Pi_h^D(\tau_2) := \exp\left(-\int_0^{\tau_2} \alpha^*_{D_T^*}(\tau)\, d\tau\right) = \exp\left(-\frac{N^0}{H^0}\frac{\int_0^{\tau_2} \alpha_0(\tau)\ d\tau}{\int_0^\infty \Pi_b^D(\tau)\, d\tau}\right). \tag{9}$$

Here, $f^*_{D_T^*}(\tau_1)$ and $\alpha^*_{D_T^*}(\tau_2)$ denote, respectively, the recurrence rate of burglars and the victimization rate of houses at equilibrium. These expressions for (N^*, H^*) are not explicit because they both depend on the deterrence factor and the mean vulnerability at equilibrium, D_T^* and $\overline{\alpha}_0^*$, respectively, that they are the solution of the system

$$D_T^* = F\left(H^0 \frac{1-e^{-\xi T}}{\xi}\left(\int_0^\infty \exp\left(-\frac{N^0}{H^0}\frac{\int_0^\tau \alpha_0(s)\, ds}{\int_0^\infty e^{-D_T^* f_1(\overline{\alpha}_0^*)\int_0^u f_0(s)\, ds}\, du}\right)\, d\tau\right)^{-1}\right), \tag{10}$$

$$\overline{\alpha}_0^* = \frac{\dfrac{H^0}{N^0}\displaystyle\int_0^\infty e^{-D_T^* f_1(\overline{\alpha}_0^*)\int_0^\tau f_0(s)\, ds}\, d\tau}{\displaystyle\int_0^\infty \exp\left(-\frac{N^0}{H^0}\frac{\int_0^\tau \alpha_0(s)\, ds}{\int_0^\infty e^{-D_T^* f_1(\overline{\alpha}_0^*)\int_0^u f_0(s)\, ds}\, du}\right) d\tau}. \tag{11}$$

The first equation is obtained by replacing $H^*(0)$ by its formal expression in (7) into the expression of D_T^*, whereas the second one is given by (2) after plugging $H^*(\tau_2)$ in. The existence of a unique solution $(D_T^*, \overline{\alpha}_0^*)$ of the system is guaranteed because of the convergence of solutions to a globally stable equilibrium. So, we can compute it numerically and, then, the equilibrium densities $(N^*(\tau_1), H^*(\tau_2))$.

We are interested now in obtaining some analytical results about the burglary activity. At the equilibrium, the expected time between two consecutive offenses committed by the same burglar, R_b, and the expected time between two consecutive burglaries of the same house, R_h, are given by

$$R_b(T) := \int_0^\infty \tau f^*_{D_T^*}(\tau) \Pi_b^D(\tau)\, d\tau = \int_0^\infty \Pi_b^D(\tau)\, d\tau, \tag{12}$$

$$R_h(T) := \int_0^\infty \tau \alpha^*_{D^*_T}(\tau) \Pi^D_h(\tau)\, d\tau = \int_0^\infty \Pi^D_h(\tau)\, d\tau. \tag{13}$$

These expected times depend on T through D^*_T. In fact, by means of the differentiation chain rule and the fact that D^*_T decreases with T, one easily sees that both expected times are increasing functions of T.

The relationship between R_h and R_b follows upon replacing Π^D_h in (13) by its expression (9) which gives

$$R_h = \int_0^\infty \exp\left(-\frac{N^0}{H^0}\frac{\int_0^{\tau_2} \alpha_0(\tau)\, d\tau}{R_b}\right) d\tau_2. \tag{14}$$

As one would expect, the relation is increasing and reflects that the longer a burglar waits to commit the following burglary, the longer a house remains safe.

4 A numerical example

In order to show the flexibility of our model, in this section, we consider a scenario where

$$f_0(\tau_1) := \frac{\tau_1}{1+\tau_1}, \quad \alpha_0(\tau_2) := \frac{\tau_2^3}{10^3+\tau_2^3}, \quad f_1(\overline{\alpha}_0) := 1 - e^{-5\,\overline{\alpha}_0}, \tag{15}$$

and,

$$D_T\left(H\left(0,\cdot\right),t\right) := \exp\left(-10^{-3}\int_{t-T}^{t} H(0,s)e^{-\xi(t-s)}\, ds\right). \tag{16}$$

Then, after solving the system (10-11) and obtaining the expected times between two consecutive offenses committed by the same burglar (12) and two consecutive burglaries of the same house (13), we can see, in Figure 1, the number of burglaries per day as a function of the police observation length T, that is, $H^*(0) = H^0/(R_h(T))$.

As the number of burglaries is a decreasing function of T, we see that the longer the observation period is the higher is the deterrence effect of the police, although it always has a positive limit. In general, it is observed that after some point it is not worth considering longer observation lengths since the decrease in the amount of burglaries is unnoticeable. However, by decreasing the value of ξ one sets a new asymptote which also decreases.

Acknowledgements

All authors are part of the Catalan Research group 2017 SGR 1392 and M.A., A.A., M.P. and J.R. have been partially supported by the MINECO grant MTM2017-84214-C2-2-P (Spain).

References

[1] Cohen, L.E., and Felson, M., Social change and crime rate trends: A routine activity approach, *Amer. Sociol, Rev.*, (44): 588–608,1979.

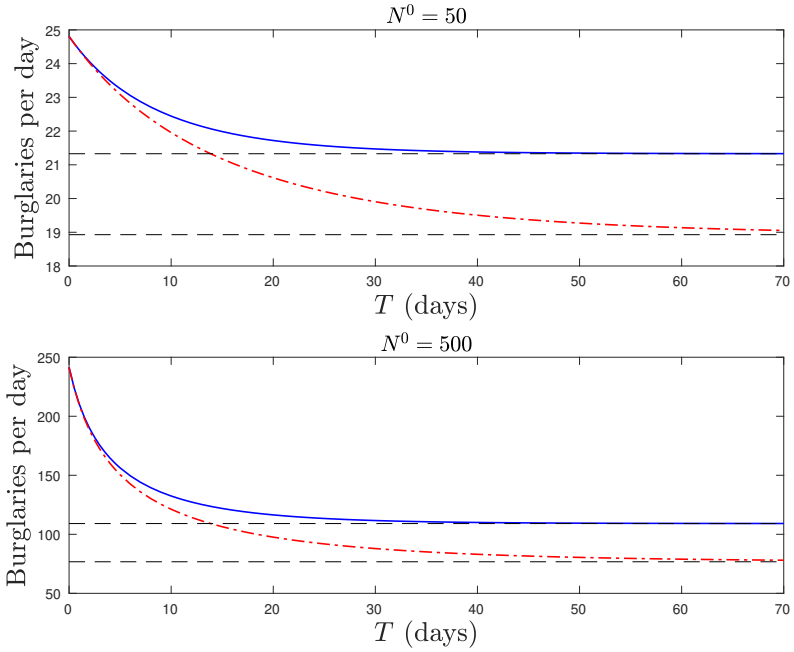


Figure 1: Rate of burglaries while varying the observation period T. Blue solid curve: $H^0 = 10^5$, $\xi = 0.1$; red dash-dotted curve: $H^0 = 10^5$, $\xi = 0.05$.

[2] Inaba H., *Age-Structured Population Dynamics in Demography and Epidemiology*. Singapore. Springer Science+Business Media, 2017.

[3] Pitcher, A. B., Adding police to a mathematical model of burglary, *Eur. J. Appl. Math.*, 21: 401–419, 2010.

[4] Poskrobko, A. and Dawidowicz, A.L., On the multispecies delayed Gurtin-MacCamy model, *Abstract and Applied Analysis*, vol. **2013**, 908768, 2013.

[5] Saldaña, J., Aguareles, M., Avinyó, A., Pellicer, M. and Ripoll, J., An age- structured population approach for the mathematical modeling of urban burglaries, *SIAM Journal on Applied Dynamical Systems*, 17: 2733- 2760, 2018.

[6] Short, M. B., D'Orsogna, M. R., Pasour, V. B., Tita, G. E., Brantingham, P. J., Bertozzi, A. L. and Chayes, L. B., A statistical model of criminal behavior, *Math. Models Methods Appl. Sci.*, 18: 1249–1267, 2008.

3

A Heuristic optimization approach to solve berth allocation problem

Clara Burgos Simón[♭1], Juan-Carlos Cortés López[♭], David Martínez-Rodríguez[♭] and Rafael-Jacinto Villanueva Micó[♭]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

One of the main consequences of Globalization is the development of international trade (imports and exports). This fact leads to an increase in vessel transports and container manipulations. To get an idea about the magnitude of the problem, while in Busan port (South of Korea) in 2011 they were operating more than 10000 twenty-foot equivalent unit (TEU, unit to measure containers), in April of 2013 they were handling 18000 TEU, almost twice.

The requirements for hub ports have also changed and developing new strategies in container manipulations is becoming really important. Some shipping lines require new performance levels from terminal as a part of the contract conditions, as the throughput rate per berth, the turnaround time of a vessel or the increment of the rate containers movements, among others [2].

The aim of this contribution is the development of a new approach to find the optimal planning of docking the vessels. Berth planning is defined as the process of establish the best outline of the vessels in the corresponding berths and the display of Quay Cranes (QC) in order to minimize the cost of the terminal and maximize the service of the containers movements. It is a complex problem because the QC deployment is closely linked with the best berth outline.

In the literature, depending on the initial display of the vessels, we can study two different situations: the first is named Static Berth Allocation Problem (SBAP) and it considers that all the vessels are in the anchoring spot waiting to be docked; the second, named by Dynamical Berth Allocation Problem (DBAP), assumes that the vessels arrive dynamically. The interesting problem for the terminal is the second one because all the vessels arrive at different times. With the technique we propose, we can solve both problems, nevertheless, as the aim of this paper is also to study the goodness of our method, we will use SBAP because it is easier to obtain the exact solution for SBAP than for DBAP.

[1]e-mail: clabursi@posgrado.upv.es

This abstract is organized as follows. In Section 2 we describe the approach to establish the best planning of the berth allocation problem. Section 3 is devoted to prove the goodness of our approach. Finally Section 4 is addressed to conclusions.

2 Procedure design

In this section we explain the procedure of assigning the vessels in the corresponding berth. It is based on an optimization technique, so we need two elements: a fitness function and an heuristic approach. Without loss of generality we consider that there is only one QC per berth. The case with more than one QC per berth is analogous but taking into account different unloaded time of the vessels.

2.1 Fitness function

In the literature we can find several fitness functions [1]. Depending on which one we use, we can benefit the terminal or the shipping line, since their interests are not the same. While the main aim of the terminal is to minimize its economic cost, the aim of the shipping line is to have its vessels unloaded in the shortest time. The fitness function we use is a simplification of the one we find in [1, Section 2]. It is defined as the sum of the waiting and operating time of the vessels in the different berths, understanding as operating time the period that the vessel needs to be docking, unloading, loading and setting sail. This function has been chosen because benefits the shipping line and terminal, as it minimizes the time that vessels are waiting for and operating also it allows that the terminal could deal with more vessels.

Figure 1 shows an example of a berthing plan with two berths m_1 and m_2 and seven vessels b_1, b_2, b_3, b_4, b_5, b_6 and b_7. The squares represent the vessels and the numbers inside the squares represent the operating time of each vessel. The red numbers are the times the vessels are waiting for and operating. Then, the cost of this berthing plan is the sum of all the red numbers

$$2+2+4+2+4+3+1+1+7+1+7+2=34.$$

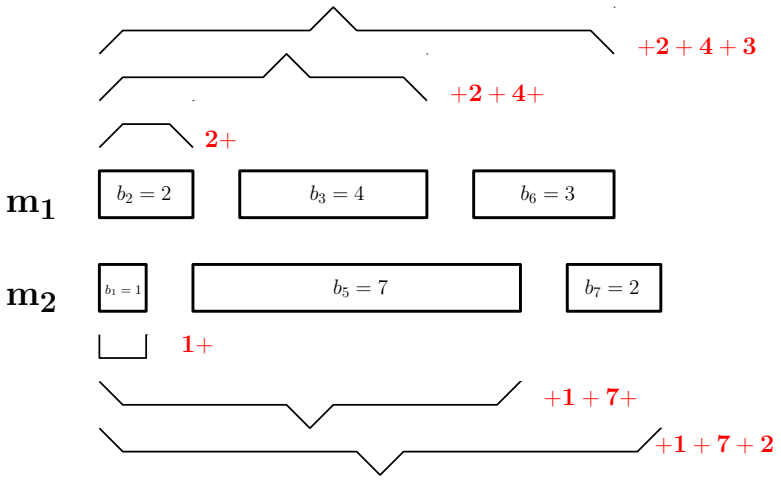


Figure 1: Example of a berthing plan and computation of its cost

2.2 Heuristic technique

This section is devoted to define the heuristic approach that allows us to minimize the fitness function given in Subsection 2.1. Our technique consists of making changes in the vessel's position and compare the new solutions with the past ones. Moreover we need to take into account some issues:

- The vessels are moving one by one, is other words, in the same iteration we can only make one movement.
- If we move a vessel and the new solution is better than the previous one, that vessel are not going to move in the next 10 iterations.

In Figure 2 we can see graphically different movements of our heuristic approach, the red lines represent the new changes in the vessels position.

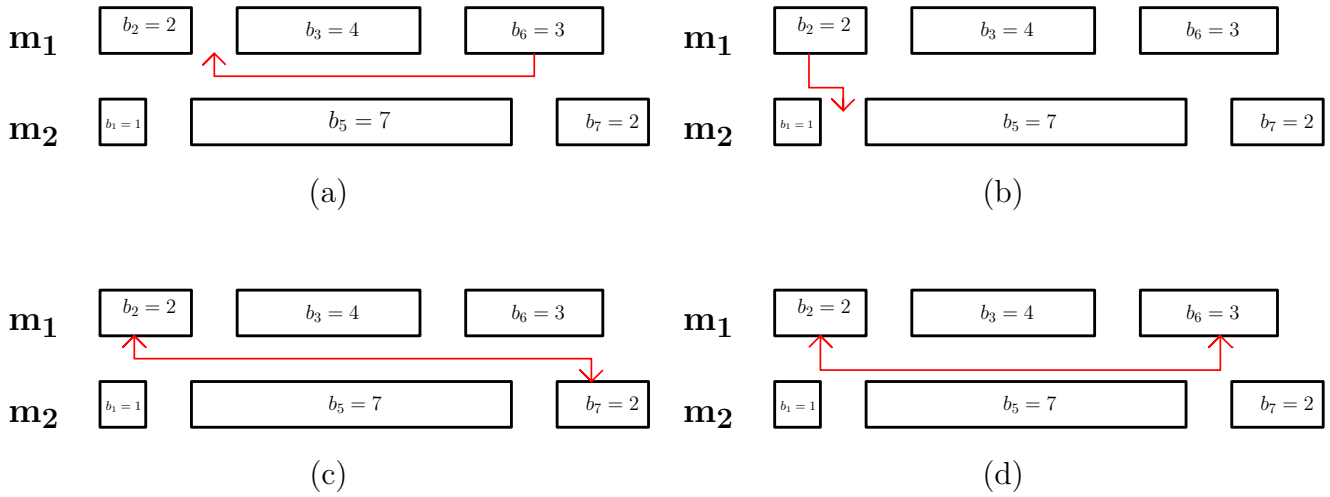


Figure 2: Different moves of the heuristic approach, for example changing the position of a vessel in the same berth (a), add a vessel in other berth (b), exchange two vessels in different berths (c), exchange two vessels in the same berth (d).

3 Results

The aim of this section is to prove the goodness of our method. To do so, we have develop an example with 4 berths and 6 vessels. Considering 6 different situations of operation times of vessels, in Table 1 we show in the first column the exact solution, in the second the best solution given by the heuristic approach and in the third the relative error. As we can see the magnitude of the relative error is 10^{-2}, thus we can verify the goodness of our approach.

	Best solution	Heuristic solution	Relative error
Situation 1	167.11	167.51	0.002387
Situation 2	165.60	165.80	0.001206
Situation 3	165.82	165.82	0
Situation 4	163.87	163.87	0
Situation 5	165.91	165.91	0
Situation 6	167.87	167.94	0.000416

Table 1: Relative error between the exact solution and the heuristic solution of the static berth allocation problem.

4 Conclusions

In this contribution we have develop a heuristic approach to find an optimal solution of berth allocation problem. In order to prove the goodness of our method we have considered SBAP and we have compared the results obtained with our heuristic algorithm and the exact solution. Their relative errors are low enough to conclude that we have develop an appropriate heuristic.

Acknowledgements

This work has been partially supported by the Ministerio de Economía y Competitividad grant MTM2017-89664-P and by the Ministerio de Ciencia, Innovación y Universidades (Retos Colaboración 2017) grant RTC-2017-6566-4, "VALKNUT".

References

[1] P. Hansen, C. Oğuz. A note on formulations of static and dynamic berth allocation problems. *Les cahiers du gerad*. ISSN: 0771–2440.

[2] K. Hwan and H. Lee. Chapter 2: Container Terminal Operation: Current Trends and Future Challenges.*Handbook of Ocean container transport logistics* International series in operations researchs and management sciences, 220.

4

Improving the efficiency of orbit determination processes

Miguel Camarasa[♭1], Alicia Cordero[♭] and J.R. Torregrosa[♭]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València.

In Classical Mechanics, the solution of the two-body problem with known initial values is well defined and finds many applications. For instance, it can be applied to approximate the position of an Earth satellite. Nowadays, achieving a certain precision in the calculation of the position of artificial satellites around the Earth is fundamental. However, as the Earth-satellite system is in practice affected by numerous other bodies, a differential correction is often necessary. It is also necessary to know, for example, the position of the International Space Station (ISS) at a given moment or the position of each of the satellites that make up the Global Navigation Satellite System (GNSS).

The problem of n-gravitational bodies is fundamental for Positional Astronomy, as it studies the motion of celestial bodies subjected to forces derived from Newton's gravitation law. An example is the motion of planets around the Sun or the motion of artificial satellites around the Earth. This problem is simplified by dispensing with the gravitational action of stars, planets and other distant celestial bodies, since it is small due to its distance. In addition, the considered stars are replaced by other material points with the same mass and whose position coincides with the corresponding center of gravity. This replacement, proposed by Newton, is justified by the classical theorems of the theory of potential.

Now, it is known that if $n \geq 3$, the problem does not have an analytical solution, so it would be necessary to resort to approximate solutions and the theory of perturbations. That is why we will focus on the two-body problem, that is, $n = 2$. In this case the problem has an analytical solution, i.e., is integrable. Performing three observations in three given times and using the solution of two-body problem, the initial position of the satellite can be adjusted. Nevertheless, we aim to correct the orbit, getting an improved position using high order iterative methods and measuring only angle coordinates.

The coordinate system used is the absolute equatorial one, which has as reference plane the celestial equator, which is an extension of the plane of the terrestrial equator to the celestial sphere, and as direction the celestial poles (North, NCP, and South, SCP, in Figure 1), which are an extension of the Earth's axes. So, the right ascension α and declination δ coordinates

[1]e-mail: miguel.camarasa.buades@gmail.com

yield the position any celestial body at an specific instant in the celestial sphere. Let us remark that the right ascension α is measured on the celestial equator with the Vernal point γ as starting position. This point is one of the intersection points between celestial equator and the ecliptic (extension to the celestial sphere of the orbit of the Earth around the Sun), that apparently crosses the Sun at Spring Equinox (seen from the Earth).

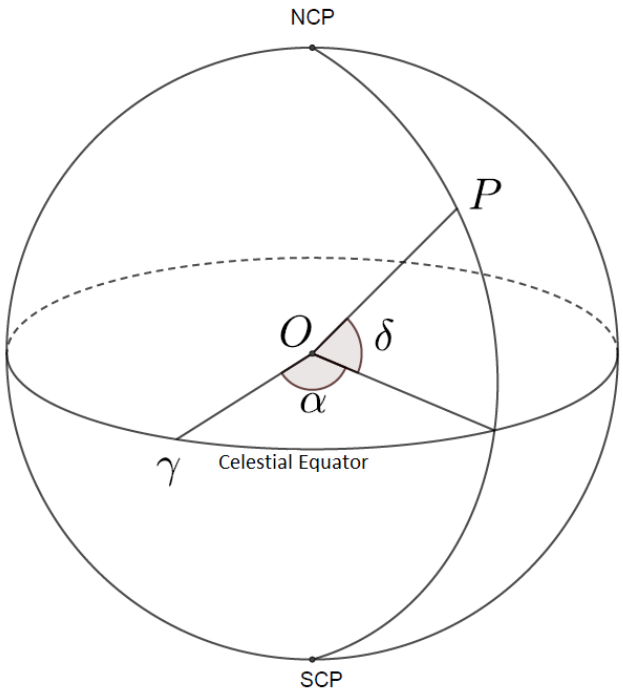


Figure 1: Right Ascension and Declination coordinates

Our next aim is to perform an algorithm (see Figure 2 allowing us to calculate the position of an artificial satellite around the Earth, at a given instant t, known two initial positions X_0 and X_1 in the instants t_0 and t_1. From the Gauss method of the areas, it is possible to obtain, from this information, the speed of the star at instant t_0. Thus, we have the position and velocity of the satellite at instant t_0, which are the initial conditions of the differential equation describing the two-body problem. Solving this equation by the classical method, one obtains the position and velocity of the star in an instant of time t.

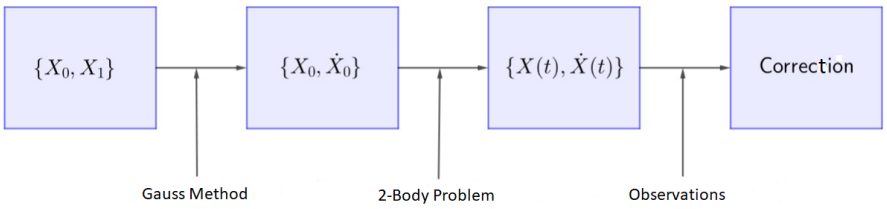


Figure 2: Orbit Correction Process

These calculations are only approximations, as only the Earth-satellite system is being taken into account. The gravitational action of the Moon, the Sun or the other planets of the Solar

System, for example, is not being considered. Nor are being taken into account the modifications that may be being made to the satellite's own trajectory, which surely would not correspond with the calculations made. That is why, if a certain precision is required, a correction is necessary.

The differential correction method used is based on least squares procedure [3] and use angular observations, since they are easier to obtain. These will be declination, δ and right ascension, α. Each of them will be observed at different times $t_0, t_1, \ldots, t_n$, with $\in \mathbb{N}$. These will be, respectively, $\alpha_0^O, \alpha_1^O, \ldots, \alpha_n^O$ right ascensions, and $\delta_0^O, \delta_1^O, \ldots, \delta_n^O$ declinations, where the superscript O represents that they are observed magnitudes. These data are grouped in a single vector λ^O. These data have been obtained from a NASA database, so we also have access to the real positions and velocities we want to estimate with the correction process, which allows us to check the efficiency of our procedure.

The aim is to compare these observations with the calculations made. The function that calculates the declination and right ascension in those instants of time, given an initial position vector X and velocity vector $\dot{X}$, is denoted by F. Therefore, we want to minimize

$$\|\lambda^O - F(X, \dot{X})\|_2^2$$

For getting this aim, it is necessary to find a solution to the following system of equations:

$$\nabla F(X, \dot{X})^T (\lambda^O - F(X, \dot{X})) = 0. \tag{1}$$

This system can be solved, for example, by the multidimensional extension of Steffensen's method, which allows us to avoid the calculation and evaluation of Jacobian matrices, needed for example when Newton's method is used. The initial guess is that obtained through the two-body problem, in order to ensure convergence. It is verified that the correction improves considerably the result obtained only with the problem of the two bodies, since we compare it with the real data of NASA.

So, when convergence is available, the following questions arise: is it possible to reduce the calculation time? Could it be done in a single iteration? To increase the number of observations, would improve numerical calculations? Perhaps it would be better to use a more efficient numerical method to ensure faster convergence? Is it important that observations are made in an instant of time close to the time you want to calculate?

To answer these questions, several studies are conducted to improve the correction:

- Reduction of the calculation time: some simplifications made in equation (1) are justified so that the system of equations can be solved more quickly.

- Increasing of the number of observations used: by a practical analysis it is concluded that four observations are the optimal amount for both right ascension and declination.

- Separation between observations: it follows that the separation between the observations made is very important, as the closer you are to the instant of time in which you want to obtain the position of the star, the better the results will be. In fact, from a certain distance, the results may not converge.

- Implementation of a high order numerical method: a higher order numerical method (than Steffensen's one) is implemented, also free of Jacobian matrices, but with fourth-order of convergence. With it, convergence is achieved in the first iteration, so it can be deduced that the correction process has been improved.

References

[1] Escobal, P. R., *Methods of Orbit Determination*. New York, Wiley & Sons, 1965.

[2] Sevilla, M. J., *Mecánica Celeste Clásica*. Madrid, Instituto de Astronomía y Geodesia, 1989.

[3] Dennis Jr, J.E. and Schnabel, R.B., *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Rice University, 1983.

5

A new three-steps iterative method for solving nonlinear systems

Raudys R. Capdevila [♭1], Alicia Cordero[♭] and Juan R. Torregrosa[♭]

(♭) Instituto de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

The problem of solving equations and systems of nonlinear equations is among the most important in theory and practice, not only of applied mathematics, but also in many branches of science, engineering, physics, computer science, astronomy, finance,... A glance at the literature shows a high level of contemporary interest. The search for solutions of systems of nonlinear equations is an old, frequent and important problem for many applications in mathematics and engineering (for example, see [1–3]).

This work deals with the approximation of a solution ξ of a system of nonlinear equations $F(x) = 0$, where $F: D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a sufficiently differentiable function on the convex set $D \subset \mathbb{R}^n$. The most commonly used techniques are iterative methods, where, from an initial estimate, a sequence is built converging to the solution of the problem under some conditions. Although not as many as in the case of equations, some publications have appeared in the recent years, proposing different iterative methods for solving nonlinear systems. They have made several modifications to the classical methods to accelerate the convergence and to reduce the number of operations and functional evaluations per step of the iterative method. Newton's method is the most used iterative technique for solving this kind of problems, whose iterative expression is

$$x^{(k+1)} = x^{(k)} - [F'(x^{(k)})]^{-1} F(x^{(k)}), \quad k = 0, 1, \ldots, \tag{1}$$

where $F'(x)$ denotes the Jacobian matrix associated to function F.

Let $\{x^{(k)}\}_{k\geq 0}$ be a sequence in $\mathbb{R}^n$ which converges to ξ, then the convergence is called of order p with $p \geq 1$, if there exists $M > 0$ ($0 < M < 1$ if $p = 1$) and k_0 such that

$$\|x^{(k+1)} - \xi\| \leq M \|x^{(k)} - \xi\|^p, \forall k \geq k_0,$$

or

$$\|e^{(k+1)}\| \leq M \|e^{(k)}\|^p, \forall k \geq k_0, \text{ where } e^{(k)} = x^{(k)} - \xi.$$

[1]e-mail: raucapbr@doctor.upv.es

Let $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a sufficiently Fréchet differentiable in D, for $\xi + h \in \mathbb{R}^n$ lying in a neighborhood of a solution ξ of $F(x) = 0$, applying Taylor expansion and assuming that the Jacobian matrix $F'(\xi)$ is non singular, we have

$$F(\xi + h) = F'(\xi)\left[h + \sum_{q=2}^{p-1} C_q h^q\right] + O(h^p) \tag{2}$$

where $C_q = (1/q!)[F'(\xi)]^{-1}F^{(q)}(\xi)$, $q \geq 2$. We take into account that $C_q h^q \in \mathbb{R}^n$ since $F^{(q)}(\xi) \in \mathcal{L}(\mathbb{R}^n \times \cdots \times \mathbb{R}^n, \mathbb{R}^n)$ and $[F'(\xi)]^{-1} \in \mathcal{L}(\mathbb{R}^n)$. We can also express F' as

$$F'(\xi + h) = F'(\xi)\left[I + \sum_{q=2}^{p-1} qC_q h^{q-1}\right] + O(h^{p-1}), \tag{3}$$

where I is the identity matrix and $qC_q h^{q-1} \in \mathcal{L}(\mathbb{R}^n)$.

If $X = \mathbb{R}^{n\times n}$ denotes the Banach space of real square matrices of size$n \times n$, we can define $H : X \to X$ such that its Fréchet derivative satisfies:

(*a*) $H'(u)(v) = H_1 uv$, where $H' : X \to \mathcal{L}(X)$ and $H_1 \in \mathbb{R}$,

(*b*) $H''(u, v)(v) = H_2 uvw$, where $H'' : X \times X \to \mathcal{L}(X)$ and $H_2 \in \mathbb{R}$.

By using different techniques: composition of known methods, Jacobian "frozen", weight matrix function procedure, etc. several Newton-type methods of different orders have been designed for improving Newton's scheme. One of the first algorithms was Jarratt's method [4] whose iterative expression is

$$\begin{aligned} y^{(k)} &= x^{(k)} - \frac{2}{3}[F'(x^{(k)})]^{-1}F(x^{(k)}), \\ x^{(k+1)} &= x^{(k)} - [6F'(y^{(k)}) - 2F'(x^{(k)})]^{-1}(3F'(y^{(k)}) - F'(x^{(k)}))[F'(x^{(k)})]^{-1}F(x^{(k)}). \end{aligned} \tag{4}$$

More recently, other authors have constructed different methods for solving nonlinear systems. For example, Cordero et al. in [5] design a three-steps iterative method of order six, by combining Newton and Jarratt's schemes; Behl et al. in [6] also construct a iterative method of order six, with two Jacobian matrix in its iterative expression.

In order to compare the different methods, we analyze the computational effort that they involve, in terms of functional evaluations d and amount of products and quotients op. By using this information, we are going to use two the multidimensional extension of the efficiency index defined by Ostrowski in [7] as $I = p^{1/d}$ and the computational efficiency index CI defined in [5] as $CI = p^{1/(d+op)}$, where p is the order of convergence, d is the number of functional evaluations per iteration and op is the number of products-quotients per iteration.

In this work, a new class of iterative methods for solving nonlinear systems of equations is presented. This family is developed by using a weight function procedure getting 6th-order of convergence. We present the convergence result and an study of the efficiency of our method in comparison with other known ones.

2 The proposed scheme: convergence order and efficiency

Our proposed family, denoted as PS6, is designed by using Newton's scheme and the weight function procedure. Let $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a real sufficiently differentiable function, H a matrix weight function that should be chosen and the three step iterative method

$$\begin{aligned} y^{(k)} &= x^{(k)} - [F'(x^{(k)})]^{-1}F(x^{(k)}), \\ z^{(k)} &= y^{(k)} - H(t^{(k)})[F'(x^{(k)})]^{-1}F(y^{(k)}), \\ x^{(k+1)} &= z^{(k)} - H(t^{(k)})[F'(x^{(k)})]^{-1}F(z^{(k)}), \end{aligned} \tag{5}$$

being $t^{(k)} = I - [F'(x^{(k)})]^{-1}[x^{(k)}, y^{(k)}; F]$.

Theorem 1 *Let $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a sufficiently differentiable function in an open neighborhood D of such that $F(\xi) = 0$, and $H : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ a sufficiently differentiable matrix function. Let us assume that $F'(x)$ is nonsingular at ξ and $x^{(0)}$ is an initial estimatione close enough to ξ. Then, the sequence $\{x^{(k)}\}_{k\geq 0}$ obtained from expression (5) converges to ξ with order 6 if the function H satisfies $H_0 = I$, $H_1 = 2$ and $|H_2| < \infty$, where $H_0 = H(0)$ and I is the identity matrix. The error equation is*

$$\begin{aligned} e^{(k+1)} = \frac{1}{4}\Big[& 120C_2^5 - 22H_2C_2^5 + H_2^2C_2^5 - 24C_2^2C_3C_2 + 2H_2C_2^2C_3C_2 \\ & +4C_3^2C_2 - 20C_3C_2^3 + 2H_2C_3C_2^3\Big]{e^{(k)}}^6 + O({e^{(k)}}^7), \end{aligned}$$

where $C_q = \frac{1}{q!}[F'(\xi)]^{-1}F^{(q)}(\xi)$, $q = 2, 3, ...$

For comparing the efficiency of this family and other known ones, we choose the weight function $H(t) = I + 2t + \frac{1}{2}H_2t^2$, that satisfies the conditions of previous result. In order to calculate de efficiency index I, we recall that the number of functional evaluations of F. F' and first order divided difference $[\cdot, \cdot, F]$ at certain iterates is n, n^2 and $n(n-1)$, respectively. The comparison of efficiency index for the different methods is shown in Table 1. $n.F$, $n.F'$ and $n.[\cdot, \cdot; F]$ denote the number of functional evaluations F, Jacobian matrix F' and divided difference $[\cdot, \cdot; F]$, respectively, per iteration. FE is the number of scalar functions per iteration.

Method	$n.F$	$n.F'$	$n.[\cdot, \cdot; F]$	FE	I
$PS6_{\{H_2\neq 0\}}$	3	1	1	$2n^2 + 2n$	$6^{1/(2n^2+2n)}$
$PS6_{\{H_2=0\}}$	3	1	1	$2n^2 + 2n$	$6^{1/(2n^2+2n)}$
Newton	1	1	0	$n^2 + n$	$2^{1/(n^2+n)}$
Jarratt	1	2	0	$2n^2 + n$	$4^{1/(2n^2+n)}$

Table 1: Efficiency indices of the new and known methods

For calculating the computational efficiency index CI, we take in account that the number of products-quotients required for solving a linear system by Gaussian elimination is $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$

where n is the system size. If required the solution by using LU decomposition of m linear systems with the same matrix of coefficients, then is necessary $\frac{1}{3}n^3+mn^2-\frac{1}{3}n$ products-quotients operations. In addition, n^2 products are necessary for a matrix-vector multiplication and n^2 quotients for first order divided differences. The notation LS($[F'(x)]^{-1}$) and LS($Others$) is the number of lineal systems with $[F'(x)]$ as the matrix of coefficients and with others matrix coefficients, respectively. The comparison of computational efficiency index for the new and known methods is shown in Table 2.

Method	FE	LS($[F'(x)]^{-1}$)	LS($Others$)	$M \times V$	CI
PS6$_1$=PS6$_{\{H_2\neq 0\}}$	$2n^2+2n$	7	0	4	$6^{1/((1/3)n^3+13n^2+(5/3)n)}$
PS6$_2$=PS6$_{\{H_2=0\}}$	$2n^2+2n$	5	0	2	$6^{1/((1/3)n^3+9n^2+(5/3)n)}$
Newton	n^2+n	1	0	0	$2^{1/((1/3)n^3+3n^2+(2/3)n)}$
Jarratt	$2n^2+n$	1	1	1	$4^{1/((2/3)n^3+5n^2+(1/3)n)}$

Table 2: Computational efficiency index of the new and known methods

It is easy to observe that for any value of n, $n \geq 2$, we have

$$CI_{PS6_2} > CI_{PS6_1} > CI_{Newton} > CI_{Jarratt},$$

so, the best method under this point of view is $PS6_2$.

References

[1] Iliev, A. and Kyurkchev, N., Nontrivial Methods in Numerical Analysis: Select Topics in Numerical Analysis, Saarbcken, LAP LAMBERT Academic Publishing, Germany, 2010.

[2] Zhang, Y. and Huang, P., High-precision Time-interval Measurement Techniques and Methods, *Progress in Astronomy*, Volume 24(1), 1–15, 2006.

[3] He, Y. and Ding, C., Using accurate arithmetics to improve numerical reproducibility and stability in parallel applications, *Journal of Supercomputing* Volume(18), 259–277, 2001.

[4] Jarratt, P., Some fourth order multipoint iterative methods for solving equations, *Math. Comput.*, Volume(20), 434–437, 1966.

[5] Cordero, A., Hueso, J.L., Martinez, E. and Torregrosa, J.R., A modified Newton-Jarrat composition, *Numer. Algor.*, Volume(55), 87–99, 2010.

[6] Behl, R., Sarría, Í., González, R. and Magreñán, Á.A., Highly efficient family of iterative methods for solving nonlinear models, *Comput. Appl. Math.*, Volume(346), 110–132, 2019.

[7] Ostrowski, A.M., Solution of Equations and System of Equations, Academic Press, 1966.

6

Adaptive modal methods to integrate the neutron diffusion equation

A. Carreño [♭1], A. Vidal-Ferrándiz[♭], D. Ginestar[♯] and G. Verdú[♭]

(♭) Instituto de Seguridad Industrial: Radiofísica y Medioambiental,
Universitat Politècnica de València,
(♯) Institut Universitari de Matemàtica Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

The distribution of the neutrons inside a reactor core along time can be described by the time dependent multigroup neutron diffusion equation. A finite element method (FEM) is used to discretize the neutron diffusion equation to get a system of semi-ordinary differential equations

$$\begin{aligned} V^{-1}\frac{\mathrm{d}\Phi}{\mathrm{d}t} + L\Phi &= (1-\beta)M\Phi + \sum_{k=1}^{K}\lambda_k^d X C_k, \\ \frac{\mathrm{d}C_k}{\mathrm{d}t} &= \beta_k M_1 \Phi - \lambda_k^d C_k, \qquad k = 1,\ldots,K. \end{aligned} \tag{1}$$

The FEM has been implemented by using the open source finite elements library `Deal.ii`. This system of ordinary differential equations is, in general, stiff. Several approaches have been studied to integrate this time-dependent equation such as the backward differential method, the quasi-static method or the modal method. In this work, we use this last approach that assumes that the solution can be described by the sum of amplitude functions multiplied by shape functions of modes. The shape functions are obtained from the solution of the λ-modes problem

$$L\psi_m = \frac{1}{\lambda_m} M\psi_m. \tag{2}$$

Generally, the eigenfunctions are updated in the time-dependent equations with a fixed time-step size to obtain accurate results [1]. This implies to select a suitable time step before the simulation. In this work, we propose an adaptive time-step control that selects an appropriate step size from a given tolerance. The time selection for the backward method and the quasi-static method is usually based on the local error in the time discretization. In the modal approach, different types of errors are studied for updating the time-step selection.

[1] e-mail:amcarsan@iqn.upv.es

2 The updated modal method

The modal methodology supposes that $\Phi(\vec{r},t)$ admits the following expansion

$$\Phi(\vec{r},t)=\sum_{m=1}^{q} n_m(t)\psi_m(\vec{r}), \tag{3}$$

where $n_m(t)$ are the amplitude coefficients and $\psi_m(\vec{r})$ are the shape functions obtained when problem (2) is solved. Moreover, the matrices L and M of the problem (2) can be expressed as $L=L_0+\delta L$, $M=M_0+\delta M$, respectively, where L_0 and M_0 are the matrices at $t=0$ and in steady-state (M_0 is divided by $k_{eff}=\lambda_1$).
Taking into account these considerations and multiplying by the adjoint eigenfunctions associated with the λ-modes problem, $\psi_l^{\dagger,i}$, over the Eq. (1), one can obtain the system of ODE's

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{N}=\mathbf{T}\mathbf{N}, \tag{4}$$

where

$$\mathbf{N}=\begin{pmatrix} n_1\cdots n_q & c_{11}\cdots c_{q1} & \cdots & c_{1K}\cdots c_{qK}\end{pmatrix}^T,$$

$$\mathbf{T}=\left(\begin{array}{c|ccc} \Lambda^{-1}((1-\beta)I-[\lambda]^{-1}-A^L+(1-\beta)A^M) & \Lambda^{-1}\lambda_1^d & \cdots & \Lambda^{-1}\lambda_K^d \\ \hline \beta_1(I+A^M) & -\lambda_1^d I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_K(I+A^M) & 0 & \cdots & -\lambda_K^d I \end{array}\right),$$

and

$$\begin{aligned} \Lambda_{lm}&=\langle\psi_l^\dagger, V^{-1}\psi_m\rangle, & A_{lm}^L&=\langle\psi_l^\dagger,\delta L\psi_m\rangle,\\ A_{lm}^M&=\langle\psi_l^\dagger,\delta M\psi_m\rangle, & c_{lk}&=\langle\psi_l^\dagger, XC_k\rangle, \end{aligned}$$

The initial conditions values are

$$\begin{aligned} &n_1(0)=1, && n_m(0)=0,\ m=2,\dots,q,\\ &c_{1k}(0)=\frac{\beta_k}{\lambda_k^d}, && c_{mk}(0)=0,\ m=2,\dots,q, \quad k=1,\dots,K, \end{aligned}$$

that are obtained from the equations in the critical state. The ODE is solved with the CVODE module from the open source library SUNDIALS.

In realistic transients, the flux can be suffered extremely spatial variations. Obtaining accurate approximations implies a high number of modes. That means a high computational cost [2]. A solution is a modal methodology where the modes are updated in a certain time interval $[t_i,t_i+\Delta t_i]=[t_i,t_{i+1}]$. In each interval $[t_i,t_{i+1}]$, the neutron diffusion equation can be integrated through the solution of the λ-modes problem associated at time t_i.

The differential equations that are needed to integrate have the same form than the problem without updating (Eq. (4)). The initial conditions for $n_m^{i,\lambda}$ at time t_i must be defined to solve

the differential problem in the interval $[t_i, t_{i+1}]$. For the expansion (3) at $t = t_i$, one could approximate the value of $n_m^{i,\lambda}(t_i)$ as

$$n_m^{i,\lambda}(t_i) \approx \frac{\langle \psi_m^{\dagger,i}, M^i \Phi(t_i) \rangle}{\langle \psi_m^{\dagger,i}, M^i \psi_m^i \rangle},$$

where $\Phi(t_i)$ is obtained from the previous modal step.
The concentration of precursors at time t_i can be computed as

$$c_{l,k}^{i,\lambda}(t_i) = \sum_{m=1}^{q} a_{lm} c_{m,k}^{i-1,\lambda}(t_i), \quad \text{where,} \quad a_{lm} = \frac{\langle \psi_l^{\dagger,i}, M^{i-1} \psi_m^{i-1} \rangle}{\langle \psi_m^{\dagger,i-1}, M^{i-1} \psi_m^{i-1} \rangle}.$$

The λ-modes problem is solved with the block inverse-free preconditioned Arnoldi method, (BIFPAM) (see more details in [5]).

3 Adaptive time-step control

The modes can be updated with fix time-step, but it implies several limitations such as selecting a time-step previously that leads to obtain results with unpredictable errors. In this work, we study an adaptive time-step control. Its implementation requires to define an error estimation due to the modal expansion assumption and a suitable constraint to select the time-step based on the error estimation.

Estimation error The modal error comes (essentially) from the assumption in the modal expansion because the eigenvalue functions do not form a complete basis. Therefore, if there are large variations in the flux along the time, the modal method will obtain large errors. The first estimation is based in the difference between eigenfunctions. The *modal difference error* is defined as

$$\varepsilon_{md} = \max_m \|\psi_m^{i-1} - \psi_m^i\|.$$

The *modal residual error* is computed thought the residual error as

$$\varepsilon_{mr} = \max_m \|L^i \psi_m^{i-1} - \lambda_m^{i-1} M^i \psi_m^{i-1}\|.$$

Finally, we assume that the flux will change according to the changing on the absorption cross-section Σ_{a1}. We define the *cross-section perturbation error* as

$$\varepsilon_{sa} = \sum_c \|\Sigma_{a1}^{i-1}(c) - \Sigma_{a1}^i(c)\|,$$

where c is a cell of the reactor.

Control Algorithm Two strategies based on the error in the previous step are proposed. The *Fixed Error Control* strategy defined as

$$\Delta t_i = \begin{cases} \Delta t_{i-1} * 2, & \varepsilon < min_{le}, \\ \Delta t_{i-1}, & min_{le} < \varepsilon < max_{le}, \\ \Delta t_{i-1} \,/\, 2, & max_{le} < \varepsilon. \end{cases}$$

The *Adaptive Error Control*, based on the control algorithms defined for differential methods. The step Δt_i is computed as

$$\Delta t_i = \Delta t_{i-1} \min\{2.0, \max\{0.5, \sqrt{3.0/\varepsilon}\}\},$$

4 Numerical results

The Langenbuch reactor [4] is chosen to study the performance of the adaptive modal methodology proposed with a local perturbation. This reactor has two sets of control rods that define the transient (Fig. 1.a). It starts with the withdrawal of one bar of C1 at 10cm/s. The rest of C1 is inserted at 3cm/s over $7.5 < t < 47.5$s. C2 is inserted at 3cm/s over $7.5 < t < 47.5$s. The evolution of the global power is represented in Figure 1. The solutions are compared with the solutions obtained with a backward differential method [3] (Fig. 1.b).

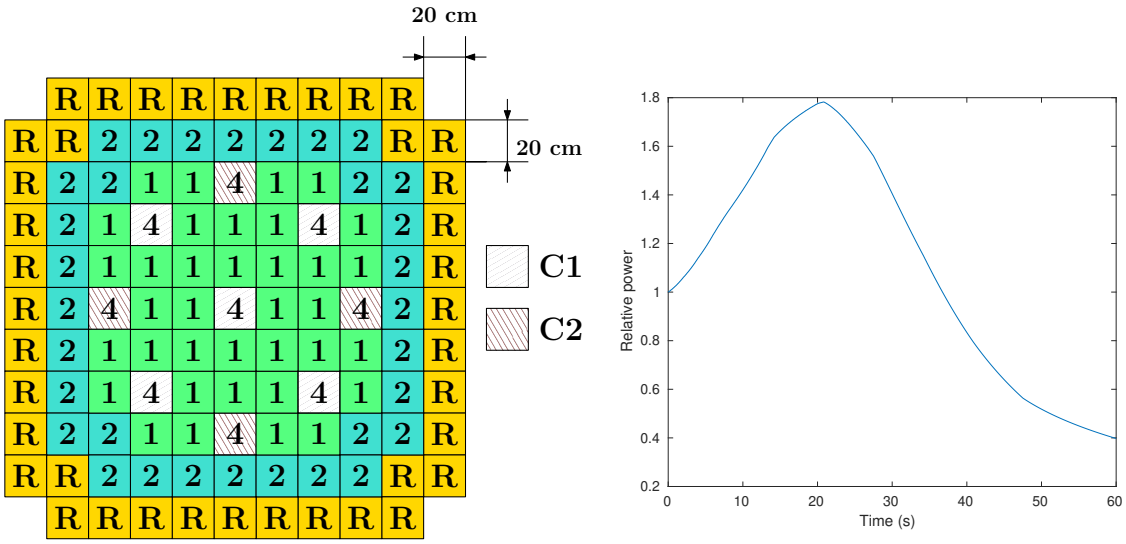


Figure 1: Geometry and global power of the Langenbuch transient.

Table 1 shows that better approximations are obtained if the step of updating is smaller. However, the computational times in these cases are also higher. Moreover, one can observe that using a high number of modes in the expansion gives more accurate results. A combination between a reasonable number of modes and a not small time-step is the most effective option. However, these parameters are dependent on the transient.

N. eigs. (q)	Updated	Error	CPU Time
1	1.0s	4.9e-03	38min
1	5.0s	1.8e-02	9min
3	5.0s	1.5e-02	10min
3	10.0s	3.5e-02	6min

Table 1: Performance of the Updated modal method with fixed time-step.

The global error is represented along the time (Fig. 2). It is observed that the large errors are produced when the local perturbation is applied and before to the updating of the modes. Fig. 2 represents the radial profile of the error at $t = 2.1$ in the case with 3 eigenvalues and time-step equal to 5 s. It is observed that this is focused around the control rod that is withdrawal at this moment. The same conclusions are deduced with other settings.

Error time-step	Control	Error	CPU Time
ε_{md}	Fixed	9e-03	53.0min
ε_{md}	Adaptive	9e-03	100.2min
ε_{mr}	Fixed	7e-03	8.0min
ε_{mr}	Adaptive	1e-02	6.0min
ε_{sa}	Fixed	8e-03	7.2min
ε_{sa}	Adaptive	1e-02	6.3min

Table 2: Errors and CPU time obtained with the adaptive time-step modal method.

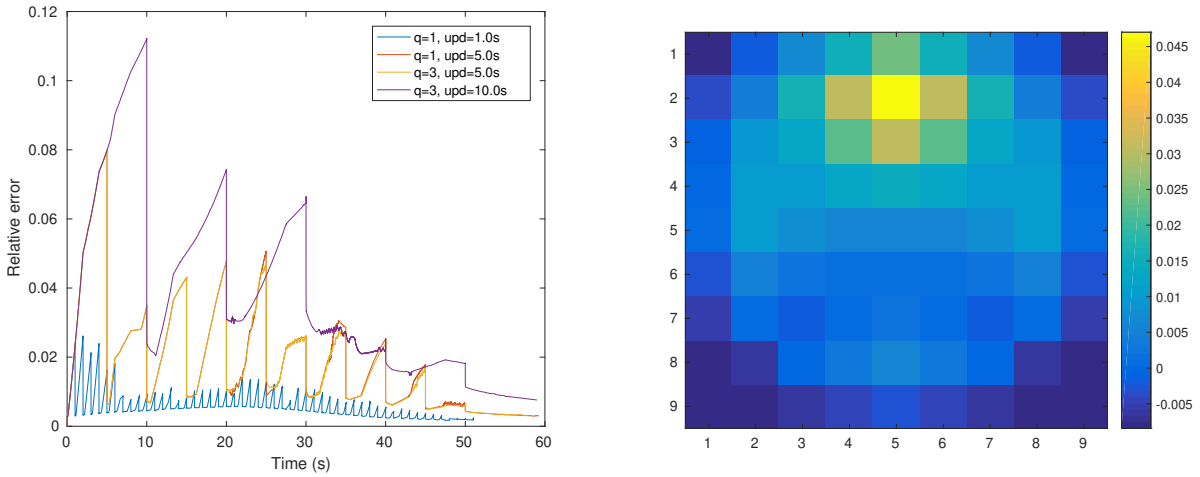


Figure 2: Error in the Langenbuch transient with fixed time-step.

Table 2 shows the error and CPU times obtained with the different error estimations and control errors obtained with 3 eigenvalues. First, one can observed that the modal difference error (md) is not a very efficient technique because it needs to compute the modes for estimating the error (that is very expensive). Regarding to other error estimations, there are not big differences between them. If now, the type of control error is compared the adaptive gives similar approximations than the fixed control but in less time.

Finally, we compare the error between the adaptive time-step modal method and the fixed time-step modal method with 3 eigenvalues and the same initial time-step (5.0 s). More accurate results are obtained with the adaptive method in less time. Moreover, if this error is analyzed over the time in both cases, an error more distributed is obtained by using the adaptive time-step control.

References

[1] Miró, R., Ginestar, D., Verdú, G. and Hennig, D., A nodal modal method for the neutron diffusion equation. Application to BWR instabilities analysis, *Annals of Nuclear Energy*, 29(10): 1171-1194, 2002.

[2] Carreño, A., Vidal-Ferràndiz, A., Ginestar, D. and Verdú, G., Modal methods for the neutron diffusion equation using different spatial modes, *Progress in Nuclear Energy*, 115: 181-193, 2019.

[3] Vidal-Ferràndiz, A., Fayez, R., Ginestar, D. and Verdú, G., Moving meshes to solve the time-dependent neutron diffusion equation in hexagonal geometry, *Journal of Computational and Applied Mathematics*, 291: 197-208, 2016.

[4] Langenbuch, S., Maurer, W. and Werner, W., Coarse-mesh flux-expansion method for the analysis of space-time effects in large light water reactor cores, *Nuclear Science and Engineering*, 63(4): 437-456, 1977.

[5] Vidal-Ferràndiz, A., Carreño, A., Ginestar, D. and Verdú, G., A block Arnoldi method for the SPN equations, *International Journal of Computer Mathematics*, 1-17, 2019.

7

Numerical integral transform methods for random hyperbolic models

M.-C. Casabán[♭1], R. Company[♭] and L. Jódar[♭]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

This work deals with the construction of analytic-numerical solutions, in the mean square sense [1], of the random heterogeneous telegraph type problem

$$u_{tt}(x,t) = (k(x)\,u_x(x,t))_x + a(x)\,u(x,t) + \psi(x,t)\,, \quad x>0,\ t>0\,, \tag{1}$$
$$u(0,t) = g_1(t)\,, \tag{2}$$
$$u_x(0,t) = g_2(t)\,, \tag{3}$$
$$u(x,0) = g(x)\,, \tag{4}$$

where $a(x)$, $k(x)$, $\psi(x,t)$, $g_1(t)$, $g_2(t)$ and $g(x)$ are stochastic processes (s.p.'s) with a finite degree of randomness [1].

Efficient methods for solving numerically deterministic problems such as finite-difference methods become unsuitable for the random case because of the computation of the expectation and the variance of the approximation solution s.p. The drawbacks are essentially of computational complexity such as the handling of big random matrices which appear throughout the iterative levels of the discretization steps and the necessity to store the information of all the previous levels of the iteration process. Then, they motive the search of non iterative alternatives. In this sense, this paper provides an approximation solution s.p. of the problem (1)–(4) which combines the random Fourier sine transform, the Gauss-Laguerre quadrature rule and the Monte Carlo method.

2 Gauss-Laguerre solution of a random hyperbolic model

The construction of an approximated solution s.p. of the problem (1)–(4) will be in two-stages. Firstly, using the Fourier sine transform, an infinite integral form solution of the theoretical

[1]email: macabar@imm.upv.es

solution is obtained. Then, using random Gauss-Laguerre quadrature formulae a random numerical solution is represented that is further computer by means of Monte Carlo simulations at appropriated root points of the Laguerre polynomials.
Let $V(x)(\xi) = \mathfrak{F}_s[u(x,\cdot)](\xi)$ be the Fourier sine transform of the unknown $u(x,\cdot)$:

$$V(x)(\xi) = \mathfrak{F}_s[u(x,\cdot)](\xi) = \int_0^{+\infty} u(x,t)\, \sin(\xi\, t)\, \mathrm{dt}\,, \quad \xi > 0\,,\ x > 0\,. \tag{5}$$

Let us denote

$$\begin{aligned} G_1(\xi) &= \mathfrak{F}_s[u(0,\cdot)](\xi) = \mathfrak{F}_s[g_1(t)](\xi)\,, \quad \xi > 0\,, && (6)\\ G_2(\xi) &= \mathfrak{F}_s[u_x(0,\cdot)](\xi) = \mathfrak{F}_s[g_2(t)](\xi)\,, \quad \xi > 0\,, && (7)\\ \Psi(x)(\xi) &= \mathfrak{F}_s[\psi(x,\cdot)](\xi)\,, \quad x > 0,\ \xi > 0\,. && (8) \end{aligned}$$

Let us assume that the s.p.'s $k(x)$, $a(x)$, $\psi(x,t)$, $g_1(t)$, $g_2(t)$ and $g(x)$ of problem (1)–(4) are mean four (m.f.) continuous with a finite degree of randomness. Let $k(x)$ be a positive s.p. 4-differentiable and let $\psi(x,t)$, $g_1(t)$, $g_2(t)$ be m.f. absolutely integrable s.p.'s in $t > 0$. By applying random Fourier sine transform to problem (1)–(4) and using the properties of the random Fourier sine transform, [2] , one gets, for $\xi > 0$ fixed

$$\frac{d^2}{dx^2}(V(x))(\xi) + \frac{k'(x)}{k(x)}\,\frac{d}{dx}(V(x))(\xi) + \frac{a(x)+\xi^2}{k(x)}\,V(x)(\xi) = \frac{\Psi(x)(\xi) - \xi\, g(x)}{k(x)}\,, \tag{9}$$

together with

$$\begin{aligned} &V(0)(\xi) = G_1(\xi)\,,\\ &\frac{d}{dx}(V(0))(\xi) = G_2(\xi)\,. \end{aligned} \tag{10}$$

Solution of problem (9)–(10) is the first component of the solution of extended random linear differential system, $V(x)(\xi) = [1,\ 0]\, X(x)(\xi)$,

$$\left.\begin{aligned} X'(x)(\xi) &= L(x)(\xi)\, X(x)(\xi) + B(x)(\xi)\,, \quad x > 0\,,\\ X(0)(\xi) &= Y_0(\xi)\,, \end{aligned}\right\} \tag{11}$$

where

$$\left.\begin{aligned} L(x)(\xi) = \begin{bmatrix} 0 & 1 \\ -\dfrac{\xi^2 + a(x)}{k(x)} & -\dfrac{k'(x)}{k(x)} \end{bmatrix},\ B(x)(\xi) = \begin{bmatrix} 0 \\ \dfrac{\Psi(x)(\xi) - \xi\, g(x)}{k(x)} \end{bmatrix},\\ Y_0(\xi) = \begin{bmatrix} G_1(\xi) \\ G_2(\xi) \end{bmatrix}. \end{aligned}\right\} \tag{12}$$

Assuming that 4-s.p.'s $a(x)$, $k(x)$ and $k'(x)$ satisfy the moment condition

$$\mathbb{E}\left[|s(x)|^r\right] \leq m\, h^r < +\infty\,, \quad \forall\, r \geq 0\,, \tag{13}$$

for every $x > 0$, it is guaranteed that the entries of the matrix s.p. $L(x)(\xi) \in L_4^{2\times 2}(\Omega)$, for $\xi > 0$ fixed, satisfy condition (13). Condition (13) guarantees that $L(x)(\xi)$ is 4-locally

absolutely integrable. Furthermore, it is verifies that vector s.p.'s both $B(x)(\xi)$ and $Y_0(x)(\xi)$ lie in $L_4^{2\times 1}(\Omega)$ and they are absolutely integrables in $x \in [0, +\infty)$. By using random inverse Fourier sine transform to $V(x)(\xi)$ one gets

$$u(x,t) = \frac{2}{\pi}\int_0^\infty V(x)(\xi)\,\sin(\xi\,t)\,\mathrm{d}\xi = \frac{2}{\pi}\int_0^\infty X_1(x)(\xi)\,\sin(\xi\,t)\,\mathrm{d}\xi\,, \tag{14}$$

where $X_1(x)(\xi) = [1,\ 0]\,X(x)(\xi)$. Now, taking advantage of the Gauss-Laguerre quadrature formula of degree N, see page 890 of [3], for a s.p. $\mathcal{J}(\xi) \in L_2(\Omega)$ being m.f.-absolutely integrable respect to $\xi > 0$, we can consider the following numerical approximation for each event $\omega \in \Omega$

$$I_N^{\text{G-L}}[\mathcal{J}](\omega) = \sum_{j=1}^N \nu_j\ \mathcal{J}(\vartheta_j;\omega)\,, \qquad \nu_j = \frac{\vartheta_j}{\left[(N+1)\,L_{N+1}(\vartheta_j)\right]^2}\,, \tag{15}$$

where ϑ_j is the j-th root of the deterministic Laguerre polynomial, $L_N(\vartheta)$, of degree N and ν_j is the weight. This quadrature formula is going to be applied to the r.v. $u(x,t)$ given by (14) taking

$$\mathcal{J}(\xi) = \mathcal{J}(x,t,\xi) = \frac{2}{\pi}\,X_1(x)(\xi)\,\sin(\xi\,t)\,e^{\xi}\,.$$

Given the degree N, let us denote by $u_N^{\text{G-L}}(x,t)$ the Gauss-Laguerre s.p. approximation of degree N of the exact solution s.p. $u(x,t)$ of the random problem (1)–(4), evaluated at (x,t) and expressed as the r.v.

$$u_N^{\text{G-L}}(x,t) = \frac{2}{\pi}\sum_{j=1}^N \nu_j\,\sin(\vartheta_j\,t)\,e^{\vartheta_j}\,X_1(x)(\vartheta_j)\,. \tag{16}$$

The exact solution $X_1(x)(\vartheta_j)$ is going to be obtained using Monte Carlo simulation because it is not available. We denoted by $\mathbb{E}_{MC}^K[\bar{X}_1(x)(\vartheta_j)]$ and $\mathrm{Cov}_{MC}^K\left[\bar{X}_1(x)(\vartheta_j)\,,\bar{X}_1(x)(\vartheta_\ell)\right]$ the expectation and the covariance, respectively, of K number of realizations used in the Monte Carlo (MC) simulation and $\bar{X}_1(x)(\vartheta_j)$ the deterministic numerical solution obtained after taking K realizations. Thus the final expressions for the approximations of the expectation and the variance of the solution s.p. take the form

$$\mathbb{E}[u_N^{\text{G-L}}(x,t)] \approx \mathbb{E}[u_{N,K}^{\text{G-L}}(x,t)] = \frac{2}{\pi}\sum_{j=1}^N \nu_j\,\sin(\vartheta_j\,t)\,e^{\vartheta_j}\,\mathbb{E}_{MC}^K[\bar{X}_1(x)(\vartheta_j)]\,, \tag{17}$$

$$\mathrm{Var}\left[u_N^{\text{G-L}}(x,t)\right] \approx \mathrm{Var}\left[u_{N,K}^{\text{G-L}}(x,t)\right] = \left(\frac{2}{\pi}\right)^2 \sum_{j=1}^N\sum_{\ell=1}^N \nu_j\,\nu_\ell\,\sin(\vartheta_j\,t)\,\sin(\vartheta_\ell\,t)\,e^{\vartheta_j+\vartheta_\ell}\,\mathrm{Cov}_{MC}^K\left[\bar{X}_1(x)(\vartheta_j)\,,\bar{X}_1(x)(\vartheta_\ell)\right]\,. \tag{18}$$

3 Numerical example

Consider the random heterogeneous telegraph type problem (1)–(4) with the following input data having a finite degree of randomness

$$\left.\begin{array}{l} k(x) = 1 + b\,\cos(\pi\,x)\,,\ a(x) = e^{-ax}\,,\ \psi(x,t) = e^{-(x+t)} \\ g_1(t) = 0\,,\ g_2(t) = 0\,,\ g(x) = 0 \end{array}\right\}\,, \tag{19}$$

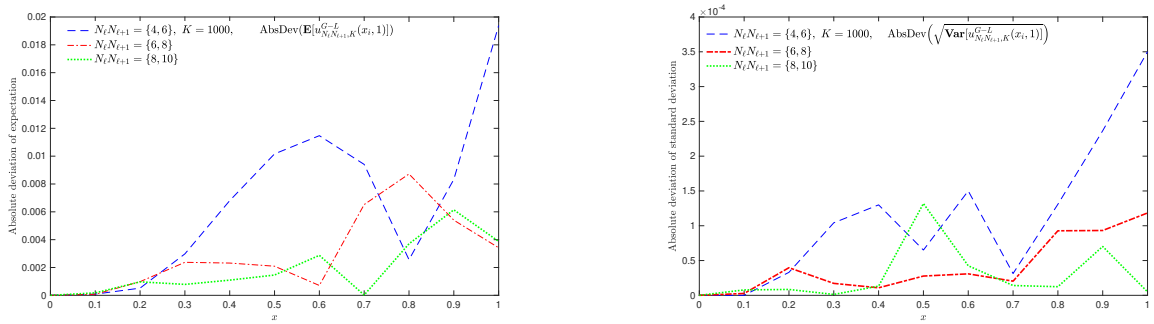
where parameters a and b are assumed to be both independent r.v.'s, specifically, a has a uniform distribution giving values in $[0,1]$, that is, $a \sim Un(0,1)$, and $b > 0$ has an exponential distribution of parameter 2 truncated on the interval $[0.1, 0.2]$, that is, $b \sim Exp_{[0.1,0.2]}(2)$. Then it is verified that s.p.'s $k(x)$, $a(x)$ and the deterministic functions $\psi(x,t)$, $g_1(t)$, $g_2(t)$ and $g(x)$ are 4-continuous and 4-absolutely integrable with respect to the time variable those depending on t. Furthermore, $k(x)$ is positive and 4-differentiable.

To study the numerical convergence of the approximations of both the expectation and the standard deviation we have studied the behaviour of their root mean square deviations (RMSD) and the absolute deviations (AbsDev), that is

$$\begin{aligned}
\text{RMSD}\left[\mathbb{E}[u^{\text{G-L}}_{N,K_\ell K_{\ell+1}}(x_i,t)]\right] &= \sqrt{\tfrac{1}{(n+1)}\sum_{\ell=0}^{n}\left(\mathbb{E}[u^{\text{G-L}}_{N,K_{\ell+1}}(x_i,t)] - \mathbb{E}[u^{\text{G-L}}_{N,K_\ell}(x_i,t)]\right)^2}, \\
\text{RMSD}\left[\sqrt{\text{Var}[u^{\text{G-L}}_{N,K_\ell K_{\ell+1}}(x_i,t)]}\right] &= \sqrt{\tfrac{1}{(n+1)}\sum_{\ell=0}^{n}\left(\sqrt{\text{Var}[u^{\text{G-L}}_{N,K_{\ell+1}}(x_i,t)} - \sqrt{\text{Var}[u^{\text{G-L}}_{N,K_\ell}(x_i,t)}\right)^2}, \\
\text{AbsDev}\left(\mathbb{E}[u^{\text{G-L}}_{N_\ell N_{\ell+1},K}(x,t)]\right) &= \left|\mathbb{E}[u^{\text{G-L}}_{N_{\ell+1},K}(x,t)] - \mathbb{E}[u^{\text{G-L}}_{N_\ell,K}(x,t)]\right|, \\
\text{AbsDev}\left(\sqrt{\text{Var}[u^{\text{G-L}}_{N_\ell N_{\ell+1},K}(x,t)]}\right) &= \left|\sqrt{\text{Var}[u^{\text{G-L}}_{N_{\ell+1},K}(x,t)]} - \sqrt{\text{Var}[u^{\text{G-L}}_{N_\ell,K}(x,t)]}\right|,
\end{aligned}$$

in two stages. Firstly, varying the number K of realizations in the Monte Carlo method but considering fixed N in the Gauss-Laguerre quadrature rule and secondly, varying N but considering the number of realizations K fixed. Table 1 and Figure 1 illustrated the numerical convergence of our approximations.

Figures



(a) (b)

Figure 1: (a): Comparative graphics of the absolute deviations for successive approximations to the expectation $\mathbb{E}[u^{\text{G-L}}_{N_\ell N_{\ell+1},K}(x_i,1)]$. (b): Comparative graphics of the absolute deviations for successive approximations to the standard deviation $\sqrt{\text{Var}[u^{\text{G-L}}_{N_\ell N_{\ell+1},K}(x_i,1)]}$. Both graphics correspond to the time $t = 1$ on the spatial interval $0 \le x \le 1$, $K = 1000$ realizations and the degrees $N = \{4,6,8,10\}$ for the Laguerre polynomials.

Tables

$K_\ell K_{\ell+1}$	RMSD $\left[\mathbb{E}[u^{\text{G-L}}_{6,K_\ell K_{\ell+1}}(x_i,1)]\right]$	RMSD $\left[\sqrt{\text{Var}[u^{\text{G-L}}_{6,K_\ell K_{\ell+1}}(x_i,1)]}\right]$
K_0K_1	$2.81922e-05$	$2.91484e-05$
K_1K_2	$1.96565e-05$	$1.18180e-05$
K_2K_3	$1.11618e-05$	$2.49163e-06$
K_3K_4	$1.07918e-05$	$4.96128e-06$
K_4K_5	$3.59937e-06$	$2.83452e-06$

Table 1: Values of the RMSDs for the approximations of the expectation, RMSD $\left[\mathbb{E}^{\text{G-L}}_{N,K_\ell K_{\ell+1}}(x_i,t)\right]$, and the standard deviation, RMSD $\left[\sqrt{\text{Var}^{\text{G-L}}_{N,K_\ell K_{\ell+1}}(x_i,t)}\right]$, at $t=1$ on the spatial domain $0 \leq x \leq 1$, $N=6$ the degree of the Laguerre polynomial and the realizations $K_0=2500$, $K_1=5000$, $K_2=10^4$, $K_3=2\times10^4$, $K_4=4\times10^4$ and $K_5=5\times10^4$.

Acknowledgements

This work has been partially supported by the Spanish Ministerio de Economía y Competitividad grant MTM2017-89664-P.

References

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

[2] Casabán, M.C., Cortés, J.C. and Jódar, L., Solving linear and quadratic random matrix differential equations: A mean square approach. *Appl. Math. Model.*, 40: 9362–9377, 2016.

[3] Abramowitz, M. and Stegun, I.A., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, New York, Dover Publications, Inc., 1972.

8

Nonstandard finite difference schemes for coupled delay differential models

M.A. Castro[♭1], M.A. García[♭], J.A. Martín[♭] and F.Rodríguez[♭]

(♭) Department of Applied Mathematics,
University of Alicante.

1 Introduction

Delay differential equations (DDE) have been increasingly used in the last decades in the mathematical modelling of scientific and technical problems that exhibit the presence of time lags, memory effects, or hereditary characteristics [1]. Nonstandard finite difference (NSFD) numerical schemes [2] have also been widely applied to obtain numerical solutions for ordinary or partial differential problems. NSFD methods might provide exact numerical solutions for particular equations and they can be able to compete in accuracy with standard methods and derive numerical solutions that are dynamically consistent with the original differential problems.

The construction of NSFD schemes for delay differential models has not been much developped. In [3], a NSFD method was proposed for the scalar linear delay problem

$$\begin{aligned} x'(t) &= \alpha x(t) + \beta x(t-\tau), \qquad t > 0, && (1)\\ x(t) &= f(t), \qquad -\tau \le t \le 0, && (2) \end{aligned}$$

which was exact only in the initial time interval $0 \le t \le \tau$. More recently, in [4], an exact scheme for the problem (1)-(2) was constructed, valid in the whole domain of definition, and a family of increasing order NSFD schemes was defined.

In the this work, NSFD schemes are proposed for the coupled linear delay system

$$\begin{aligned} X'(t) &= AX(t) + BX(t-\tau), \qquad t > 0, && (3)\\ X(t) &= F(t), \qquad -\tau \le t \le 0, && (4) \end{aligned}$$

where $X(t)$ and $F(t)$ are d-dimensional vector functions, and A and B are $d \times d$ matrices, in general not simultaneously diagonalizable. Asymptotic and delay stability properties of the new schemes are illustrated using numerical experiments.

[1]e-mail: ma.castro@ua.es

2 Results

The NSFD schemes proposed in this work are based on the next expression for the exact solution of problem (3)-(4).

Lemma 1 *Consider problem (3)-(4) with A and $I+C$ inversible, where $C = A^{-1}B$. The solution of (3)-(4) for any continuous initial function $F(t)$ is given by*

$$\begin{aligned} X(t) &= (G(t) + G(t-\tau)C)\,(I+C)^{-1}F(0) \\ &+ \int_{-\tau}^{0} G'(t-\tau-s)(I+C)^{-1}CF(s)ds, \end{aligned} \tag{5}$$

where $G(t)$ is the solution of the matrix delay initial value problem

$$G'(t) = AG(t) + BG(t-\tau), \qquad t > 0, \tag{6}$$

$$G(t) = I, \qquad \tau \le t \le 0. \tag{7}$$

From the expression given in Lemma 1, the next exact numerical solution is obtained.

Theorem 1 *Consider a uniform mesh of size h such that $Nh = \tau$, for some integer $N > 0$, and write $t_n \equiv nh$, and $X_n \equiv X(t_n)$, for $n \ge -N$. Then, the numerical solution given by $X_n = F(t_n)$, for $-N \le n \le 0$, and for $(m-1)\tau \le nh < m\tau$ and $m \ge 1$ by*

$$\begin{aligned} X_n &= (G(t_n) + G(t_n-\tau)C)\,(I+C)^{-1}F(0) \\ &+ \int_{-\tau}^{0} G'(t_n-\tau-s)(I+C)^{-1}CF(s)ds, \end{aligned} \tag{8}$$

coincides with the exact solution of (3)-(4) for the mesh points.

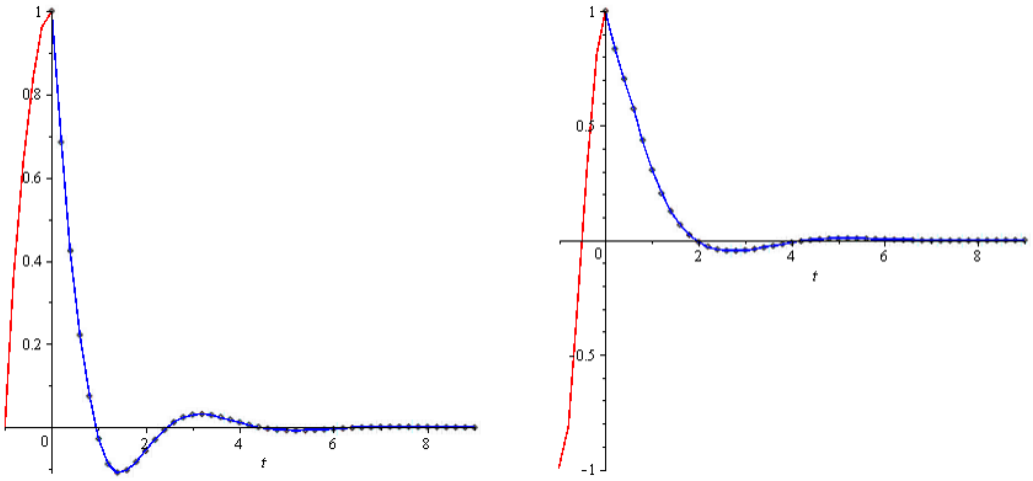


Figure 1: Exact solutions (lines) and numerical solutions provided by the exact scheme (points) for the problem in Example 1. Left: First component, $X_1(t)$. Right: Second component, $X_2(t)$.

Example 1 *Figure 1 shows the continuous solution given by Lemma 1 (lines) and the exact numerical solution of Theorem 1 with* $N = 5$ *(points), for the problem (3)-(4) with parameters* $\tau = 1$ *and*

$$A = \begin{pmatrix} -1 & -1/2 \\ -1/2 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} -1/2 & 1/5 \\ 1/5 & -1/2 \end{pmatrix}, \quad F(t) = \begin{pmatrix} 1 - t^2 \\ \cos(\pi t) \end{pmatrix}.$$

From the exact expression of Theorem 1, the following family of numerical schemes of increasing order can be derived.

Theorem 2 *Fix* $M \geq 1$*, and compute the numerical solution of (3)-(4) in the intervals* $(m-1)\tau \leq nh \leq m\tau$*, for* $0 \leq m \leq M$*, with the exact method given in Theorem 1 or with any other numerical method of order at least* $O(h^{M+1})$*. Then, for* $m \geq M+1$ *and* $(m-1)\tau \leq nh < m\tau$*, the expression*

$$X_{n+1} = e^{Ah} X_n + \sum_{p=1}^{M} \sum_{r=p}^{M} \frac{h^r}{r!} K_{r,p} X_{n-pN}, \tag{9}$$

where the matrix constants $K_{r,p}$ *are defined by*

$$\begin{aligned} K_{r,s} = 0, \quad s > r, \qquad K_{r,0} = A^r, \quad r \geq 0 \\ K_{r+1,s} = K_{r,s-1} B + K_{r,s} A, \quad 1 \leq s \leq m-1, \\ K_{r+1,m} = K_{r,m-1}, \end{aligned} \tag{10}$$

defines a nonstandard numerical scheme of global order M*.*

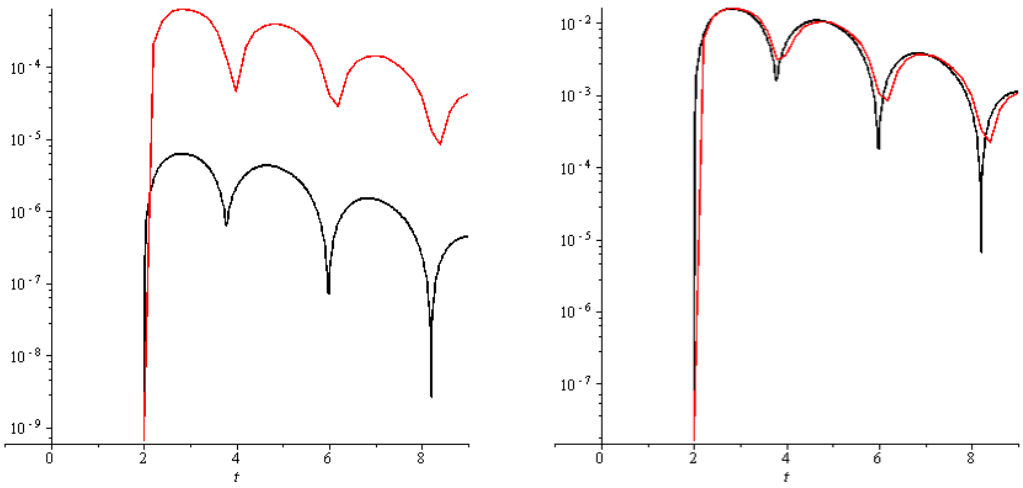


Figure 2: Left: Absolute error (log-scale) of the numerical solutions for the first component of Example 1 provided by the nonstandard scheme of second order ($M = 2$) defined in Theorem 2, for two different mesh sizes ($h = 0.2$, red, and $h = 0.02$, black). Right: Errors divided by h^2.

Numerical experiments show that the approximated solutions obtained using the NSFD schemes defined in Theorem 2 preserve basic dynamical properties of the corresponding exact continuous solution.

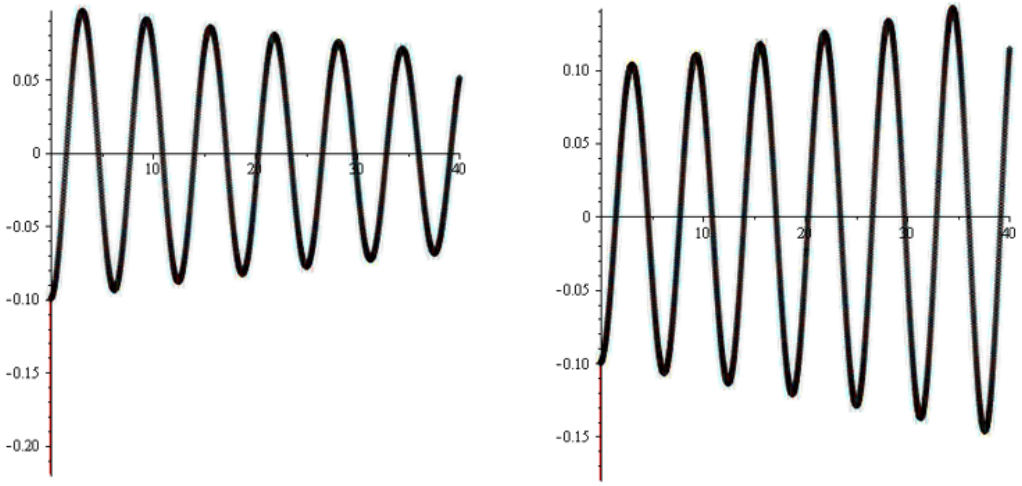


Figure 3: Numerical solutions for the first component of Example 2 provided by the nonstandard scheme of third order ($M = 3$) defined in Theorem 2, for two different delay values. Left: $\tau = 0.12$. Right: $\tau = 0.08$.

Example 2 *Figure 3 illustrates the preservation of asymptotic stability by the numerical solutions obtained with the third order ($M = 3$) NSFD scheme defined in Theorem 2, with $N = 5$ (points), for the first component of problem (3)-(4) with parameters*

$$A = \begin{pmatrix} 0 & 1 \\ -2 & 0.1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad F(t) = \begin{pmatrix} 1 - t^2 \\ (t+1)^2 \end{pmatrix},$$

and two different delay values, $\tau = 0.12$ (left), and $\tau = 0.08$ (right). The exact solution for this problem is asymptotically stable if and only if $\tau \in (0.1002, 1.7178)$.

3 Conclusions

We have proposed a family of NSFD schemes of increasing order for the couple system of delay initial value problem (3)-(4). Numerical experiments have shown that the proposed schemes exhibit the right order, as stated in Theorem 1, and preserve asymptotic stability and other dynamical properties, as oscillation and positivity, of the exact continuous solutions.

The general expressions presented in Theorems 1 and 2 can be notably simplified when the coefficient matrices A and B commute, allowing to derive formal proofs of the dynamic consistency properties of the new NSFD schemes.

We note that higher order linear delay differential equations can be transformed into first order linear systems, and hence the results of this work can also be applied to them. In particular, it is expected that special forms of the so obtained systems may led to simplified expressions for the NSFD schemes, thus allowing a more detailed analysis of their properties.

References

[1] Kolmanovskii, V. and Myshkis, A., *Introduction to the Theory and Applications of Functional Differential Equations*. Kluwer Academic Publishers, Dordrecht, 1999.

[2] Mickens, R.E., *Nonstandard Finite Difference Models of Differential Equations*. World Scientific, Singapore, 1994.

[3] Garba, S.M., Gumel, A.B., Hassan, A.S. and Lubuma, J.M.-S. Switching from exact scheme to nonstandard finite difference scheme for linear delay differential equation, *Applied Mathematics and Computation* 258: 388–403, 2015.

[4] García, M.A., Castro, M.A., Martín, J.A. and Rodríguez, F., Exact and nonstandard numerical schemes for linear delay differential models, *Applied Mathematics and Computation* 338: 337-345, 2018.

9

Semilocal convergence for new Chebyshev-type iterative methods

Abhimanyu Kumar ♭, D.K. Gupta♮, Eulalia Martínez♯, Jose L. Hueso♯ and Fabricio Cevallos*[1]

(♭) Department of Mathematics,
Lalit Narayan Mithila University,
(♮) Department of Mathematics,
Indian Institute of Technology Kharagpur,
(♯) Instituto de Matemática Multidisciplinar,
Universitat Politècnica de València,
(∗) Facultad de Ciencias Económicas,
Universidad Laica Eloy Alfaro de Manabí.

1 Introduction

In this paper, the convergence of improved Chebyshev-Secant-type iterative methods are studied for solving nonlinear equations in Banach space settings. Its semilocal convergence is established using recurrence relations under weaker continuity conditions on first order divided differences. Convergence theorems are established for the existence-uniqueness of the solutions.

Consider approximating a locally unique solution ρ^* of

$$\mathrm{F}(x) = 0, \tag{1}$$

where F is a continuous nonlinear operator defined on a non-empty open convex subset D of a Banach space X with values in another Banach space Y. This is one of the most important problems in applied mathematics and engineering.

The next family of iterative methods used for the solution of (1) is known as the Chebyshev-Secant-type methods (CSTM).

$$\begin{aligned} y_k &= x_k - [x_{k-1}, x_k; \mathrm{F}]^{-1}\mathrm{F}(x_k), \\ z_k &= x_k + \alpha(y_k - x_k), \\ x_{k+1} &= x_k - [x_{k-1}, x_k; \mathrm{F}]^{-1}(\beta\mathrm{F}(x_k) + \gamma\mathrm{F}(z_k)), \end{aligned} \tag{2}$$

where $x_{-1},\ x_0 \in \mathrm{D}$ are two starting iterates and $[x, y; \mathrm{F}] \in \mathrm{L}(\mathrm{X}, \mathrm{Y})$ satisfies $[x, y; \mathrm{F}](x - y) = \mathrm{F}(x) - \mathrm{F}(y)$ for $x, y \in \mathrm{D}$ and $x \neq y$, for $x = y$, $[x, y; \mathrm{F}] = \mathrm{F}'(x)$. Here, α, β and γ are nonnegative real parameters carefully chosen so that the sequence $\{x_k\}$ converges to ρ^*.

[1]e-mail: alfa2205@gmail.com

The improved Chebyshev-Secant-type method (ICSTM) proposed by us is given for $k \geq 0$ by

$$\begin{aligned} x_{k+1} &= x_k - \mathrm{B}_k^{-1}\mathrm{F}(x_k), \quad \mathrm{B}_k = [x_k, y_k; \mathrm{F}], \\ z_k &= x_k + \alpha(x_{k+1} - x_k), \\ y_{k+1} &= x_k - \mathrm{B}_k^{-1}(\beta\mathrm{F}(x_k) + \gamma\mathrm{F}(z_k)), \end{aligned} \tag{3}$$

where $x_0, y_0 \in \mathrm{D}$ are two starting iterates and $\alpha,\ \beta$ and γ are nonnegative real parameters. Considering $\alpha = \beta = \gamma = 1$ we obtain the double step Secant method [1, 2] with order of convergence $1 + \sqrt{2}$. It can be easily seen that the number of functions evaluations and the corresponding divided differences used in CSTM and ICSTM are equal. The importance of the ICSTM lies in the fact that for $\alpha = \beta = \gamma = 1$, its convergence order is $1 + \sqrt{2}$, while the convergence order of the CSTM is 2.

2 Semilocal convergence of ICSTM

In this section, the semilocal convergence of ICSTM for solving (1) is established. Let $\mathcal{B}(x, r)$ and $\overline{\mathcal{B}}(x, r)$ denote open and closed balls with center at x and radius r, respectively. For suitably chosen initial approximations x_0 and y_0, we define a class $\mathrm{S}(\Theta, \delta, \eta, \sigma)$, where $\Theta > 0,\ \delta > 0,\ \eta > 0$ are some positive real numbers and σ is to be defined. The triplet $(\mathrm{F}, x_0, y_0) \in \mathrm{S}(\Theta, \delta, \eta, \sigma)$ if

[C_1] $\|x_0 - y_0\| \leq \Theta$ for $x_0, y_0 \in \mathrm{D}$.

[C_2] $\mathrm{B}_0^{-1} \in \mathrm{L}(\mathrm{Y}, \mathrm{X})$ such that $\|\mathrm{B}_0^{-1}\| \leq \delta$.

[C_3] $\|\mathrm{B}_0^{-1}\mathrm{F}(x_0)\| \leq \eta$.

[C_4] $\|([x, y; \mathrm{F}] - [u, v; \mathrm{F}])\| \leq \sigma(\|x - u\|, \|y - v\|)$, where $\sigma : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous and non decreasing function in its both arguments for $x, y, u, v \in \mathrm{D}$.

[C_5] $(1 - \beta) = (1 - \alpha)\gamma$ and $\alpha \in (0, 1]$.

[C_6] The equation

$$(1 - g(t))t - \eta = 0$$

where, $g(t) = \dfrac{M}{1 - \delta\sigma(t, t + \Theta)}$,
and $M = \max(\alpha\gamma\delta\sigma(\eta, \Theta), \alpha\delta\sigma(\eta, \Theta), \delta\sigma(\eta, \Theta), \alpha\gamma\delta\sigma(\eta, (1+p)\eta))$, where $p = \alpha\gamma\delta\sigma(\eta, \Theta)$, has at least one positive root. The smallest positive root is denoted by R.

[C_7] $g(R) \in (0, 0.618034...)$.

[C_8] $\overline{\mathcal{B}}(x_0, R) \subseteq \mathrm{D}$.

Lemma 1 *For the improved Chebyshev-Secant-type method (ICSTM) proposed in (3) it is verified:*

(i) $\mathrm{F}(z_k) = \alpha([z_k, x_k; \mathrm{F}] - \mathrm{B}_k)(x_{k+1} - x_k) + (1 - \alpha)\mathrm{F}(x_k)$.

(ii) $\mathrm{F}(x_{k+1}) = ([x_{k+1}, x_k; \mathrm{F}] - [x_k, y_k; \mathrm{F}])\,(x_{k+1} - x_k)$.

Proof: The proof follows obviously by (3) and the application of the usual property of the divided difference operator, $[x, y, \mathrm{F}](x - y) = \mathrm{F}(x) - \mathrm{F}(y)$, hence omitted here. □

Lemma 2 *For method ICSTM proposed in (3) under conditions* $[\mathrm{C}_1]-[\mathrm{C}_8]$ *and for* $(\mathrm{F}, x_0, y_0) \in \mathrm{S}(\Theta, \delta, \eta, \sigma)$*, we obtain the following bounds:*

(i) *There exists* B_k^{-1} *satisfying* $\|\mathrm{B}_k^{-1}\| \leq \frac{\delta}{1-\delta\sigma(R,R+\Theta)}$,

(ii) $\|x_{k+1} - x_k\| \leq g(R)\|x_k - x_{k-1}\|$,

(iii) $\|y_{k+1} - x_k\| \leq (1 + g(R))\|x_{k+1} - x_k\|$,

(iv) $\|y_{k+1} - x_{k+1}\| \leq g(R)\|x_{k+1} - x_k\|$,

(v) $\|x_{k+1} - x_0\| \leq \sum_{j=0}^{k} g(R)^j \eta < R$,

(vi) $\|y_{k+1} - x_0\| \leq \sum_{j=0}^{k+1} g(R)^j \eta < R$,

(vii) $\|z_k - x_0\| \leq \sum_{j=0}^{k} g(R)^j \eta < R$.

Proof: The above inequalities can be proved by using mathematical induction. Using Lemma 1 and the definition of class $\mathrm{S}(\Theta, \delta, \eta, \sigma)$, we get $\|x_1 - x_0\| \leq \eta$, $\|z_0 - x_0\| \leq \eta$ and

$$\begin{aligned} \|y_1 - x_0\| &= \|x_1 - x_0 - \alpha\gamma \mathrm{B}_0^{-1}\sigma(\|z_0 - x_0\|, \|x_0 - y_0\|)(x_1 - x_0)\| \\ &\leq (1 + \alpha\gamma\delta\sigma(\eta, \Theta))\,\|x_1 - x_0\| < (1 + g(R))\eta < R. \end{aligned}$$

with $\|y_1 - x_1\| \leq \alpha\gamma\delta\sigma(\eta, \Theta)\|x_1 - x_0\| \leq g(R)\|x_1 - x_0\|$. Thus, lemma holds for $n = 0$. Suppose that it holds for some $n \leq k$. Now,

$$\|I - \mathrm{B}_0^{-1}\mathrm{B}_k\| \leq \delta\sigma(\|x_k - x_0\|, \|y_k - x_0\| + \|y_0 - x_0\|) \leq \delta\sigma(R, R + \Theta) < 1.$$

So using Banach's lemma on invertible operators [3], it is verified

$$\|\mathrm{B}_k^{-1}\| \leq \frac{\delta}{1 - \delta\sigma(R, R + \Theta)}.$$

Using Lemma 1 once more, we get

$$\begin{aligned} \|x_{k+1} - x_k\| &\leq \|\mathrm{B}_k^{-1}\|\|\mathrm{F}(x_k)\| \\ &\leq \frac{\delta\sigma(\|x_k - x_{k-1}\|, \|x_{k-1} - y_{k-1}\|)}{1 - \delta\sigma(R, R + \Theta)}\|x_k - x_{k-1}\| \\ &\leq g(R)\|x_k - x_{k-1}\|. \end{aligned}$$

Now,

$$\begin{aligned} \|y_{k+1} - x_k\| &\leq \|x_{k+1} - x_k - \alpha\gamma \mathrm{B}_k^{-1}\sigma(\|z_k - x_k\|, \|x_k - y_k\|)(x_{k+1} - x_k)\| \\ &\leq \left(1 + \frac{\alpha\gamma\delta\sigma(\|z_k - x_k\|, \|x_k - y_k\|)}{1 - \delta\sigma(R, R + \Theta)}\right)\|x_{k+1} - x_k\| \\ &\leq (1 + g(R))\|x_{k+1} - x_k\|. \end{aligned}$$

This gives

$$\|y_{k+1} - x_{k+1}\| \leq \|\alpha\gamma \mathrm{B}_k^{-1}\sigma(\|z_k - x_k\|, \|x_k - y_k\|)(x_{k+1} - x_k)\| \leq g(R)\|x_{k+1} - x_k\|.$$

Thus this proves (i)-(iv). (v), (vi) and (vii) can easily be obtained with the recursive use of (i)-(iv). Hence, this proves the lemma. □

Theorem 1 *Let $F : D \subseteq X \to Y$ be a continuous nonlinear operator, and consider the triplet $(\mathrm{F}, x_0, y_0) \in \mathrm{S}(\Theta, \delta, \eta, \sigma)$ defined in section 2, with $x_0, y_0 \in \mathrm{D}$ verifiying conditions $[\mathrm{C}_1] - [\mathrm{C}_8]$. Then, by taking x_0, y_0 as starting points, the sequences $\{x_k\}$, $\{y_k\}$ and $\{z_k\}$ generated by (3) are well defined and belong to $\mathcal{B}(x_0, R) \subseteq \mathrm{D}$. Also, the iterate x_k, y_k and z_k converge to $\rho^* \in \overline{\mathcal{B}}(x_0, R) \subseteq \mathrm{D}$, where ρ^* is the unique solution of (1) in $\overline{\mathcal{B}}(x_0, R) \cap \mathrm{D}$.*

Proof: Using Lemma 1 and Lemma 2, we see that the iterates x_k and y_k are well defined and belong to $\mathcal{B}(x_0, R)$. It is sufficient to show that $\{x_k\}$ is a Cauchy sequence. For fixed k and $m \geq 1$, we get

$$\begin{aligned}
\|x_{k+m} - x_k\| &\leq \|x_{k+m} - x_{k+m-1}\| + \ldots + \|x_{k+1} - x_k\| \\
&\leq \left(g(R)^{m-1} + g(R)^{m-2} + \ldots + g(R) + 1\right)\|x_{k+1} - x_k\| \\
&\leq \left(g(R)^{m-1} + g(R)^{m-2} + \ldots + g(R) + 1\right)\|x_{k+1} - x_k\| \\
&\leq \left(\frac{1 - g(R)^m}{1 - g(R)}\right) g(R)^k \|x_1 - x_0\|.
\end{aligned}$$

Therefore $x_k \to \rho^*$ as $k \to \infty$. Now, we show that ρ^* is a solution of (1). From Lemma 1, we get

$$\|\mathrm{F}(x_{k+1})\| \leq \|[x_{k+1}, x_k; \mathrm{F}] - [x_k, y_k; \mathrm{F}]\|\|x_{k+1} - x_k\| \to 0 \text{ as } k \to \infty.$$

From the continuity of F, it is assured that $\mathrm{F}(\rho^*) = 0$. To show the uniqueness of ρ^*, let $\widehat{\rho}$ be another solution of (1) in $\overline{\mathcal{B}}(x_0, R)$ such that $\mathrm{F}(\widehat{\rho}) = 0$. For $\mathrm{B}^* = [\rho^*, \widehat{\rho}; \mathrm{F}]$, we get

$$\|I - \mathrm{B}_0^{-1}\mathrm{B}^*\| \leq \delta\sigma(\|x_k - x_0\|, \|y_k - x_0\| + \|y_0 - x_0\|) \leq \delta\sigma(R, R + \Theta) < 1.$$

This shows that B^* is invertible and from the identity $[\rho^*, \widehat{\rho}; \mathrm{F}](\rho^* - \widehat{\rho}) = \mathrm{F}(\rho^*) - \mathrm{F}(\widehat{\rho})$, taking norms on both sides, we get $\rho^* = \widehat{\rho}$. This implies the uniqueness of ρ^*. □

Keywords: Nonlinear equations, Divided differences, Semilocal convergence.

References

[1] Ren, H. and Argyros, I., On the convergence of King-Werner-type methods of order $1+\sqrt{2}$ free of derivatives, *Applied Mathematics and Computation*, 256: 148–159, (2015).

[2] Kumar, A., Gupta, DK, Martínez, E. and Singh, S., Semilocal convergence of a Secant-type method under weak Lipschitz conditions in Banach spaces, *Journal of Computational and Applied Mathematics*, 330: 732–741, (2018).

[3] Rall, Louis B, *Computational solution of nonlinear operator equations*, Wiley New York (1969).

10

Mathematical modeling of Myocardial Infarction

B. Chen-Charpentier [♭1] and H. Kojouharov [♭]

(♭) Department of Mathematics,
University of Texas at Arlington.

1 Introduction

Coronary artery disease is the leading cause of death worldwide. According to the World Health Organization there are more than seven million deaths every year due to cardiac infarction. Myocardial Infarction (MI), commonly known as a heart attack, occurs when a lack of blood flow and oxygen to a region of the heart results in the death of cardiac muscle cells. The heart is a very complex organ composed of a wide variety of cells: Cardiomyocytes which account for 20-35% of cells, endothelial cells that make up the largest proportion (60%) of non-cardiomyocytes and are involved in angiogenesis, fibroblasts which are among the most represented cell populations of the heart and have a key homeostatic role in the synthesis of the cardiac extracellular matrix, pericytes are smooth muscle like cells and immune cells, including monocytes and neutrophils, that are rapidly recruited in large numbers to the heart following injury [3]. Even though there is a large variety of immune cells including neutrophils, monocytes, macrophages both pro-inflammatory M1 and anti-inflammatory M2, dendritic cells, lymphocytes and mast cells, the first three are the most common in MI [5]. These cells interact among themselves directly, for example neutrophils ingesting dead myocytes, and indirectly through molecular mediators such as cytokines, chemokines and growth factors. After MI, the complex myocardial healing process can be divided into four distinct phases:

1. Necrotic phase: acute cell death (immediately after MI).
2. Acute inflammatory phase: inflammatory response in order to absorb necrotic tissue (1-7 days).
3. Sub-acute granulation phase: formation of granulation tissue which consists of proliferating myofibroblasts, which help stabilize the heart by increasing blood vessels (1-3 weeks).
4. Chronic scar phase: fibroblasts formation and generation of the final collagen-rich scar tissue (1 month).

In this paper we are concerned with the first two phases. That is, with the death of cardiomyocytes, and on the effects of the immune system in cleaning the heart of death cells.

[1]e-mail: bmchen@uta.edu

We construct two mathematical models. The first one deals with the second phase: cardiomyocytes have died and the immune system is cleaning the death cells. In the second model we add the death process of the cardiomyocytes. Both models are based on ordinary differential equations.

2 Mathematical models

We assume the following hypotheses: We consider a small part of the heart tissue where the distribution of the involved cells is homogeneous. Communication between cells is done through a series of chemical factors and we assume that the amount of each factor is proportional to the number of cells that produce it. Following MI, neutrophils initiate pro-inflammatory response. They phagocytosize the dead cardiomyocytes. After the neutrophil infiltration, circulating monocytes migrate from blood to the site of the lesion, where they differentiate into macrophages of both types, gradually replacing neutrophils. M1 macrophages continue eliminating dead cardiomyocytes and also dead neutrophils. The rate of conversion of monocytes to M1 macrophages depends on the number of dead dells. Macrophage conversion from a pro-inflammatory (M1) to an anti-inflammatory (M2) phenotype is a central process and several molecular and cellular processes are involved. Macrophage phagocytosis is one of the mechanisms leading to M1/M2 conversion. M2 macrophages inhibit M1 macrophages and their pro-inflammatory response. After all the dead cells have been removed, remaining neutrophils and M1 macrophages leave the system [1, 2].

We first model the second phase in the MI process. The mathematical model consists of six ordinary nonlinear differential equations in the unknowns: dead cardiomyocytes, C_d, neutrophils, N, death neutrophils, N_d, monocytes, M, macrophages M1, M_1, and macrophages M2, M_2. The dead cardiomyocytes attract neutrophis and monocytes both of which eliminate the dead cardiomyocytes. Neutrophils ingest dead cardiomyocytes and die. Dead neutrophils are either phagocytosized by M1 macrophages or are flushed out of the system. Monocytes are recruited by dead cardiomyocytes and neutrophils and transform into macrophages M1 and M2 and are also eliminated from the system. Macrophages M1 and M2 can switch phenotype and are also eliminated from the system. The presence of macrophages M2 inhibits the changing of monocytes into macrophages M1. The evolution in time of the six populations is given by Eq. (1).The model involves sixteen parameters. All parameters are assumed to be positive.

$$\frac{dC_d}{dt} = -d_4 M_1 C_d - d_5 N C_d \tag{1a}$$

$$\frac{dN}{dt} = k_0 C_d - d_0 N \tag{1b}$$

$$\frac{dN_d}{dt} = d_0 N - d_9 M_1 N_d - d_{10} N_d \tag{1c}$$

$$\frac{dM}{dt} = k_4 N + k_5 C_d - \frac{k_1}{M_2 + k_6} C_d M - \frac{k_2}{M_2 + k_6} N_d M - d_8 M \tag{1d}$$

$$\frac{dM_1}{dt} = \frac{k_1}{M_2 + k_6} C_d M + \frac{k_2}{M_2 + k_6} N_d M - d_6 C_d M_1 - d_7 N_d M_1 - d_1 M_1 \tag{1e}$$

$$\frac{dM_2}{dt} = d_6 C_d M_1 + d_7 N_d M_1 - d_2 M_2. \tag{1f}$$

We show that the model 1 has only one biologically realistic fixed point: $C_d = 0$, $N = 0$, $N_d = 0$, $M = 0$, $M_1 = 0$ and $M_2 = 0$. For non-negative initial conditions system 1 has non-negative solutions; the solutions of system are bounded above; and that the steady solution is stable. The model involves sixteen parameters. Measurements of the parameters are few and they have large variations. The decay rates are the best known.

The second model simulates phases 1 and 2. That is, we consider the cardiac infarction since the flow of blood and oxygen is interrupted. Cardiomyocytes start dying and stop dying when the flow are restored. The model has one extra population C, healthy cardiomyocytes. The only modification to Model 1 is to add and equation for heathy cardiomyocytes and change the equation dead cardiomyocytes:

$$\frac{dC}{dt} = -d_{11}C\text{Heaviside}(\tau - t)$$
$$\frac{dC_d}{dt} = d_{11}C\text{Heaviside}(\tau - t) - d_4 M_1 C_d - d_5 N C_d$$

Here Heaviside(.) is the Heaviside function and τ is the time in days until oxygen flow is restored. For $T \leq \tau$ there is no steady solution since C is decaying. For $t > \tau$, C is constant and the other populations are zero as shown above.

3 Sensitivity Analysis

Mathematical models depend on a number of parameters. Determining parameter values is a difficult issue since there are large variations and uncertainties associated with heir measurement, or since they have to be estimated indirectly. Therefore it is important to determine how sensitive the model output is to variations in the parameter values. A local sensitivity analysis determines the model sensitivity to parameter variations over a localised region around a given set of parameter values. A global sensitivity analysis (GSA) investigates the sensitivity over the entire parameter space. Since there is sensitivity index for each type of cell with respect to each parameter, and they are functions of time, there is a large amount of information. We investigate ways of reducing it by looking only at the steady solutions, and by just considering them at a few selected times, and by taking time averages over the simulation time. We also investigate the effects on only considering the indices for one relevant variable such as the number of death cardiomyocytes, and by looking at the average, maximum and minimum values of the indices.

Local sensitivity analysis is fast to solve, but does not take into account the range of variability of the parameters. And furthermore, it changes the values of the parameters one at a time. Our results suggest that a good strategy for time-dependent sensitivity indices is to use the average over the time period of interest.

In global sensitivity analysis methods values of the parameters are varied over their ranges of interest. All parameters are varied simultaneously so the methods are expensive. There is the question of how to choose these combinations of values. A common alternative is to use the latin hypercube sampling method to generate random points.

There are two main classes of global sensitivity methods: Regression-based methods with rank transformation such as the partial rank correlation coefficient (PRCC), and variance-based methods such as the Sobol's method and Fourier amplitude sensitivity test (FAST). Some general references are [4, 6]. PRCC is faster to calculate but it can be inaccurate. The larger local sensitivity indices are d_{10}, k_0, k_6. For Sobol's method and for the FAST method, the largest global sensitivity indices are d_0, d_2, d_5, k_0.

4 Numerical simulations

We assume that initially there is a given number of dead cardiomyocytes. The initial conditions used for the simulations are $C_d = 2000, N = N_d = M = M_1 = M_2 = 0$. The units are cells/mm^3. The value of C_d corresponds to assuming that 25% of the cardiocytes died from lack or oxygen. The simulations show that changing some parameters by a factor of two changes significantly the number of cells and the time it takes the immune system to get rid of the dead cells.

For simulations using Model 2, the same parameter values for Model 1 were used adding the dead rate of cardiomyocytes due to lack of oxygen $d_{11} = .35$. The initial conditions are $C = 8000, C_d = N = N_d = M = M_1 = M_1 = 0$ since we are starting with a healthy heart and cutting the flow of oxygen at time $t = 0$. We varied the length of time τ before the flow of oxygen was restored, from .1 day to 1 day. As expected, the longer it takes to restore the flow of oxygen, the larger the number of dead cardiocytes and the longer it takes the immune system to clean the dead cells.

5 Conclusions

The models give the quantitative behavior of the immune system after MI. Model 1 shows that the only steady solution is the zero solution and it is stable. That is, all the dead cardiocytes are cleaned. For $t > \tau$, Model 2 has the same zero solution with the exception that the number of live cardiomyocytes is a constant. Sensitivity analysis shows that parameter values can produce significant changes, specifically that the values of the phagocytosis rates are very important. Sensitivity analysis helps determine the most important parameters but needs caution. Local sensitivity analysis, even though simple and fast doesn't always give the parameters that produce the largest changes. In our case, the sensitivity analysis results when the solutions are averaged over time are very similar as those at some fixed time values. Sobol's method and FAST give similar results that validate both sets of results, but PRCC doesn't. The simulation results show that changing some parameters by a factor of two significantly changes the cell populations and the time it takes the immune system to eliminate the dead cells. So it is very important to measure those parameters accurately. Finally, Restoring the oxygen flow as fast as possible is critical.

References

[1] Bonvini, R.F., Hendiri, T. and Camenzind, E., Inflammatory response post-myocardial infarction and reperfusion: a new therapeutic target? *European Heart Journal Supplements*,

7(suppl_I):I27–I36, Oct. 2005.

[2] Dunster, J.L., The macrophage and its role in inflammation and tissue repair: mathematical and systems biology approaches: Macrophage and its role in inflammation and tissue repair. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(1):87–99, Jan. 2016.

[3] Gray, G., Toor, I., Castellan, R., Crisan, M., and Meloni, M., Resident cells of the myocardium: more than spectators in cardiac injury, repair and regeneration. *Current Opinion in Physiology*, 1:46–51, Feb. 2018.

[4] Marino, S., Hogue, I.B., Ray, C.J., and Kirschner, D.E., A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of theoretical biology*, 254(1):178–196, 2008.

[5] Nahrendorf, M. and Swirski, F.K., Monocyte and Macrophage Heterogeneity in the Heart. *Circulation Research*, 112(12):1624–1633, June 2013.

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

11

Symmetry relations between dynamical planes

Francisco I. Chicharro[♭1], Alicia Cordero[♮], Neus Garrido[♮] and Juan R. Torregrosa[♮]

(♭) Escuela Superior de Ingeniería y Tecnología,
Universidad Internacional de La Rioja,
(♮) Instituto de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

There is a large number of problems in Science and Engineering that can not be solved analytically. One example is the solution of the nonlinear equation $f(x) = 0$. A way to proceed with this kind of problems is the application of iterative methods. Many techniques regarding this topic can be found in the literature [1], and so many classifications based on their own features. Some criteria to sort the iterative methods are based on its order of convergence p, the number of functional evaluations per step d or the optimality of the method [2] when $p = 2^{d-1}$, amongst others. However, a key point of an iterative scheme is the stability of the method for every initial guess. This study can be performed with a dynamical analysis.

During the last years, the analysis of the stability by means of complex dynamics has been widely extended between the authors of this topic. In order to understand where and why the iterative method fails or succeeds, a deep study on complex dynamics needs to be performed [3, 4]. However, a smoother analysis can be performed if the objective is only the knowledge of where the iterative scheme fails or succeeds. In both cases, the complex dynamics analysis is assisted by graphical tools. The main representation is the dynamical plane [5, 6]. The dynamical plane represents, for a set of initial guesses in the complex plane, the final state of the orbit of each initial guess.

Let $R : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ be a rational function. Moreover, this rational function represents the fixed point operator of an iterative method when it is applied on a polynomial $p(z)$. In [7] it was stated that the dynamical planes of the introduced iterative methods satisfied the property

$$R(\bar{z}) = \overline{R(z)}. \tag{1}$$

One consequence can be found by means of symmetry . If $R(z)$ satisfies (1), then the dynamical planes are symmetric about polar axis. Regarding the computational cost side, a symmetry directly involves the reduction of the number of operations that need to be performed to obtain the dynamical planes.

[1]e-mail: francisco.chicharro@unir.net

2 Symmetries in one-point iterative methods of order two

Any iterative method of one point of order two can be expressed as [8]

$$z_{k+1} = z_k - H(t(z_k))\frac{f(z_k)}{f'(z_k)}, \tag{2}$$

where $H(t)$ is a function of variable $t = f(z)/f'(z)$ that satisfies $H(0) = 1$.

Proposition 1 *Let $f(z) : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ and $H(t) = \hat{\mathbb{C}} \to \hat{\mathbb{C}}$. If $f(z)$ and $H(t)$ satisfy*

1. *$f(\bar{z}) = \overline{f(z)}$ and $f'(\bar{z}) = \overline{f'(z)}$,*
2. *$H(\bar{t}) = \overline{H(t)}$,*

then $R(\bar{z}) = \overline{R(z)}$.

On the one hand, some functions $f(z)$ that satisfy the previous property are the polynomials of real coefficients $p(z)$

$$p(z) = \sum_{i=0}^{n} p_i z^i, \quad p_i \in \mathbb{R}.$$

In addition, some functions $H(t)$ that satisfy both $H(\bar{t}) = \overline{H(t)}$ and $H(0) = 1$ are the rational functions of real coefficients, such that

$$H(t) = \frac{a_0 + \sum_{i=1}^{n} a_i t^i}{a_0 + \sum_{j=1}^{m} b_j t^j}, a_i, b_j \in \mathbb{R}.$$

Some examples can be found in the literature, such that the methods of

- Newton, $H(t) = 1$.
- Kanwar-Tomar [9], $H(t) = \frac{1}{1+\beta t}, \beta \in \mathbb{R}$.
- Kou-Li [10], $H(t) = 1 + \frac{\lambda t}{(1+\beta t)(1+2\beta t)}, \lambda, \beta \in \mathbb{R}$.

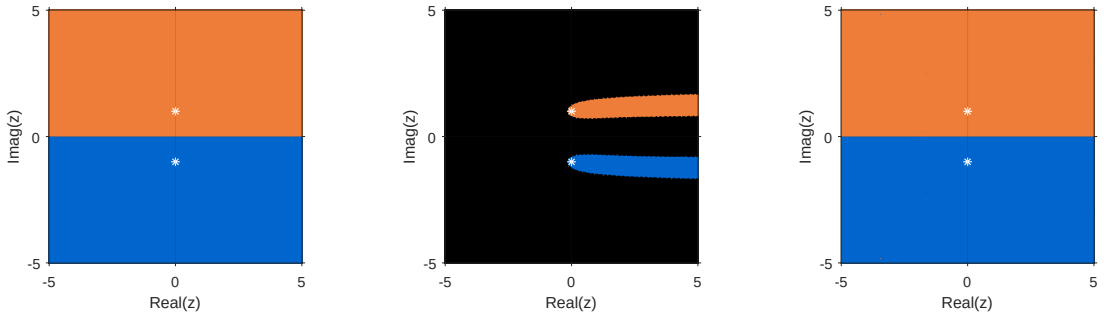
2.1 Application of $p(z) = z^2 + 1$

The polynomial $p(z) = z^2+1$ has two complex roots. The application of the iterative methods of Newton, Kanwar-Tomar and Kou-Li on $p(z)$ satisfy the symmetry property, since $p(\bar{z}) = \overline{p(z)}$. The corresponding fixed point operators of Newton's, Kanwar-Tomar's and Kou-Li's methods are

$$R(z) = \frac{z^2 - 1}{2z}, R(z) = \frac{\beta\left(z^3 + z\right) + z^2 - 1}{\beta z^2 + \beta + 2z}, R(z) = z - \frac{\left(z^2 + 1\right)\left(\frac{\lambda\left(z^3+z\right)}{(\beta z^2+\beta+z)(\beta z^2+\beta+2z)} + 1\right)}{2z},$$

respectively.

Figure 1 represents the dynamical planes of these methods when $p(z) = z^2 + 1$, for $\beta = 5, \lambda = -1/2$. The initial guesses in blue converge to the root $z_1 = -i$, while the orange corresponding ones converge to the root $z_2 = i$. In the three dynamical planes, the symmetry property can be observed. The knowledge of the semiplane $Im(z) > 0$ involves directly the knowledge of the behavior of the semiplane $Im(z) < 0$.



(a) Newton (b) Kanwar-Tomar (c) Kou-Li

Figure 1: Dynamical planes of methods of order two applied on $p(z) = z^2 + 1$.

2.2 Application of $p(z) = z^2 - 1$

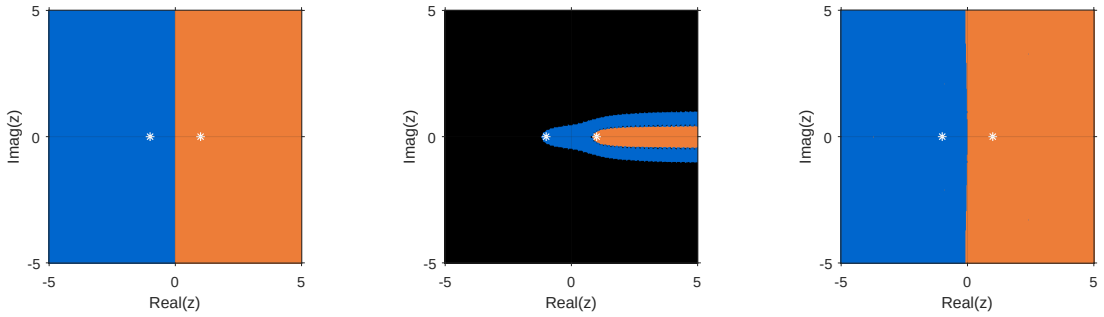
The polynomial $p(z) = z^2 - 1$ has two real roots, and satisfies the condition $p(\bar{z}) = \overline{p(z)}$, since $p(z)$ is a polynomial of real coefficients. In this particular case, the fixed point operators of Newton', Kanwar-Tomar' and Kou-Li's methods on $p(z)$ are

$$R(z) = \frac{z^2+1}{2z}, R(z) = \frac{\beta\left(z^2-1\right)z + z^2 + 1}{\beta\left(z^2-1\right) + 2z}, R(z) = z - \frac{\left(z^2-1\right)\left(\frac{\lambda\beta z\left(z^2-1\right)}{(\beta(z^2-1)+z)(\beta(z^2-1)+2z)} + 1\right)}{2z},$$

respectively.

Figure 2 represents the dynamical planes of these methods when $p(z) = z^2 - 1$, for $\beta = 5, \lambda = -1/2$. The initial guesses in blue converge to the root $z_1 = -1$, while the orange corresponding ones converge to the root $z_2 = 1$.

As in the previous case, there exists a conjugated symmetry in the dynamical planes. Obtaining the dynamical plane of the semiplane $Im(z) > 0$ is enough to know the aspect of the semiplane $Im(z) < 0$.



(a) Newton (b) Kanwar-Tomar (c) Kou-Li

Figure 2: Dynamical planes of methods of order two applied on $p(z) = z^2 - 1$.

3 Conclusions

The representation of the dynamical plane associated to an iterative method when a nonlinear function is applied gives information about the initial guesses where the method fails or succeeds. Its obtention requires a computational cost that depends on the number of initial guesses, amongst other parameters. If the nonlinear function $f(z)$ and the weight function $H(t)$ that implements the iterative method satisfy the described properties, the computational cost of representing the dynamical plane can be reduced.

Acknowledgement

This research was partially supported by MCIU/AEI/FEDER, UE PGC-2018-095896-B-C22 and Generalitat Valenciana PROMETEO/2016/089.

References

[1] Amat, S., and Busquier, S., *Advances in Iterative Methods for Nonlinear Equations*, Springer, 2016.

[2] Kung, H. T. and Traub, J. F., Optimal order of one-point and multipoint iteration, *J. Assoc. Comput. Math.*, Volume(21), 643–651, 1974.

[3] Fagella, N., Invariants en dinàmica complexa, *Butlletí Societat Catalana Matemàtiques*, Volume(23), 29–51, 2008.

[4] Cordero, A., Gutiérrez, J. M., Magreñán, Á. A. and Torregrosa, J. R., Stability analysis of a parametric family of iterative methods for solving nonlinear models, *Appl. Math. Comput.*, Volume(285), 26–40, 2016.

[5] Varona, J. L., Graphic and numerical comparison between iterative methods, *Mathematical Intelligencer*, Volume(24), 37–46, 2002.

[6] Chicharro, F. I., Cordero, A. and Torregrosa, J. R., Drawing Dynamical and Parameters Planes of Iterative Families and Methods, *The Scientific World Journal*, Volume(2013), ID 780153, 1–11, 2013.

[7] Chicharro, F. I., Cordero, A. and Torregrosa, J. R., Dynamics and fractal dimension of Steffensen-type methods, *Algorithms*, Volume(8), 271–279, 2015.

[8] Cordero, A., Jordán, C. and Torregrosa, J. R., One-point Newton-type iterative methods: a unified point of view, *J. Comput. Appl. Math.*, Volume(275), 366–374, 2015.

[9] Kanwar, V. and Tomar, S. K., Modified families of Newton, Halley and Chebyshev methods, *Appl. Math. Comput.*, Volume(192), 20–26, 2007.

[10] Kou, J. and Li, Y., A family of new Newton-like methods, *Appl. Math. Comput.*, Volume(192), 162–167, 2007.

12

Econometric methodology applied to financial systems

S. Climent-Serrano[♭1],

(♭) Economía Financiera y Actuarial,
Universitat de València.

In this paper we review the main characteristics of econometric models, focusing on their application in financial systems. The objective is to give an overview of the econometric technique and its practical application. The methodology used is a description of the bases of econometrics and the properties it assumes. The different types of adjustments are developed so that they can be made (logarithmic, linear, etc.), and the goodness-of-fit measurements. Finally, they are analysed by the practical application of the different types of variables and their real application to regression models on Spanish credit institutions.

1 Introduction: Economic and econometric models

In a market economy, we can assume that the higher the GDP, the greater will be the deposits in credit institutions. At first, we can also assume that this relationship is linear. In this case, an example of the relationship between two variables can be seen in Fig. 1. In this case we are dealing with an economic model.
The representation will be: $D = \beta_1 + \beta_2 GDP$, $0 < \beta_2 < 1$.
In the representation, D are the bank deposits, GDP the gross domestic product, and β_1, β_2 the parameters of the model.
However, when we look at the real economy, the relationship between GDP and deposits is not exact and may be more like the representation shown in Fig. 2.

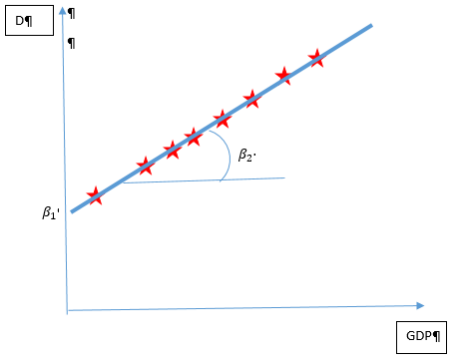


Figure 1: Linear adjustment.

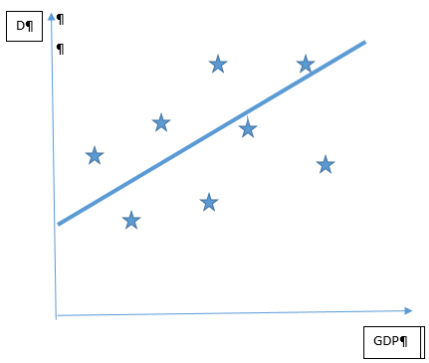


Figure 2: Real situation.

[1]e-mail: Salvador.climent@uv.es

The relationships between economic variables are usually not exact, since they depend on multiple factors, such as the propensity to save or to spend, age, ease of access to bank branches, even random factors. To take all these factors into consideration, a random term "u" is included in the model, which we will call error or resid, which is not observable and which represents the factors that are not in the deterministic part of the model. In this case it will be an econometric model. So the model will be:

$$D = \beta_1 + \beta_2 GDP + u, \quad 0 < \beta_2 < 1.$$

The objective of the regression is to obtain the parameters β_1 and β_2 from an available sample. Estimates should be sought for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ (as the results obtained are estimates, a circumflex accent is introduced above the betas, or of what we estimate).
If we look at Fig. 2, the distance from each point to the line will be $D_i - \widehat{D}_i = \widehat{u}_i$. This error or resid is the difference between the observed value of the endogenous variable (dependent variable) and the adjusted value.

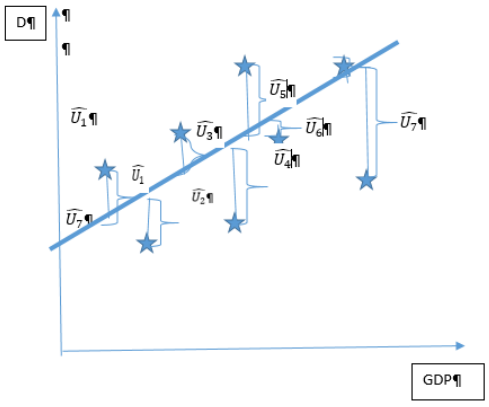


Figure 3: Adjustments

The objective of the regression is to adjust the line to the points so that all the residuals or errors, which are shown in Fig. 3, are as small as possible.
Mathematically as in the $\sum_{i=1}^{N} \widehat{u}_i$, we have positive and negative errors; so that they are not compensated, we modify the equation to: $\sum_{i=1}^{N} \widehat{u}_i^2$.
This adjustment is called ordinary least squares and the intention is to minimize the expression:

$$\min \sum_{i=1}^{N} \hat{u}_i^2 = \min \sum_{i=1}^{N} \left(GDP_i - \widehat{GDP}_i \right)^2.$$

The goal is to find $\widehat{\beta_1}$ and $\widehat{\beta_2}$ that minimize the expression:

$$\min \sum_{i=1}^{N} \hat{u}_i^2 = \min \sum_{i=1}^{N} \left(GDP_i - \widehat{\beta_1} - \widehat{\beta_2 GDP}_i \right)^2.$$

To minimize, firstly it is derived partially with respect to $\widehat{\beta_1}$ and $\widehat{\beta_2}$. The first order condition is obtained by equating the expressions obtained to zero. Solving the equations, we obtain the

normal equations and solving these we arrive at the following expressions:

$$\widehat{\beta_1} = \bar{D} - \widehat{\beta_2}\overline{GDP} \quad \text{and} \quad \widehat{\beta_2} = \frac{Cov(GDP, D)}{Var(GDP)}.$$

The estimated line will be $D_i = \beta_1 + \beta_2 GDP_i U_i$. This means that $\widehat{D_i} = \beta_1 + \beta_2 GDP_i$, for example $\widehat{D_i} = 5 + 0.30GDP$.

Properties:

1. The sum of the residuals is zero: $\sum_{i=1}^{N} \widehat{U} = 0$.
2. The adjusted line passes through the point of means: $G\bar{D}P, \bar{D}$.
3. The sum of the cross products of the residuals with the explanatory variable is zero: $\sum_{i=1}^{N} \widehat{U}GDP = 0$.
4. The sum of the cross products of the resid with $\widehat{D_t}$ is zero: $\sum_{i=1}^{N} \widehat{U}\widehat{D} = 0$.

2 Non-linear adjustments

In the event that the variables are not linear, non-linear estimates can be made with alternative functional forms, for example: Logarithmic adjustment: $\lambda^n \widehat{D_i} = \beta_1 + \beta_2 \lambda^n GDP_i + U_i$. In this case, the interpretation of the coefficient is related to the elasticity. In other words, if GDP changes 1% (Δ GDP / GDP) = 1%, then D will change $(\Delta D/D) = \beta_2\%$.
As well as this functional form, others can be estimated. The summary of the alternative functional forms, and the interpretation of each of the parameters, is shown in Table 1.

	MODEL	**MARGINAL PROPENSION**	**ELASTICITY**
LINEAR	$D_I = \beta_1 + \beta_2 GDP_i + U_i$	$\widehat{\beta_2}$	$\frac{\widehat{\beta_2} \cdot \overline{GDP}}{\bar{D}}$
INVERSE	$D_I = \beta_1 + \frac{\beta_2 \cdot 1}{GDP_I} + U_i$	$\frac{\widehat{-\beta_2} \cdot 1}{\overline{GDP}^2}$	$\frac{\widehat{-\beta_2} \cdot 1}{\overline{GDP} \cdot \bar{D}}$
LINEAR-LOG	$D_I = \beta_1 + \beta_2 lnGDP_i + U_i$	$\frac{\widehat{\beta_2} \cdot \bar{D}}{\overline{GDP}}$	$\frac{\widehat{\beta_2} \cdot 1}{\bar{D}}$
LOG - LOG	$lnD_I = \beta_1 + \beta_2 lnGDP_i + U_i$	$\frac{\widehat{\beta_2} \cdot 1}{\overline{GDP}}$	$\widehat{\beta_2}$
LOG - LINEAR	$lnD_I = \beta_1 + \beta_2 GDP_i + U_i$	$\widehat{\beta_2} \cdot \bar{D}$	$\widehat{\beta_2} \cdot \overline{GDP}$
LOG - INVERSE	$lnD_I = \beta_1 + \beta_2 \frac{1}{\overline{GDP_{I_i}}} + U_i$	$\frac{\widehat{-\beta_2} \cdot \bar{D}}{\overline{GDP}^2}$	$\frac{\widehat{-\beta_2} \cdot 1}{\overline{GDP}}$

Table 1: Functional forms of the econometric regressions.

Coefficient of determination or R^2 being $R^2 = \frac{\sum_{i=1}^{N}(\widehat{D_i} - \widehat{\bar{D}})^2}{\sum_{i=1}^{N}(D_i - \bar{D})^2}$; therefore, R^2 (R squared) is the proportion of the sample variance of D explained by the regression; the greater R squared is, the better the adjustment.

Dependent Variable: _2_DEPOSITS
Method: Panel Least Squares
Date: 12/12/18 Time: 00:22
Sample: 2004 2017
Periods included: 14
Cross-sections included: 77
Total panel (unbalanced) observations: 569
White diagonal standard errors & covariance (no d.f. correction)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.720250	0.002407	299.2344	0.0000
_6_REAL_GDP	0.296165	0.083927	3.528834	0.0005

Effects Specification

Cross-section fixed (dummy variables)

R-squared	0.798783	Mean dependent var	0.724545
Adjusted R-squared	0.767228	S.D. dependent var	0.113928
S.E. of regression	0.054966	Akaike info criterion	-2.837470
Sum squared resid	1.483447	Schwarz criterion	-2.241999
Log likelihood	885.2602	Hannan-Quinn criter.	-2.605117
F-statistic	25.31367	Durbin-Watson stat	0.981463
Prob(F-statistic)	0.000000		

Figure 4: Measures of goodness of the estimation by econometric regression.

3 Multiple linear regression

A usual extension of the simple linear regression model is to include more than one explanatory variable in the model. For example, that bank deposits depend on the GDP and the interest rate (I). In this case, the econometric model would be the following:

$$D_I = \beta_1 + \beta_2 GDP_i + \beta_3 I_i + U_i.$$

β_2 will be expected to be positive, with more GDP plus bank deposits. β_3 is expected to be negative; the higher the interest rate, the lower the incentive to have the money in bank deposits and the higher in other financial assets.

For the estimation of the coefficients, β, with a simple linear adjustment, with two variables it has been possible to intuit on the two-dimensional Cartesian axis. In the same way as with two variables, it can also be done with three variables, so one can intuit on a three-dimensional axis. With n variables you cannot intuit visually, since an n dimensional axis will be used, but the process is the same. In this case, the procedure for estimating betas is the same as that developed with two variables, but with matrix notation.

4 Basic assumptions that must comply with residuals or errors, u_i

- $E(u_i) = 0$ residuals are balanced between positive and negative.
- **Homoscedasticity:** $Var(u_i) = E(u_i^2) = \sigma^2$. The variance is constant. This is verified by the White test; if homoscedasticity persists, the Generalized Moment Method (MCM) procedure can be used since this procedure allows the presence of the homoscedasticity to estimate the coefficients of the model when using panel data [1].

- **No autocorrelation:** $cov(u_i, u_j) = E(u_i, u_j) = 0$. This is checked by the Durbin Watson test.

- **The explanatory variables are not related to the residual:** $cov(GDP_i, u_i) = 0$.

- **No multicollinearity:** The independent variables are linearly independent of each other. The vector of the residuals "u" has a **normal distribution**. This is checked by the Jarque-Bera Test. The model is well specified, in the form and in the variables.

To validate the model, the level of significance of each of the betas or parameters of the model is used. The level of significance or P-value is used, normally using levels of significance of 1%, 5% or 10%. With these levels of significance the variable is considered valid and for the whole model, the F test is used.

In this case, as shown in Fig. 5, the real GDP and the interest rate are significant variables, with a level of significance lower than 1% (confidence level above 99%), while the IPAP is not significant, since its p-Value is 0.3268.

Dependent Variable: _2_DEPOSITS
Method: Panel Least Squares
Date: 12/16/18 Time: 00:13
Sample: 2004 2017
Periods included: 14
Cross-sections included: 77
Total panel (unbalanced) observations: 569
White diagonal standard errors & covariance (no d.f. correction)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.741834	0.005231	141.8253	0.0000
_6_REAL_GDP	0.426065	0.086496	4.925809	0.0000
_6_EURIBOR	-0.703335	0.253538	-2.774075	0.0057
_6_IAPC	-0.290112	0.295553	-0.981591	0.3268

Effects Specification

Cross-section fixed (dummy variables)

R-squared	0.808275	Mean dependent var	0.724545
Adjusted R-squared	0.777301	S.D. dependent var	0.113928
S.E. of regression	0.053764	Akaike info criterion	-2.878762
Sum squared resid	1.413468	Schwarz criterion	-2.268023
Log likelihood	899.0077	Hannan-Quinn criter.	-2.640452
F-statistic	26.09531	Durbin-Watson stat	1.012126
Prob(F-statistic)	0.000000		

Figure 5: Regression model using the Eviews application.

5 Type or classes of variables

The variables that can be used in the regression models can be:

- **Metric variables:** for example: GDP, bank deposits, interest margin, capital, profits, etc.

- **Variables with ordinal scale:** for example: small, medium, large, systemic.

- **Dummy variables:** they are artificially constructed variables that collect this type of qualitative information, for example: bank, savings bank. Normally one is coded with zero (0) and the alternative with one (1). The interpretation is the difference between one of the variables and the alternative.

Fig. 6 shows the change that would occur if the dummy variable corresponding to a beta were positive. The economic interpretation to be given would be, as an example, if the bank is assigned the 1 and the savings banks the 0, that with the same GDP, the bank gets more deposits than the savings banks. It can also be used for more than two categories, for example to differentiate sectors or autonomous communities etc.

Fig. 7 shows an example of a model with metric variables and a dummy variable. On the left, the results are shown as they appear in the Eviews application. In this case, the model studies the losses by default of the Spanish banks with a sample that goes from 2004 to 2015.

The explanatory variables are all significant. This is verified with the p-value (Prob.), which is obtained from the t-statistic.

The different types of variables are:

1. Variable with positive sign. BDE, the money that the banks have requested from the ECB. In this case, the results indicate that the greater the need for central bank funds, the greater the losses due to late payments.

2. Variable with negative sign. The relationship of equity assets; in this case the results indicate that the credit entities that have more equity funds have fewer losses due to delinquency.

3. Control variable. To give robustness to the model, control variables are introduced, which are variables whose sign is predicted in advance. In this case they are the adjudications that the credit entities have. Here it is obvious that the result has to be positive and with a significant value. In this case we see that it is positive and with a coefficient of 0.16, much higher than the rest of the coefficients.

4. Dummy variable. Dummy_12. This variable represents the change in regulations that occurred in 2012 with respect to the allocation of provisions for impairments by what are colloquially known as "Guindos" decrees. The results indicate that the impairment losses in 2012 were 2.67% of the total assets as a consequence of this change. The way to include it in the model is to give the value zero to all the years of the sample, except for the year 2012 which is assigned the value one.

5. Variables with delay. In some cases, the dependent variable may have inertia, that is, it depends on the same variable from the previous year. This case occurs in delinquency, where the results of the variable DET_CREDITOS (-1) indicate that the losses due to delinquency depend 16% on the delinquency of the credit institution of the previous year. The way to include this variable is by introducing the dependent variable as explicative, with a delay of one year; if it is significant, this indicates that there is inertia. When entering the variable with a delay, one year of the sample is lost.

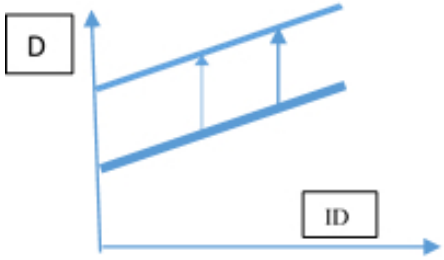


Figure 6: Linear adjustment.

Equation: Z_DET_BALANCE Workfile: PANEL 2015_TRAB::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: _0_DET_CREDITOS
Method: Panel EGLS (Cross-section weights)
Date: 12/17/18 Time: 20:31
Sample (adjusted): 2005 2015
Periods included: 11
Cross-sections included: 74
Total panel (unbalanced) observations: 455
Linear estimation after one-step weighting matrix
White diagonal standard errors & covariance (no d.f. correction)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.011423	0.001487	7.680269	0.0000
_1_BDE	0.041622	0.007776	5.352466	0.0000
_1_CRED_DEP	-0.002492	0.000705	-3.537298	0.0004
_1_FP_ACTIVO	-0.075622	0.013459	-5.618596	0.0000
_1_ADJUDICACIONES	0.167371	0.025980	6.442228	0.0000
_3_DUMMY_12	0.026733	0.002909	9.189745	0.0000
_0_DET_CREDITOS(...	0.163841	0.035502	4.614991	0.0000

Weighted Statistics

R-squared	0.653781	Mean dependent var	0.021166
Adjusted R-squared	0.649144	S.D. dependent var	0.024472
S.E. of regression	0.012634	Sum squared resid	0.071510
F-statistic	140.9966	Durbin-Watson stat	1.721722
Prob(F-statistic)	0.000000		

Figure 7: Real situation.

6 Type or classes of data

- **Time series data:** data of the same variable at different moments of time; for example: the profits of Banco Santander in 2010, 2011, 2012 etc.

- **Cross-section data:** data of the same variable in the same period of time; for example: the profits in 2015 of Banco Santander, BBVA, Bankia etc.

- **Panel data:** Combination of the two previous ones, which is usually used in the studies of the financial entities when the data is available; for example: the profits of Banco Santander, BBVA, Bankia, etc. in 2010, 2011, 2012 etc.

In addition, the form of the series can be modified; for example, if it is used to see how interest rates affect non-performing loans and loan defaults, the relationship may not be linear but quadratic, so the variable of interest rates is introduced to the table, etc.

It can also be introduced in a multiplicative way with fictitious variables, for example, interest rates multiplied by bank (1) and savings bank (0). In this case, the information it provides is how a variation in interest rates affects banks and savings banks in a different way.

7 Other checks

The Chow test [2] is used to test that there is no structural break in the model, that is, that the linearity does not break. But it can also be used to the contrary, to show that when a certain effect occurs, the variables do not behave in the same way. For example, if you study how GDP affects delinquency, the effect may not be the same during stages of growth as in recession and it can be contrasted that in year X (moment of change) there is a structural change, so that the coefficients of the parameters during growth stages will be different to the same parameter, but during recession stages.

Acknowledgement

The authors wishes to thank the support of the Chair of International Finance-Banco Santander.

References

[1] Arellano, M. and Bond, S., Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, 58(2): 277-297, 1991.

[2] Chow, G.C., Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28(3): 591-605, 1960.

[3] Granger, C. W. J., Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 37(3): 424-438, 1969

[4] Uriel Jiménez, E. and Gea Rosat, I. *Econometría aplicada*. Madrid : AC, 1967.

13

New matrix series expansions for the matrix cosine approximation

Emilio Defez [♭1], Javier Ibáñez[♮], José M. Alonso[♮], Jesús Peinado[♯] and Pedro Alonso-Jordá[*]

(♭) Institut Universitari de Matemàtica Multidisciplinar,
Universitat Politècnica de València,
(♮) Instituto de Instrumentación para Imagen Molecular,
Universitat Politècnica de València,
(♯) Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València,
(∗) Grupo Interdisciplinar de Computación y Comunicaciones,
Universitat Politècnica de València.

1 Introduction and notation

The computation of matrix trigonometric functions has received remarkable attention in the last decades due to its usefulness in the solution of systems of second order linear differential equations. Recently, several state-of-the-art algorithms have been provided for computing these matrix functions, see [1–4], in particular for the matrix cosine function.

Among the proposed methods for the approximate computation of the matrix cosine, two fundamental ones stand out: those based on rational approximations [1, 5–7], and those related to polynomial approximations, using either Taylor series developments [8,9] or serial developments of Hermite matrix polynomials [10]. In general, polynomial approximations showed to be more efficient than the rational algorithms in tests because they are more accurate despite a slightly higher cost.

Bernoulli polynomials and Bernoulli numbers have been extensively used in several areas of mathematics (an excelent survey about Bernoulli polynomials and its applicacions can be found in [11]).

In this paper, we will present a new series development of the matrix cosine in terms of the Bernoulli matrix polynomials. We are going to verify that its use allows obtaining a new and competitive method for the approximation of the matrix cosine.

The organization of the paper is as follows: In Section 2, we will obtain two serial developments of the matrix cosine in terms of the Bernoulli matrix polynomials. In Section 3, we will present

[1]e-mail: edefez@imm.upv.es

the different numerical tests performed. Conclusions are given in Section 4.

Throughout this paper, we denote by $\mathbb{C}^{r\times r}$ the set of all the complex square matrices of size r. Besides, we denote I as the identity matrix in $\mathbb{C}^{r\times r}$. A polynomial of degree m is given by an expression of the form $P_m(t) = a_m t^m + a_{m-1}t^{m-1} + \cdots + a_1 t + a_0$, where t is a real variable and a_j, for $0 \leq j \leq m$, are complex numbers. Moreover, we can define the matrix polynomial $P_m(B)$ for $B \in \mathbb{C}^{r\times r}$ as $P_m(B) = a_m B^m + a_{m-1}B^{m-1} + \cdots + a_1 B + a_0 I$. As usual, the matrix norm $\|\cdots\|$ denotes any subordinate matrix norm; in particular $\|\cdots\|_1$ is the usual $1-$norm.

2 On Bernoulli matrix polynomials

The Bernoulli polynomials $B_n(x)$ are defined in [**?**, p.588] as the coefficients of the generating function

$$g(x,t) = \frac{te^{tx}}{e^t - 1} = \sum_{n\geq 0} \frac{B_n(x)}{n!} t^n \ , \ |t| < 2\pi, \tag{1}$$

where $g(x,t)$ is an holomorphic function in $\mathbb{C}$ for the variable t (it has an avoidable singularity in $t = 0$). Bernoulli polynomials $B_n(x)$ has the explicit expression

$$B_n(x) = \sum_{k=0}^{n} \binom{n}{k} B_k x^{n-k}, \tag{2}$$

where the Bernoulli numbers are defined by $B_n = B_n(0)$. Therefore, it follows that the Bernoulli numbers satisfy

$$\frac{z}{e^z - 1} = \sum_{n\geq 0} \frac{B_n}{n!} z^n \ , \ |z| < 2\pi, \tag{3}$$

where

$$B_0 = 1, B_k = -\sum_{i=0}^{k-1} \binom{k}{i} \frac{B_i}{k+1-i}, k \geq 1. \tag{4}$$

Note that $B_3 = B_5 = \cdots = B_{2k+1} = 0$, for $k \geq 1$. For a matrix $A \in \mathbb{C}^{r\times r}$, we define the $m - th$ Bernoulli matrix polynomial by the expression

$$B_m(A) = \sum_{k=0}^{m} \binom{m}{k} B_k A^{m-k}. \tag{5}$$

We can use the series expansion

$$e^{At} = \left(\frac{e^t - 1}{t}\right) \sum_{n\geq 0} \frac{B_n(A)t^n}{n!} \ , \ |t| < 2\pi, \tag{6}$$

to obtain approximations of the matrix exponential. A method based in (6) to approximate the exponential matrix has been presented in [13].

From (6), we obtain the following expression for the matrix cosine and sine:

$$\left.\begin{aligned} \cos(A) &= (\cos(1)-1)\sum_{n\geq 0}\frac{(-1)^n B_{2n+1}(A)}{(2n+1)!}+\sin(1)\sum_{n\geq 0}\frac{(-1)^n B_{2n}(A)}{(2n)!}, \\ \sin(A) &= \sin(1)\sum_{n\geq 0}\frac{(-1)^n B_{2n+1}(A)}{(2n+1)!}-(\cos(1)-1)\sum_{n\geq 0}\frac{(-1)^n B_{2n}(A)}{(2n)!}. \end{aligned}\right\} \tag{7}$$

Note that unlike the Taylor (and Hermite) polynomials that are even or odd, depending on the parity of the polynomial degree n, the Bernoulli polynomials do not verify this property. Thus, in the development of $\cos(A)$ and $\sin(A)$, all Bernoulli polynomials are needed (and not just the even-numbered ones).

Replacing in (6) the value t for it and $-it$ respectively and taking the arithmetic mean, we obtain the expression

$$\sum_{n\geq 0}\frac{(-1)^n B_{2n}(A)}{(2n)!}t^{2n}=\frac{t}{2\sin\left(\frac{t}{2}\right)}\left(\cos\left(tA-\frac{t}{2}I\right)\right)\ ,\ |t|<2\pi. \tag{8}$$

Taking $t=2$ in (8) it follows that

$$\cos(A)=\sin(1)\sum_{n\geq 0}\frac{(-1)^n 2^{2n} B_{2n}\left(\frac{A+I}{2}\right)}{(2n)!}, \tag{9}$$

Note that in formula (9) only even grade Bernoulli's polynomials appear.

3 Numerical Experiments

Having in mind expressions (7) and (9), two different approximations are given to compute cosine matrix function.

To test the proposed method and the two distinct approximations, and to compare them with other approaches, the following algorithms have been implemented on MATLAB R2018b:

- *cosmber*. New code based on the new developments of Bernoulli matrix polynomials (formulae (7) and (9)). The maximum value of m to be used is $m=36$, with even and odd terms.
- *cosmtay*. Code based on the Taylor series for the cosine [8]. It will provide a maximum value of $m=16$, considering only the even terms, which would be equivalent to $m=32$ using the even and odd terms.
- *cosmtayher*. Code based on the Hermite series for the cosine [10]. As mentioned before, it will provide a maximum value of $m=16$.
- *cosm*. Code based on the Padé rational approximation for the cosine [7].

The following sets of matrices have been used:

a) Diagonalizable matrices. The matrices have been obtained as $A = V \cdot D \cdot V^T$, where D is a diagonal matrix (with complex or real values) and matrix V is an orthogonal matrix, $V = H/16$, where H is a Hadamard matrix. We have $2.18 \leq \|A\|_1 \leq 207.52$. The matrix cosine is exactly calculated as $\cos(A) = V \cdot \cos(D) \cdot V^T$.

b) Non-diagonalizables matrices. The matrices have been computed as $A = V \cdot J \cdot V^{-1}$, where J is a Jordan matrix with complex eigenvalues with module less than 10 and random algebraic multiplicity between 1 and 5. Matrix V is a random matrix with elements in the interval $[-0.5, 0.5]$. We have $1279.16 \leq \|A\|_1 \leq 87886.4$. The matrix cosine is exactly calculated as $\cos(A) = V \cdot \cos(J) \cdot V^{-1}$.

c) Matrices from the Matrix Computation Toolbox [14] and from the **Eigtool Matlab package** [15]. These matrices have been chosen because they have more varied and significant characteristics.

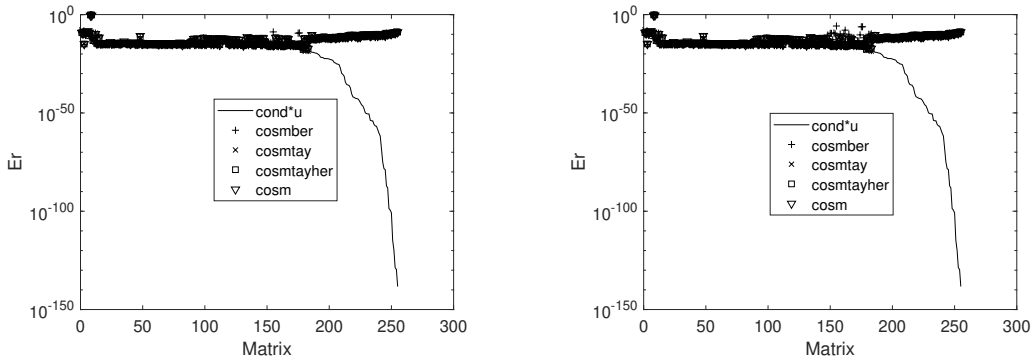
In the numerical test, we used 259 matrices of size 128×128: 100 from the diagonalizable set, 100 from the non-diagonalizable set, 42 from Matrix Computation Toolbox and 17 from Eigtool Matlab package. Results are given in Tables 1 and 2. The rows of each table show the percentage of cases in which the relative errors of `cosmber` (Bernoulli) is lower, greater or equal than the relative errors of `cosmtay` (Taylor), `cosmtayher` (Hermite) and `cosm` (Padé). Graphics of the Normwise relative errors and the Performance Profile are given in Figures 1 and 2. The total number of matrix products was: 3202 (*cosmber*), 2391 (*cosmtay*), 1782 (*cosmtayher*) and 3016 (*cosm*). Recall that in the Bernoulli implementation, the maximum value of m to be used was $m = 36$ considering all the terms and, in the rest of algorithms, was $m = 32$ but just having into account the even terms.

$E(cosmber) < E(cosmtay)$	55.60%
$E(cosmber) > E(cosmtay)$	44.40%
$E(cosmber) = E(cosmtay)$	0%
$E(cosmber) < E(cosmtayher)$	50.97%
$E(cosmber) > E(cosmtayher)$	49.03%
$E(cosmber) = E(cosmtayher)$	0%
$E(cosmber) < E(cosm)$	76.83%
$E(cosmber) > E(cosm)$	23.17%
$E(cosmber) = E(cosm)$	0%

Table 1: Using approximation (7)

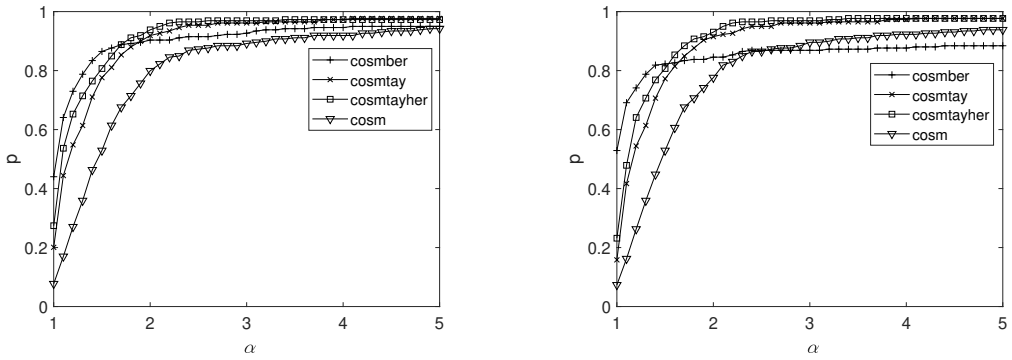
$E(cosmber) < E(cosmtay)$	65.64%
$E(cosmber) > E(cosmtay)$	34.36%
$E(cosmber) = E(cosmtay)$	0%
$E(cosmber) < E(cosmtayher)$	60.62%
$E(cosmber) > E(cosmtayher)$	39.38%
$E(cosmber) = E(cosmtayher)$	0%
$E(cosmber) < E(cosm)$	73.75%
$E(cosmber) > E(cosm)$	26.25%
$E(cosmber) = E(cosm)$	0%

Table 2: Using approximation (9)



(a) Using formula (7) (b) Using formula (9)

Figure 1: Normwise relative errors.



(a) Using formula (7) (b) Using formula (9)

Figure 2: Performance Profile.

4 Conclusions

In general, the implementation based on the new Bernoulli series (9) is more accurate than (7), comparing it with the one based on the Taylor series, algorithm (`cosmtay`) and Hermite series, algorithm (`cosmtayher`), and the one based in Padé rational approximation, algorithm (`cosm`).

Acknowledgement

This work has been partially supported by Spanish Ministerio de Economía y Competitividad and European Regional Development Fund (ERDF) grants TIN2017-89314-P and by the Programa de Apoyo a la Investigación y Desarrollo 2018 of the Universitat Politècnica de València (PAID-06-18) grants SP20180016.

References

[1] Serbin, S.M. and Blalock, S.A., An algorithm for computing the matrix cosine. *SIAM Journal on Scientific Computing*, 1(2): 198–204, 1980.

[2] Dehghan, M. and Hajarian, M., Computing matrix functions using mixed interpolation methods. *Mathematical and Computer Mod- elling*, 52(5-6): 826–836, 2010.

[3] Higham, N.J., *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, PA, USA, 2008.

[4] Alonso-Jordá, P., Peinado, J., Ibáñez, J., Sastre, J. and Defez, E., Computing matrix trigonometric functions with GPUs through Matlab. *The Journal of Supercomputing*, pages 1–14, 2018.

[5] Tsitouras, Ch. and Katsikis, V., Bounds for variable degree ratio- nal L_∞approximations to the matrix cosine. *Computer Physics Communications*, 185(11): 2834–2840, 2014.

[6] Serbin, S.M., Rational approximations of trigonometric matrices with application to second-order systems of differential equations. *Applied Mathematics and Computation*, 5(1): 75–92, 1979.

[7] Al-Mohy, A.H., Higham, N.J. and Relton, S.D., New algorithms for computing the matrix sine and cosine separately or simultaneously. *SIAM Journal on Scientific Computing*, 37(1): A456–A487, 2015.

[8] Sastre, J., Ibáñez, J., Alonso-Jordá, P., Peinado, J. and Defez, E., Two algorithms for computing the matrix cosine function. *Applied Mathe- matics and Computation*, 312: 66–77, 2017.

[9] Sastre, J., Ibáñez, J., Alonso-Jordá, P., Peinado, J. and Defez, E., Fast Taylor polynomial evaluation for the computation of the matrix cosine. *Journal of Computational and Applied Mathematics*, 354: 641–650, 2019.

[10] Defez, E., Ibáñez, J., Peinado, J., Sastre, J. and Alonso-Jordá, P., An efficient and accurate algorithm for computing the matrix cosine based on new Hermite approximations. *Journal of Computational and Applied Mathematics*, 348: 1–13, 2019.

[11] Kouba, O., Lecture Notes, Bernoulli Polynomials and Applications. *arXiv preprint arXiv*:1309.7560, 2013.

[12] WJ Olver, F., W Lozier, D., F Boisvert, R. and W Clark, C., *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge University Press, 2010.

[13] Defez, E., Ibáñez, J., Peinado, J., Alonso-Jordá, P. and Alonso, J.M., Computing matrix functions by matrix Bernoulli series. In *19th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE-2019)*, From 30th of Juny to 6th of July 2019. Poster presented at some conference, Rota (Cádiz), Spain.

[14] Higham, N.J., *The test matrix toolbox for MATLAB (Version 3.0)*. University of Manchester Manchester, 1995.

[15] Wright, TG., Eigtool, version 2.1. URL: `web.comlab.ox.ac.uk/pseudospectra/eigtool`, 2009.

14

Modeling the political corruption in Spain

Elena de la Poza-Plaza[♭1], Lucas Jódar[♮] and Paloma Merello[♯]

(♭) Centro de Ingeniería Económica,
Universitat Politècnica de València,
(♮) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València,
(♯) Department of Accounting,
University of Valencia.

1 Introduction

Political corruption is a universal problem, but it affects countries very differently, because it is influenced by culture and religion, the political system and the parties' law, the time in political office, the professional experience of politicians, the highly protectionist legal system, the type and structure of the public administration, the short democratic tradition, the economy, the proportion of women participating in politics, the lack of independence among the judiciary and the media among others [15].

The effect of political corruption is corrosive because it deteriorates the image of the country, the confidence of foreign investors, the quality of life of citizens, and worsens the future of the country. The lack of trust in institutions generates a moral disengagement, which makes it easier for citizens to excuse the corruption of the political class, considering it as alien but at the same time using it as an excuse to commit it themselves. A kind of contagion is very counterproductive because of its social, economic and moral impact on society [1, 14]. The concept of political corruption is susceptible to ambiguity, so we must specify it. Political corruption is any act or omission, legal or illegal, of a person who, based on a public office (elected or appointed) embracing political position but also a position in a labor union or business association favors a particular interest causing public harm (not necessarily monetary) [8, 9].

The cases of demonstrated political corruption that have been echoed by the media in recent years due to their economic, judicial and social significance, are just the tip of the iceberg of a problem hidden by many interested parties and the shortage of means in the fight against it. The most important factors that explain the current situation are the party system and its laws, where political offices do not respond to the citizen, but to the political chief who has appointed him, where there is a lack of self-criticism, transparency, and accountability. Imperfect judicial and media independence does not favor the end of the problem, much less when political parties

[1]e-mail: elpopla@esp.upv.es

are not able to make decisions against their partisan interests, even when these decisions are for the good of society Spanish and the economic and social future of the country. Likewise, an intoxicating and generalized state of moral relaxation has been established in Spanish society that excuses the phenomenon of political corruption as inevitable and inherent to the political class and therefore irremediable. This thought not only does not slow down, but it perpetuates and amplifies the dimension of the problem [1, 3].

In this work we quantify the level of risk of committing political corruption of the population residing in Spain between 16 and 70 years old. In addition to classify the population according to their level of risk of committing political corruption, we also take into account their employment situation at the time of the analysis. Thus, we study the evolution of subpopulations over time during the period 2015-2023, taking into account the annual dynamic transits. The external variables that determine the transits of individuals between populations during the period of study are: elections, time in office, gender, moral disconnection, economy, religion and the effect of "revolving doors" [11]. The relevance of this study relies on reporting the problem to public authorities responsible for addressing policies to stop this trend.

2 Model

The dynamic population model [6, 7, 10, 12] quantifies the amount of people from 16-70 years old in risk to commit political corruption in Spain. Four levels of risk of committing political corruption are established: zero risk (people who do not hold or are in contact with public office), low risk (less than 10%), individuals likely to collaborate with public office (member of political parties, unions or business associations); medium risk (up to 25%) people who are public representatives directly elected, or indirectly and manage public budget; high risk (more than 50%) high positions that handle large budgets and/or decision-making capacity, remaining in office since previous Administration.

Thus, 5 types of work situation have been considered: $j = 1$ pre-labour (young people up to 26 years old); $j = 2$ unemployed (26,70); $j = 3$ employed by a private company aged (26,70); $j = 4$ employed by a public company or administration aged (26,70); and $j = 5$ civil servant (26,70). Hence, the target population is divided into 20 subpopulations, taking into account their level of risk of committing political corruption and their alternative or complementary professional life to hold public office.

$\rightarrow$ $Z_j(n)$= zero risk subpopulation.

$\rightarrow$ $B_j(n)$= Low-risk subpopulation.

$\rightarrow$ $M_j(n)$= Medium-risk subpopulation.

$\rightarrow$ $A_j(n)$= High risk subpopulation.

The individuals transit to lower or higher levels of probability to commit political corruption by the conjunction of factors; those factors are explained by vector transits: demography (birth & death rates), time in office, contagion effect, elections, fear to loss the office, revolting doors effect but also by environmental factors: gender, culture & religion, economy, lack of political

transparency, controlled press, lack of independent justice.

Regarding the demographic transit, the variables considered are the birth rate (I_{ij}), the death rate (d_{ij}), and the retirement rate, individuals who retired or become over 70 years old, (R_{ij}). (Spanish Statistics Institute). These transit coefficients are assumed constant for the period of study (2015-2023).

Following is the economic transit explained by the political disenchantment, which drives to the loss of members of traditional political parties and unions. $\gamma = 0.81 \cdot 0.01 = 0.0081$. This transit coefficient affects $B_j(n)$ subpopulation that transit to $Z_j(n)$. This transit is assumed constant for the period of study. Next it is the change in election results μ, which explains the $A_j(n)$ individuals transit to $B_j(n)$ but also $B_j(n)$ individuals transit to $M_j(n)$ because of the emergence of new political parties [4]. This transit only takes place the next year after general elections (2016, 2020). It is assumed 40% position remain in office [11]. Related to the time in office of politicians, its effect is double: politicians who do not keep their seat (60%) transit from $A_j(n)$ to $B_j(n)$ but 50% politicians who keep their seat transit to higher categories $M_j(n)$ to $A_j(n)$. The transit coeffiecient increases progressively to the closer time to elections.

In addition, the fear to loss the seat of individuals impacts negatively on their probability to commit political corruption for $j = 2, 3, 4$. (ρ_{ij}). This transit affects to $B_j(n)$ individuals who transit to $M_j(n)$ but also $M_j(n)$ individuals who transit to $A_j(n)$. The transit coeffiecient increases progressively to the closer time to elections. Other transit coefficient is the moral disengagement (α_i) experienced by individuals which makes them transit to a higher risk category. $\alpha_Z = 0.005 \cdot 0.9 = 0.0045$; $\alpha_B = \alpha_M = 3\alpha_Z = 0.135$. This transit affects 90% population [1, 2, 5, 15].

Finally, it is considered the revolting doors effect (D_{A_j}) explained by those politicians who leave their political seat and join a board corporation, mainly belonging to the IBEX35. This transit affects $j = 2, 3, 4$. Approximately represents 23 positions per year transit [11].

Following, the compartment dynamic model to quantify the precarious population is expressed:

$$\begin{aligned}
Z_j(n+1) - Z_j(n) = {} & (I_{Z_1} - R_{Z_j}) - d_j(n)Z_j(n) - \alpha_Z Z_j(n) + \gamma B_j(n) \\
B_j(n+1) - B_j(n) = {} & (I_{B_1} - R_{B_j}) - d_j(n)B_j(n) - \alpha_B B_j(n) + \alpha_Z Z_j(n) \\
& - \rho_{B_j} B_j(n) + D_{A_j} - \gamma B_j(n) + \mu A_j(n) \\
M_j(n+1) - M_j(n) = {} & -R_{M_j} - d_j(n)M_j(n) + \alpha_B B_j(n) - \alpha_M M_j(n) - \tau_j M_j(n) \\
& - \rho_{M_j} M_j(n) + \rho_{B_j} B_j(n) \\
A_j(n+1) - A_j(n) = {} & -d_j(n)A_j(n) - \mu A_j(n) + \tau_j M_j(n) + \alpha_M M_j(n) \\
& + \rho_{M_j} M_j(n) - R_{A_j} - D_{A_J}
\end{aligned} \tag{1}$$

3 Results

By computing the model, the subpopulation values are estimated for each year. Table 1 shows the results at the beginning of the study, $n = 1$ after the political renovation occurred in 2016

(general elections were in 2015). Table 2 shows the results at the end of the study, $n = 8$ (2023).

	TOTAL	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$\boldsymbol{Z_j}$	20,648,347	4,206,758	5,291,150	10,355,955	228,273	566,212
$\boldsymbol{B_j}$	2,901,249	97,178	28,352	1,067,261	795,750	912,708
$\boldsymbol{M_j}$	346,885	1,051	0	249,704	67,587	28,543
$\boldsymbol{A_j}$	33,930	0	0	7,904	25,739	287
TOTAL	23,930,412	4,304,987	5,319,502	11,680,824	1,117,349	1,507,749

Table 1: Subpopulations forecast at $n = 1$, (2016).

	TOTAL	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$\boldsymbol{Z_j}$	20,250,555	3,329,507	5,900,978	9,382,973	1,136,106	500,991
$\boldsymbol{B_j}$	2,684,850	127,717	250,884	1,036,815	599,419	670,015
$\boldsymbol{M_j}$	441,168	8,039	16,038	233,271	103,937	79,882
$\boldsymbol{A_j}$	174,554	609	742	110,872	47,910	14,420
TOTAL	23,551,127	3,465,872	6,168,643	10,763,931	1,887,372	1,265,309

Table 2: Subpopulations forecast at $n = 8$, (2023).

Results show how the population at high risk to commit political corruption grows for the period of study representing 0.7% of the Spanish population in 2023. Even when this percentage can seem low, the socio-economic and moral impact on the Spanish society is dramatic.

4 Conclusions

The study quantifies the population at risk of committing political corruption in Spain by identifying and quantifying the drivers explaining the political corruption. The stop to this social problem requires the policy makers' action. In concrete, it is advisable the change of the electoral law of parties to increase the transparency and the accountability of politicians. It should be much more controlled in hiring "advisers" in office but also regulating the wages of local administration (Small city councils).

Also, it is necessary to make cuts on public funding for entities of doubtful nature such as certain non-profit organizations and/or public companies. Finally, increase of funding for the judicial system (district attorneys and judges).

References

[1] Bandura, A., Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3: 193–209, 1999.

[2] Christakis, NA. and Fowler, JH., *Connected: the Surprising Power of Our Social Networks and How They Shape Our Lives*. Boston, MA, USA: Little Brown and Company, 2009.

[3] Damasio, A., *The Strange Order of Things. Life, Feelings, and the Making of Cultures*. New York, USA: Pantheon Books, 2018.

[4] De la Poza, E., Jódar, L. and Pricop, A., Mathematical modeling of the propagation of democratic support of extreme ideologies in Spain:Causes, Effects and Recommendations for its stop. *Abstract and Applied Analysis*, 2013. `http://dx.doi.org/10.1155/2013/729814`

[5] Girard, R., *Mimesis and Theory: Essays on Literature and Criticism*, 1953–2005. Palo Alto, CA, USA: Stanford University Press; 2008.

[6] Goldthorpe, J.H., *Sociology as a Population Science*. Cambridge: Cambridge Univ. Press, 2016.

[7] Haddad, WM., Chellaboina, V. and Nersesov, SG., Hybrid nonnegative and compartmental dynamical systems, *Math. Probl. Eng.* 8(6): 493–5, 2002.

[8] La Porta, R. López de Silanes, F., Shleifer, A. and Visnhy, R., Trust in Large Organizations, *American Economic Review*, 87(2): 333-8, 1997.

[9] Lederman, D., Loayza, N. and Soares, R., Accountability and Corruption: Political Institutions Matter, *Economics & Politics*, 17(1): 1-35, 2005.

[10] MacCluer, CR., *Industrial Mathematics, Modeling in Industry, Science and Government*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[11] Montero, L.M., *El club de las puertas giratorias*, La Esfera de los libros, Madrid, 2016.

[12] Raafat, RM., Chater, N. and Frith, C., Herding in Humans. *Trends Cogn Sci*. 13(10): 420–428, 2009.

[13] Spanish Institute of Statistics `www.ine.es`.

[14] Standing, G., *The Precariat. The New Dangerous Class*. London: Bloomsbury, 2012.

[15] Veblen, T., *The Theory of the Leisure Class*. New York, USA: Macmillan, 1899.

15

Exponential time differencing schemes for pricing American option under the Heston model

R. Company [♭], F. Fuster[♮1] and L. Jódar [♭]

([♭]) Institut Universitari de Matemàtica Multidisciplinar,
Universitat Politècnica de València,
([♮]) Banco Santander,
Av. de Cantabria, s/n, 28660 Boadilla del Monte, Madrid.

1 Introduction

The classic Black-Scholes model makes assumptions that are not empirically valid. The model is widely employed as a useful approximation to reality, but proper application requires understanding its limitations and constant volatility of the stock returns is one of them. In fact, this assumption is one of the biggest source of weakness, because the variance has been observed to be non-constant leading to models, such as GARCH, to model volatility changes. There are other approaches to model the asset volatility, as consider that follows a random process or, in other words, consider the volatility as a stochastic process. This point of view lead us to a Partial Differential Equation (PDE) different from the classic Black-Scholes, now there are involved two different variables, apart of the time: asset level S and variance ν. Deal with this PDE and the presence of cross-derivatives is a challenging task. It is even more difficult to deal with American options which allows to exercise the option at any time before the expiration date. But the solution to this problem is of great interest to the financial markets.

2 The pricing problem

To the pricing of American options we use the Heston model [5]:

$$\begin{aligned} dS(t) &= \mu S(t)dt + \sqrt{\nu(t)}S(t)dW_1, \\ d\nu(t) &= \kappa(\theta - \nu(t))dt + \sigma\sqrt{\nu(t)}dW_2, \\ dW_1 dW_2 &= \rho dt, \end{aligned} \tag{1}$$

and a penalty method similar as in [3]. With this assumptions, applying Itô's lemma and standard arbitrage arguments we achieve the following PDE:

$$\frac{\partial U}{\partial t} + \frac{1}{2}\nu S^2\frac{\partial^2 U}{\partial S^2} + \rho\sigma\nu S\frac{\partial^2 U}{\partial S\partial\nu} + \frac{1}{2}\sigma^2\nu\frac{\partial^2 U}{\partial\nu^2} + rS\frac{\partial U}{\partial S} + \bar{\kappa}(\bar{\theta} - \nu)\frac{\partial U}{\partial\nu} - rU + f(E, S, U) = 0, \tag{2}$$

[1]e-mail: ferran.ffv@gmail.com

at which we will remove the cross-derivatives with the classical technique for the reduction of second order linear PDE to canonical form [4, chapter 3]. It is well known that, using finite differences, cross-derivatives involves negative coefficients. So, like we are talking about prices we must guarantee the solution's positivity. This fact motivates the transformation of the problem.

The following step of the semi-discretization. We apply centered finite difference to the spatial derivatives, letting alone the temporal-derivatives, achieving a system of ODEs:

$$\frac{dP}{dt} = A(\xi)P(t) + f\left(\xi, P\right). \tag{3}$$

Now we apply the ETD method [2] and the temporal discretization. Finally, making some assumptions to provide solutions, we achieve a numerical scheme to the PDE (2):

$$P^{n+1} = e^{Ak}P^n + k\ \varphi(A,k)\ f(\xi, P^n). \tag{4}$$

3 Positivity and stability

Like we are computing prices, we must assure the positivity an stability of the provided solutions. And in the case that we were interested in computing put prices, we also must assure that our numerical scheme provides bounded profits.

We can assure the positivity of our numerical scheme bounding the numerical derivative's step-size of the spatial variables. Specifically:

$$h \leq \frac{\alpha}{\delta}, \tag{5}$$

where α is the minimum main diagonal coefficient of matrix $A(\xi)$ and δ the maximum of non-diagonal elements.

The stability condition is fulfilled if the temporal step-size verify the following:

$$k \leq \frac{h^2}{(\lambda + r)h^2 + 2\alpha_m\left(\frac{1+m^2}{m^2}\right)}, \tag{6}$$

where α_m is the maximum main diagonal coefficient of matrix $A(\xi)$, r the risk-free rate, m the relationship between the spatial step-sizes and λ a constant dependent of the penalty term.

It can be verified for put options, using the induction principle, that at any time step:

$$\parallel P^n \parallel_\infty \leq E. \tag{7}$$

4 Numerical experiments

Fig. 1 shows the numerical solution for American put options under the set of parameters: $S_1 = 0.25,\ S_2 = 40,\ \nu_1 = 0.002,\ \nu_2 = 1.2,\ r = 0.1,\ \rho = 0.1,\ E = 10,\ T = 0.25,\ \lambda = 200,\ \kappa =$

5, $\theta = 0.16$, $\sigma = 0.9$ for k and h verifying the stability condition.

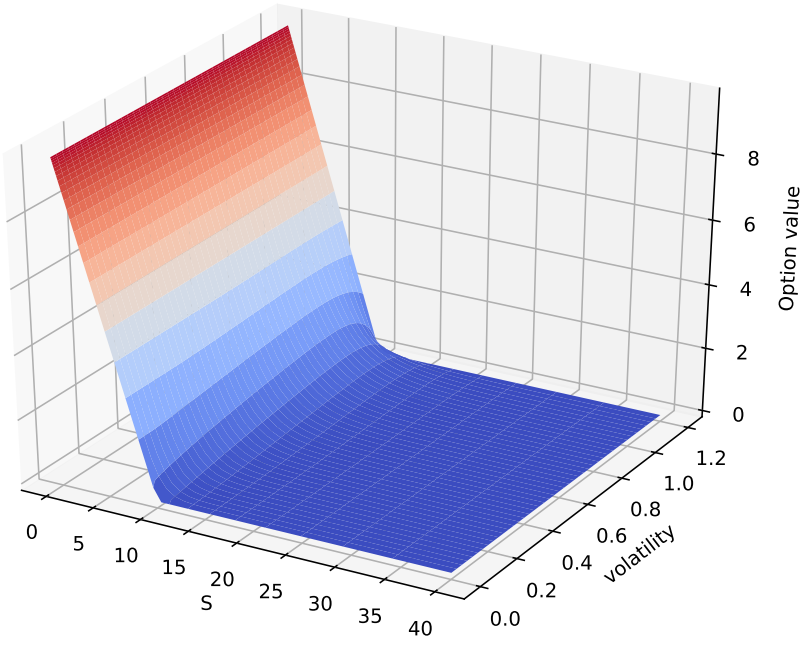


Figure 1: Numerical solution for $\tau = T$, $h = 0.07$ and $k = 5 \cdot 10^{-5}$.

We can see that for a big values of the underlying asset, the option values tends to zero. On the other hand, when the asset tends to zero the option value tends to the strike price E, as we expect because of (7). Other relevant issue that our numerical solution catches is that for a big values of the volatility the option value is bigger than for low values, but this is only relevant when the asset is near to the strike price. Proposed numerical solution are competitive with other approaches in the literature [1, 6–10].

Acknowledgements

This work has been partially supported by the Ministerio de Ciencia, Innovación y Universidades Spanish grant MTM2017-89664-P.

References

[1] Clarke, N. and Parrott, K. The multigrid solution of two-factor American put options. *Oxford Computing Laboratory, Research Report*, 96-16, 1996.

[2] Cox, S.M. and Matthews, P.C. Exponential Time Differencing for Stiff Systems. *Journal of Computational Physics*, 176(2):430-455, 2002.

[3] Forsyth, P. A. and Vetzal, K. R. Quadratic Convergence for Valuing American Options Using a Penalty Method. *SIAM Journal on Scientific Computing*, 23(6):2095-2122, 2002.

[4] Garabedian, P. R. *Partial Differential Equations*. Springer Berlin Heidelberg, 1998.

[5] Heston, S.L. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *Review of Financial Studies*, 6(2):327-343, 1993.

[6] Ikonen, S. and Toivanen, J. Efficient numerical methods for pricing American options under stochastic volatility. *Numerical Methods for Partial Differential Equations*, 24(1):104-126, 2007.

[7] Oosterlee, C.W. On multigrid for linear complementarity problems with application to American-style options. *ETNA. Electronic Transactions on Numerical Analysis [electronic only]*, 15:165-185, 2003.

[8] Yousuf, M. and Khaliq, A.Q.M. An efficient ETD method for pricing American options under stochastic volatility with nonsmooth payoffs. *Numerical Methods for Partial Differential Equations*, 29(6):1864-1880, 2013.

[9] Zhu, S.-P and Chen, W.-T. A predictor–corrector scheme based on the ADI method for pricing American puts with stochastic volatility. *Computers & Mathematics with Applications*, 62(1):1-26, 2011.

[10] Zvan, R., Forsyth, P. and Vetzal, K. Penalty methods for American options with stochastic volatility. *Journal of Computational and Applied Mathematics*, 91(2):199-218, 1998.

16

Chromium layer thickness forecast in hard chromium plating process using gradient boosted regression trees: a case study

P.J. García Nieto[♭1], E. García Gonzalo[♭], F. Sánchez Lasheras[♭] and G. Fidalgo Valverde[♮]

(♭) Department of Mathematics, Faculty of Sciences,
University of Oviedo,
(♮) Department of Business Administration, ETSIMO,
University of Oviedo.

The hard chromium plating process aims at creating a coating of hard and wear-resistant chromium with a thickness of some micrometers directly on the metal part without the insertion of copper or nickel layers. Chromium plating features high levels of hardness and resistance to wear and it is due to these properties that they can be applied in a huge range of sectors. The hard chromium plating process is one of the most effective ways of protecting the base material against a hostile environment or improving the surface properties of the base material. However, in the electroplating industry, electroplaters are faced with many problems and undesirable results with chromium plated materials. These common problems include matt deposition, milky white chromium deposition, rough or sandy chromium deposition and insufficient thickness and hardness. This paper presents a nonparametric machine learning approach using a gradient boosted regression tree model (GBRT) for prediction of the thickness of the layer in a hard chromium plating process. The optimization of the GBRT hyper-parameters was performed using the Differential Evolution (DE) technique. GBRT model is a powerful machine learning algorithm that seeks and obtains good predictions in a wide range of data-driven nonlinear problems, like the one treated here, where the studied variable presents low concentrations mixed with high concentration peaks. Two types of results have been obtained: firstly, the model allows the ranking of the dependent variables according to its importance in the model. Finally, the high performance of the model makes the gradient boosted tree (GBRT) method attractive compared to other conventional forecasting machine learning techniques.

Keywords: Gradient boosting regression trees (GBRTs); Differential evolution (DE); Hard chromium plating process; Regression analysis

[1]e-mail: lato@orion.ciencias.uniovi.es

1 Introduction

The hard chromium plating process under study in the present work is widely used on many mechanical parts and plastic moulds due to its good mechanical properties, good aesthetic appearance and superior resistance to corrosion [1]. Although it is well known that hard chromium plating is a process that has serious disadvantages from an environmental point of view, its replacement is not a simple matter due to the great performance of the chromium plated pieces. This process consists mainly of three operations (see Fig. 1), together with some intermediate quality inspections [1]:

- Vapour degreasing: This is mainly a cleaning operation. It is performed over the piece in order to assure the cleanliness of the surface.
- Electropolishing: It is an electrochemical process that removes material from the work piece. The performance of this operation before the hard chromium plating, helps to ensure a good roughness of the surface that will be coated.
- Hard chromium plating: This is the operation on which a thin layer of chromium is deposited onto the workpiece.

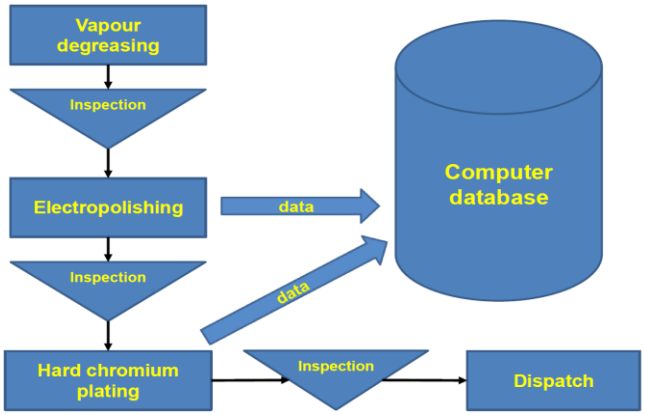


Figure 1: Scheme of the hard chromium plating industrial process.

The aim of the present work is to create a gradient boosted regression tree (GBRT) predictive model [2] capable of predicting the thickness of the chromium layer deposited over the pieces taking into account not only those variables related with the hard chromium plating operation but also with the previous process of electropolishing. The optimization of the GBRT hyperparameters has been performed using the Differential Evolution (DE) technique [3].

2 Mathematical model

Gradient boosting is a machine learning technique for regression and classification problems, which produces a model able to predict in the form of an ensemble of weak prediction models, typically decision trees [2]. Gradient boosting combines weak learners into a single strong

learner in an iterative fashion. It is easiest to explain in the least-squares regression setting, where the goal is to teach a model F to predict values of the form $\hat{y} = F(x)$ by minimizing *the mean squared error*, $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ where i indexes over some training set of size n of actual values of the output variable y.

In order to optimize the GBRT hyperparameters, the Differential Evolution (DE) technique [3] has been employed. In evolutionary computation, differential evolution (DE) is a method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. Such methods are commonly known as metaheuristics as they make few or no assumptions about the problem being optimized and can search very large spaces of candidate solutions. DE is used for multidimensional real-valued functions but does not use the gradient of the problem being optimized, which means that it does not require the optimization problem to be differentiable, as is required by classic optimization methods such as gradient descent and quasi-newton methods.

3 Results and discussion

Table 1 shows the determination and correlation coefficients for the hybrid DE/GBRT–based model fitted for the thickness of hard chromium layer in this manuscript.

Model	**Coef. of determination (R^2)/correlation coef. (r)**
DE/GBRT	0.9882/0.9941

Table 1: Coefficient of determination (R^2) and correlation coefficient (r) for the hybrid DE/GBRT–based model fitted in this study for the thickness of the hard chrome layer.

Next, Fig. 2 indicates the comparison between the observed and predicted values of the thickness of the hard chrome layer by using this hybrid DE/GBRT–based model.

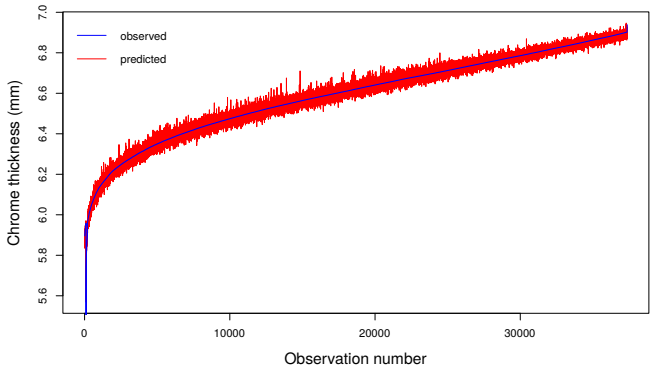


Figure 2: Comparison between the chrome thickness values observed and predicted by using the DE/GBRT–based model ($R^2 = 0.9882$).

4 Conclusions

The principal findings of this study are given as follows:

- The foretold values of the thickness of the hard chrome layer are in concordance with the observed ones since applying this DE/GBRT model, high coefficient of determination equal to 0.9882 was accomplished.

- The DE/GBRT–based model used the XGBoost algorithm [4] in combination with the DE optimization technique [5]. Please note that XGBoost is also termed regularized boosting technique as its implementation means regularization and thus it is helpful in order to reduce overfitting.

References

[1] Schlesinger, M. and Paunovich, M., *Modern Electroplating*. New York, Wiley-Interscience, 2000.

[2] Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Berlin, Springer, 2009.

[3] Storn, R. and Price, K., Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces *J. Global Optim.*, 11:341–359, 1997.

[4] Chen, T. and Guestrin, C., XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 785–794, 2016.

[5] Ardia, D., Mullen K.M., Peterson, B.G. and Ulrich, J., DEoptim: Differential evolution in R. R package, version 2.2-4. *R Foundation for Statistical Computing,* 2016.

17

Design and convergence of new iterative methods with memory for solving nonlinear problems

F.I. Chicharro[♭], A. Cordero[♮], N. Garrido[♮1] and J.R. Torregrosa[♮]

(♭) Escuela Superior de Ingeniería y Tecnología,
Universidad Internacional de La Rioja,
(♮) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

One of the most common applied problems appearing in any scientific field is to calculate a solution of a nonlinear system of equations, i.e., the problem to obtain the solution $x^* \in \mathbb{R}^n$ of $F(x) = 0$, where F is a nonlinear function, $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$, sufficiently differentiable in an open convex set D. Iterative methods have shown to be a good tool in order to approximate the solution of this kind of problems. Starting from an initial estimation, they generate a sequence of iterates that approximates the unknown x^*. There are many works in the literature devoted to the design of new iterative methods for the scalar case. Good overviews can be found in [2] and [4]. However, for the multidimensional case the amount of works is more limited.

Iterative schemes can be compared taking into account different criteria: their order of convergence p, the number of functional evaluations d done on each iteration of the method, the computational cost, among others. In addition, the efficiency of a method can be measured by relating these criteria through indices such as the efficiency index presented by Ostrowski in [1]. This index, defined by $I = p^{1/d}$, provides a relationship between the order of convergence and the number of functional evaluations of a method, being a high value an indicator that the corresponding method is suitable for solving problems efficiently.

Focusing on the number of previous iterates needed, iterative schemes are classified as methods with or without memory, being the first ones those that use more than one previous iterate to obtain the following estimation. Many researchers have shown that an adequate introduction of memory on iterative methods for solving nonlinear equations produces an increasing of its order of convergence without adding new functional evaluations. The extension of these results to nonlinear systems is a research area still in development, see for example [5, 6].

[1]e-mail: neugarsa@upvnet.upv.es

2 Inclusion of memory on Traub-Steffensen's family

The starting point of this work is the Traub-Steffensen's family of iterative methods [2],

$$x^{(k+1)} = x^{(k)} - [w^{(k)}, x^{(k)}; F]^{-1} F(x^{(k)}), \qquad k = 0, 1, 2, \ldots, \tag{1}$$

where $w^{(k)} = x^{(k)} + bF(x^{(k)})$, and b is an arbitrary constant, $b \neq 0$. Note that $b = 1$ in (1) reduces to the well-known Steffensen's method [3].

Traub-Steffensen's family converges quadratically with independence on the value of b, as shows its error equation

$$e^{(k+1)} = C_2(I + bF'(x^*))(e^{(k)})^2 + \mathcal{O}((e^{(k)})^3), \tag{2}$$

where $e^{(k)} = x^{(k)} - x^*$ is the error on each iteration, I denotes the identity matrix and $C_j = \frac{1}{j!}[F'(x^*)]^{-1}F^{(j)}(x^*)$, $j \geq 2$. Let us also remark that (1) is a derivative-free scheme that only uses information from the current iteration, so defines a family of iterative methos without memory.

Following the same derivative-free iterative structure than (1), we can found in the literature the Kurchatov's iterative scheme [7]

$$x^{(k+1)} = x^{(k)} - [2x^{(k)} - x^{(k-1)}, x^{(k-1)}; F]^{-1} F(x^{(k)}), \qquad k = 1, 2, \ldots, \tag{3}$$

that is a quadratically convergent method with memory, as in the Kurchatov's divided difference operator, $[2x^{(k)} - x^{(k-1)}, x^{(k-1)}; F]$, the current and the previous iterates are used.

By studying the error equation (2), we develop two iterative methods with memory using the iterative structure (1) and reaching higher order of convergence after the inclusion of previous iterates. The inclusion of memory is made by a properly approximation of the parameter b and the use of the Kurchatov's divided difference operator.

First, we can observe from equation (2) that $b = -[F'(x^*)]^{-1}$ provides a method of the family that has order of convergence three. This value of the parameter cannot be used because of the unknown x^* but an approximation of it can be made. In this sense, we propose the use of the Kurchatov's divided difference operator as an approximation of $F'(x^*)$. Then, the following approximation for the parameter $b := B^{(k)}$ is made:

$$B^{(k)} = -[2x^{(k)} - x^{(k-1)}, x^{(k-1)}; F]^{-1}. \tag{4}$$

The replacement of parameter (4) in (1) gives rise to a method with memory that we have denoted by M3. The order of convergence of the method is set in the following result.

Theorem 1 *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^* the solution of $F(x) = 0$, such that F' is continuous and nonsigular in x^*. Let us suppose that $x^{(0)}$ and $x^{(1)}$ are close enough to x^*. Then, the sequence of iterates $\{x^{(k)}\}$ generated by method M3 converges to x^* with order of convergence three.*

Consequently, Method M3 is a scheme with memory belonging to the Traub-Steffensen's family that reaches higher order of convergence without adding new additional functional evaluations of the nonlinear function.

The second method developed in this work is obtained from the composition of the iterative expression of Traub-Steffensen's family, resulting in the multipoint scheme

$$\begin{array}{rcll} y^{(k)} & = & x^{(k)} - [w^{(k)}, x^{(k)}; F]^{-1} F(x^{(k)}) & \\ x^{(x+1)} & = & y^{(k)} - [w^{(k)}, y^{(k)}; F]^{-1} F(y^{(k)}), & k = 0, 1, 2, \ldots, \end{array} \tag{5}$$

being $w^{(k)} = x^{(k)} + bF(x^{(k)})$, $b \neq 0$. The next result shows the error equation of family (5) and its order of convergence.

Theorem 2 *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^* the solution of $F(x) = 0$, such that F' is continuous and nonsigular in x^*. When the initial estimation $x^{(0)}$ is close enough to x^*, the uniparametric family of iterative methods (5) has order of convergence three for any value of b and with error equation*

$$e^{(k+1)} = C_2(I + bF'(x^*))C_2(I + bF'(x^*))(e^{(k)})^3 + \mathcal{O}((e^{(k)})^4), \tag{6}$$

where $e^{(k)} = x^{(k)} - x^*$ and $C_j = \frac{1}{j!}[F'(x^*)]^{-1}F^{(j)}(x^*)$, $j \geq 2$.

The previous composition allows the design of multipoint methods without memory with cubic order of convergence. Althought the order is higher than in the original family, the number of functional evaluations in (5) is also higher, so an increase in the order of convergence is necessary in order to design methods more efficiently.

From the error equation (6), the term $I + bF'(x^*)$ shows that the same approximation for the parameter as previously can be made in order to grow up the order of the family. Then, when we set parameter (4) in the iterative expresion (5), and we denote the resulting method with memory by M5. Under the same conditions than in the previous theorems, method M5 has order of convergence five, as shows the result below.

Theorem 3 *Let $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a sufficiently differentiable function in an open convex set D. Let us denote by x^* the solution of $F(x) = 0$, such that F' is continuous and nonsigular in x^*. If $x^{(0)}$ and $x^{(1)}$ are close enough to x^*, method M5 converges to x^* with order of convergence five.*

After the theoretical analysis of the convergence of the proposed methods, in Section 3 we focus on the numerical results in order to compare computationally the performance of the methods.

3 Numerical results

In this section, the designed methods with memory M3 and M5 are tested numerically in order to show their performance for solving a nonlinear problem. The numerical results obtained for the proposed schemes are also compared with the results obtained for Kurchatov's method. The nonlinear system solved in the numerical implementation is the following system of ten nonlinear equations:

$$x_i - \cos\left(2x_i - \sum_{j=1}^{10} x_j\right) = 0, \qquad i = 1, 2, \ldots, 10.$$

The numerical tests are made using the software Matlab R2018b with variable precission arithmetics with 2000 digitis of matissa. As M3, M5 and Kurchatov's method are schemes with memory, two initial estimations are required. However, for the computational implementation we use, instead of $x^{(1)}$, an initial value for the parameter $B^{(0)} = -0.01I$ and an initial estimation $x^{(0)}$ to start the iterations of the methods, where I denotes the identity matrix. The iterative process begins with an initial $x^{(0)}$ and ends when the difference between two consecutive iterations $||x^{(k+1)} - x^{(k)}||$ or the value of the function in an iterate $||F(x^{(k+1)})||$ is lower than 10^{-50}, with a maximum of 50 iterations.

Table 1 summarizes the results for each method when different initial estimations $x^{(0)}$ are considered. For each method, we show the number of iterations required to reach the convergence (*iter*) and the values of the stopping criteria when the iterative process finishes. As is was expected, in all cases method M3 and M5 need less iterations, so they converge fastly. This results confirm that the proposed methods in this work have higher order of convergence than Kurchatov's method.

$\mathbf{x^{(0)}}$	**Method**	iter	$\|\|\mathbf{x^{(k+1)}} - \mathbf{x^{(k)}}\|\|$	$\|\|\mathbf{F(x^{(k+1)})}\|\|$
$\begin{pmatrix} 0.9 \\ \vdots \\ 0.9 \end{pmatrix}$	Kurchatov	7	9.301e–36	2.685e–53
	M3	5	7.201e–45	2.058e–62
	M5	3	4.453e–31	1.129e–65
$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$	Kurchatov	8	1.537e–40	4.392e–58
	M3	5	4.198e–41	1.2e–58
	M5	3	2.589e–22	6.553e–57
$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	Kurchatov	8	2.847e–39	8.157e–57
	M3	5	5.624e–38	6.37e–55
	M5	4	6.551e–49	1.302e–82

Table 1: Numerical results for the nonlinear system

4 Conclusions

From the quadratically convergent Traub-Steffensen's family, two new iterative schemes with memory are presented in this work with orders three and five. We show that the inclusion of memory on the original scheme and the composition of iterative structures allow to increase the order of convergence from two up to five. In addition, the numerical experiments performed in

Section 3 confirm the theoretical results of Section 2, as methods M3 and M5 converge fastly than Kurchatov's method.

Acknowledgements

This research was partially supported by MCIU/AEI/FEDER, UE PGC-2018-095896-B-C22 and Generalitat Valenciana PROMETEO/2016/089.

References

[1] Ostrowski, A.M., *Solution of equations and systems of equations*. Prentice-Hall, 1964.

[2] Traub, J.F., *Iterative methods for the solution of equations*. Chelsea Publishing Company, 1977.

[3] Steffensen, J.F., Remarks on iteration, *Scandinavian Actuarial Journal*, 1933 (1):64–72, 1933.

[4] Amat, S. and Busquier, S., *Anvances in iterative methods for nonlinear equations*, SEMA SIMAI Springer, 2016.

[5] Petković, M.S. and Sharma, J.R., On some efficient derivative-free iterative methods with memory for solving systems of nonlinear equations, *Numerical Algorithms*, 71:457–474, 2016.

[6] Ahmad, F., Soleymani, F., Haghani, F.K. and Serra-Capizzano, S., Higher order derivative-free iterative methods with and without memory for systems of nonlinear equations, *Applied Mathematics and Computation*, 314:199–211, 2017.

[7] Kurchatov, V.A., On a method of linear interpolation for the solution of functional equations, *Doklady Akademii Nauk SSSR (Russian)*, 198(3):524-526, 1971. Translation in Soviet Doklady Mathematics, 12:835–838, 1971.

18

Study of the influence falling friction on the wheel/rail contact in railway dynamics

J. Giner-Navarro[♭1], V. Andrés-Ruiz[♭], J. Martínez-Casas[♭] and F. D. Denia[♭]

(♭) Centro de Investigación en Ingeniería Mecánica,
Universitat de València.

1 Introduction

The complexity of the railway interaction comes from the coupling between the train and the track introduced through the forces appearing in the wheel/rail contact area. These forces are governed by the friction coefficient through the Coulomb's law, characterised by the static and kinematic values, although most of the contact models in railway dynamics consider a constant friction along the simulations. Nevertheless, it is well known that the friction coefficient falls with the slip velocity [1, 2] from a maximum point determined by the static value to a point of saturation corresponding to the kinematic value. The question under debate is if the slope of this fall since recent test-rig experiments seem to reduce it ostensibly compared to friction curves generally estimated in the literature [3].

Rudd [4] proposed this negative slope as mechanism responsible for the generation of an instability phenomenon called railway curve squeal, which has received special attention from researchers [5–9]. The self-excited oscillations that characterise this phenomenon occur when the train is passing along a narrow curve, generating a strong tonal noise in the high-frequency domain. Although falling friction is the most widely accepted mechanism, other possibilities have been proposed to explain, getting more credit the mode-coupling mechanism [10, 11]. For this instability, the oscillation frequencies of two structural modes of an undamped system come closer and closer together until they merge and a pair of an unstable and a stable mode results [12, 13].

This work proposes a model based on a mass-spring-damper oscillator to evaluate its stability when submitted to a variable friction curve. Considering a single-dof (degrees of freedom) model, the paper studies the unstable conditions of the slip-dependent friction that can make the steady-state unstable. The study is extended to a two-dof case with two different geometric configurations in order to analyse the influence of the geometric coupling between the normal and tangential directions arisen from the contact and if it may instabilise the system even considering constant friction.

[1]e-mail: juanginer@upv.es

2 Overview of the mathematical approach

Fig. 1 shows a single degree of freedom oscillator excited by friction over a moving belt [2]. As mentioned in the previous section, there exists two means to get sustained oscillations with an oscillator: either by a decreasing slope of the creepage-creep force phenomenological law, or by a variation of the vertical force applied to the moving mass. In the former case, the motion of the belt is transformed into self-excited vibrations of the mass. In the latter case, the mass is subjected to a forced vibration imposed by the variation of the vertical force.

First, consider the case of a decreasing slope. The equation of motion reads

$$m\ddot{x} + c\dot{x} + kx = F_x, \tag{1}$$

where m is the mass of the oscillator, k the stiffness of the spring, c the damping coefficient and F_x the creep force. The dependency between the relative speed between the mass and the belt and the friction force is given by the Coulomb's law

$$F_x = (\mu_s - \delta_\mu v_x) N_0 sign(v_x), \tag{2}$$

where $v_x = \frac{(V-\dot{x})}{V}$, and N_0 is the static load, and δ_μ the decreasing slope of the friction curve. From the convenient variable transformation, Eq. (1) can be adimensionalised and expressed as

$$q'' + 2\zeta q' + q = (\mu_s - \delta_\mu \widetilde{v}_x) sign(\widetilde{v}_x), \tag{3}$$

where $q = \frac{kx}{N}$, $q' = \frac{dq}{d\tau} = \omega_n \frac{dq}{dt}$, $\widetilde{v}_x = \widetilde{V} - q'$, $\widetilde{V} = \frac{kV}{\omega N}$, $\omega_n = \sqrt{\frac{k}{m}}$ is the natural frequency and $\zeta = \frac{c}{(2m\omega_n)}$ is the damping rate.

For a given dimensionless sliding velocity $\widetilde{V}$, the equilibrium state is associated with a stationary slip where the conveyor belt moves at speed $\widetilde{V}$ but not the oscillator ($q_0' = 0$). The equilibrium may be stable or unstable. As it is well known for this kind of friction-induced self-excited oscillator that the equilibrium state can undergo instability through a Hopf bifurcation leading to a cycle solution, i.e. a periodic vibration. This stability problem may be analysed by the first Lyapunov method reconsidering the problem in the phase space.

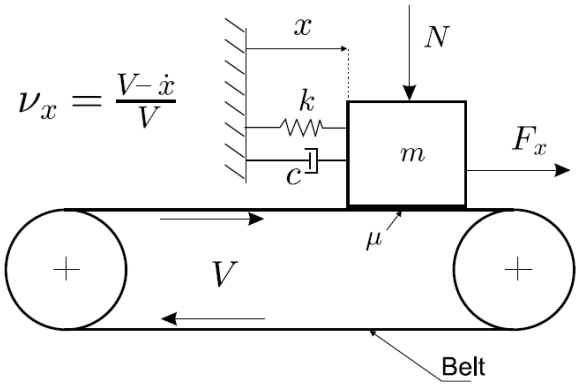


Figure 1: Single-dof oscillator excited by friction.

There is an increase of interest in the direction of mode-coupling phenomena in addressing curve squeal, which have been explained in a simplified form by Hoffmann et al. [10, 12] and Sinou and Jezequel [13], through frequency-domain models. This type of instability can occur even considering a constant coefficient of friction, arising from non-conservative displacement-dependent forces.

Fig. 2 shows the typical system adopted to illustrate this mechanism, in which the friction coefficient μ is constant. Here the mass has two dof and two springs. As the mass vibrates, variations in the normal load occur, leading to variations in the friction force. The modes of the wheel may have both vertical and lateral components and the contact angle of the wheel with the rail may vary. At least two modes are necessary to initiate this mechanism.

By considering small oscillations around the equilibrium of steady-state sliding, the system in Fig. 2 can be mathematically described as

$$\begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix} \begin{Bmatrix} \ddot{x} \\ \ddot{y} \end{Bmatrix} + \begin{pmatrix} k_{11} & k_{12} - \mu K_H \\ k_{21} & k_{22} \end{pmatrix} \begin{Bmatrix} x \\ y \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}, \tag{4}$$

where the terms $k_{ij}(\alpha_1, \alpha_2)$ in the stiffness matrix depend on the orientation and stiffness of the springs which in turn depend on angles α_1 and α_2 [13]. K_H represents the linearised Hertzian contact stiffness; x and y are the vibration displacements in tangential and normal directions, respectively, and F and N are the corresponding friction and normal forces. The most important feature of Eq. (4) is that the stiffness matrix is non-symmetric, making the system unstable if the upper diagonal term of the stiffness matrix $k_{12} - \mu K_H \leq 0$ due to the value of friction coefficient μ.

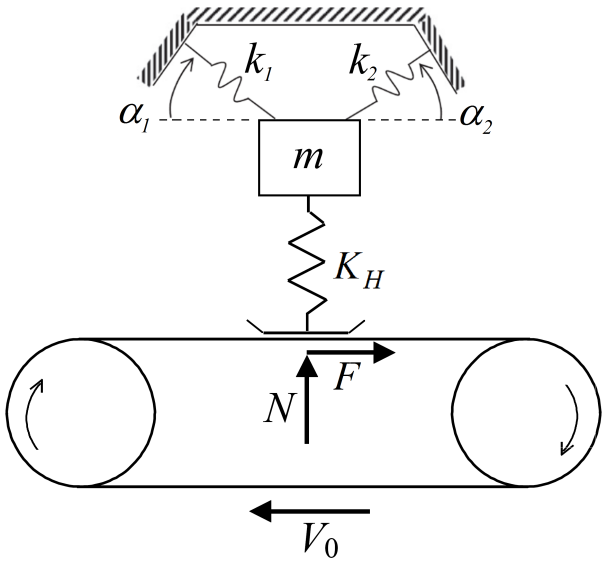


Figure 2: Two-dof system on moving belt.

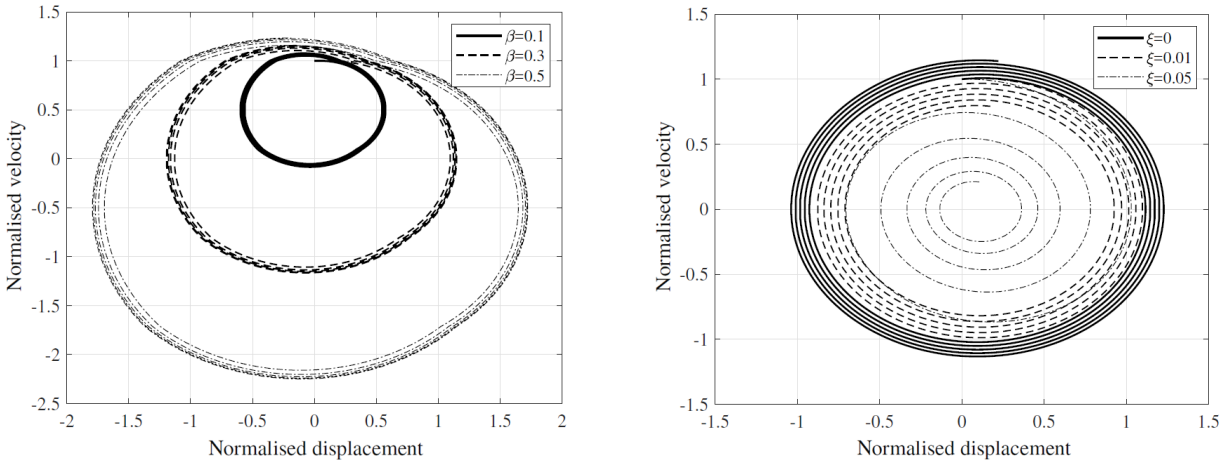
3 Results

As shown in [14], the following non-dimensional parameter indicates the relative importance of the stick and slip phases:

$$\beta = (\mu_s - \mu_k)\frac{N}{Vm\omega_n}. \tag{5}$$

This value is usually in the range 0.1–1 [14] for curve squeal situations. The parameter permits to evaluate the stick-slip motion of the single-dof oscillator, as seen in Fig. 3a for three values of β, in which the velocity (normalised by the belt velocity V_0) is plotted against the displacement (normalised by V_0/ω_0). It can be seen a 'limit cycle' as the formation of a stable periodic motion from different initial conditions. For small values of β, the slip phase predominates since the motion is close to elliptical on the phase plane and the oscillation frequency is close to the natural frequency. The stick phase predominates for large values of β and the oscillation frequency is lower than the natural frequency [14].

The effect of damping is also assessed in Fig. 3b for $\beta = 1$ and three values of damping ratio. It is observed a small effect on the amplitude of the limit cycle when the damping is increased, until the damping reaches a value where the oscillations are suppressed. For $\zeta = 0.05$ in this case, the damping exceeds the limiting value and the oscillations decay. The limiting value of damping ratio can be approximated as $\zeta > \beta^2/4\pi$ [14].

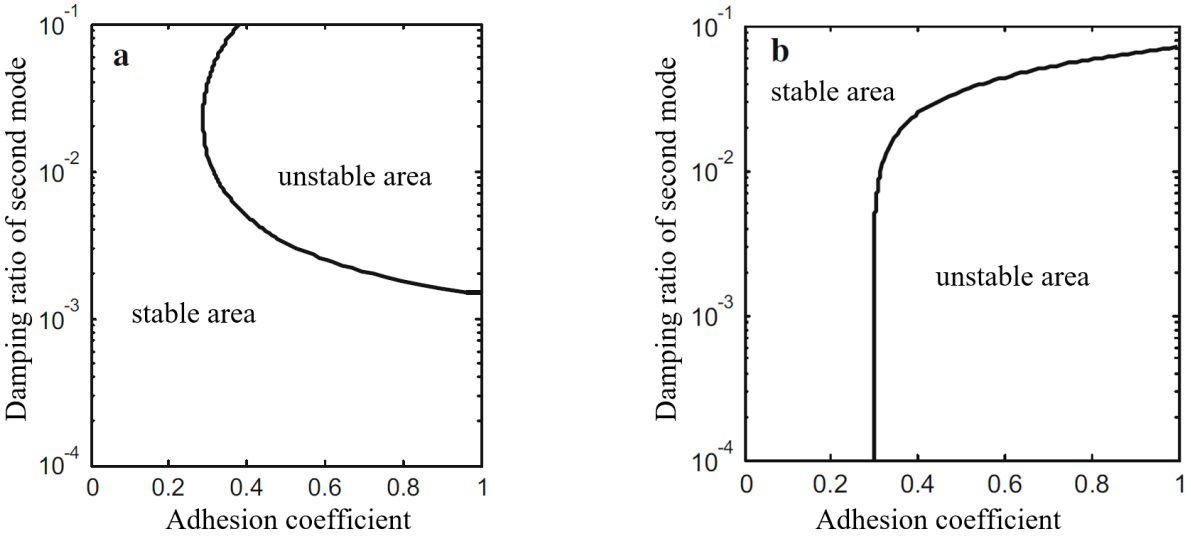


(a) (b)

Figure 3: Normalised displacement vs. non-dimensional velocity of a simplified stick-slip mechanism: $\mu_s = 0.4$, $\mu_k = 0.3$. (a) Without damping; (b) for different damping levels ($\beta = 1$).

Using now the two-dof oscillator model to assess the mode coupling, the effect of damping is evaluated. It can be observed from Fig. 4 that an increase in damping can favour instability in some situations or can improve stability in others. On the one hand, Fig. 4a shows the stability map for varying friction coefficient when the damping ratio of only the second mode of the system is varied, while the damping ratio of the first mode is kept at 10^{-4}. For low values of damping, the system remains stable. Nevertheless, it becomes more unstable when the damping of the second mode is between about 2×10^{-3} and 10^{-1}. On the other hand, increasing together the damping ratios of both modes while keeping their ratio fixed, Fig. 4b

shows that damping has no effect on the stability up to about 10^{-2}, while the system is quickly stabilised above this value.



(a) (b)

Figure 4: Stability maps for two-mode system for contact angle 3° and lateral contact position of 8 mm showing effect of damping ratio. (a) Damping ratio of second mode only is varied; (b) damping ratio of both modes is varied, keeping the ratio between them fixed.

Acknowledgements

The authors gratefully acknowledge the financial support of FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación (project TRA2017-84701-R), as well as Conselleria d'Educació, Investigació, Cultura i Esport (project Prometeo/2016/007) and European Commission through the project "RUN2Rail - Innovative RUNning gear soluTiOns for new dependable, sustainable, intelligent and comfortable RAIL vehicles" (Horizon 2020 Shift2Rail JU call 2017, grant number 777564).

References

[1] Koch, J.R., Vincent, N., Chollet, H. and Chiello, O., Curve squeal of urban rolling stock – Part 2: Parametric study on a 1/4 scale test rig, *Journal of Sound and Vibration*, 293: 701-709, 2006.

[2] Collette, C., Importance of the wheel vertical dynamics in the squeal noise mechanism on a scaled test bench, *Shock and Vibrations*, 19: 145-153, 2012.

[3] Alonso, A., Guiral, A., Baeza, L. and Iwnicki, S.D., Wheel–rail contact: experimental study of the creep forces–creepage relationships, *Vehicle System Dynamics 52 Suppl.*, 1: 469-487, 2014.

[4] Rudd, M.J., Wheel/rail noise – Part II: Wheel squeal, *Journal of Sound and Vibration*, 46(3): 381-394, 1976.

[5] Heckl, M.A. and Abrahams, I.D., Curve squeal of train wheels, part 1: mathematical model for its generation, *Journal of Sound and Vibrations*, 229(3): 669-693, 2000.

[6] Heckl, M.A., Curve squeal of train wheels, part 1: which wheel modes are prone to squeal?, *Journal of Sound and Vibrations*, 229(3): 695-707, 2000.

[7] Hsu, S.S., Huang, Z., Iwnicki, S.D., Thompson, D.J., Jones, C.J.C., Xie, G. and Allen, P.D., Experimental and theoretical investigation of railway wheel squeal, Proceedings of the Institution of Mechanical Engineers. Part F, *Journal of rail and rapid transit*, 221: 59-73, 2007.

[8] Le Rouzic, J., Le Bot, A., Perret-Liaudet, J., Guibert, M., Rusanov, A., Douminge, L., Bretagnol, F. and Mazuyer, D., Friction-induced vibration by Stribeck's law: application to wiper blade squeal noise, *Tribology Letters*, 49: 563-572, 2013.

[9] De Beer, F.G., Janssens, M.H.A. and Kooijman, P.P., Squeal noise of rail-bound vehicles influenced by lateral contact position, *Journal of Sound and Vibration*, 267: 497-507, 2003.

[10] Hoffmann, N., Fischer, M., Allgaier, R. and Gaul, L., A minimal model for studying properties of the mode-coupling type instability in friction induced oscillations, *Mechanics Research Communications*, 29: 197-205, 2002.

[11] Ding, B., Squicciarini, G., Thompson, D.J. and Corradi, R., An assessment of mode-coupling and falling-friction mechanisms in railway curve squeal through a simplified approach, *Journal of Sound and Vibration*, 423: 126-140, 2018.

[12] Hoffmann, N. and Gaul, L., Effects of damping on mode-coupling instability in friction induced oscillations, *Journal of Applied Mathematics and Mechanics*, 83(8): 524-534, 2003.

[13] Sinou, J.J. and Jezequel, L., Mode coupling instability in friction-induced vibrations and its dependency on system parameters including damping, *European Journal of Mechanics-A/Solids*, 26(1): 106-122, 2007.

[14] Thompson, D.J., *Railway Noise and Vibration: Mechanisms, Modelling and Mitigation*. Elsevier, Oxford (2009).

19

Extension of the modal superposition method for general damping applied in railway dynamics

J. Giner-Navarro[♭1], V. Andrés-Ruiz[♭], J. Martínez-Casas[♭] and F. D. Denia[♭]

(♭) Centro de Investigación en Ingeniería Mecánica,
Universitat de València.

1 Introduction

The frequency response function (FRF) permits to characterise in the frequency domain the systems governed by linear dynamics by means of a relationship between an excitation applied at one degree of freedom (dof) and the consequent output response in a particular dof. A modal approach is widely extended in the engineering fields as efficient method of computing the FRF of matrix second-order linear equations of motion derived from the application of the Finite Element Method (FEM) [1]. This approach is based on the truncation of the number of vibration modes that conform the base of the new modal coordinates. The criterion for the truncation is linked to the frequency range of the dynamic study, ordering the vibration modes with respect to the eigenvalues (the square of natural frequencies) associated. The natural frequency associated with the last vibration mode selected establishes the maximum frequency that can describe the time response of the system. The truncation permits to reduce the dimension of the system from N number of dofs in physical coordinates to m truncated vibration modes in modal coordinates.

The fundamental numerical problem derived from the truncation is the resulting non-square vibration modes matrix, used as transformation matrix in the physical to modal change of variable [1, 2]. This change should allow the diagonalisation of the matrices involved in the equation of motion: mass, stiffness and damping matrices. The diagonalisation is essential to decouple the system in m second-order linear differential equations that can be solved analytically in the time domain. Nevertheless, the inverse of vibration modes matrix required for the diagonalisation cannot be applied from its non-square nature and it can only be replaced by the transpose matrix if both mass and stiffness matrices are symmetric. The complexity increases in a case of general damping instead of proportional or spectral ones [3], in which the damping matrix must be included in the eigenproblem in order to diagonalise the whole modal system.

This work proposes a methodology to overcome the issues abovementioned and applies this in the field of railway dynamics in order to compute the modal properties of a railway wheel

[1]e-mail: juanginer@upv.es

modelled by using FEM [4]. The paper includes a study of the numerical performance of this method and its comparison with other numerical procedures to find the FRF of the wheel.

2 Overview of the mathematical approach

The matrix equation of motion of a mechanical system can be formulated as

$$\mathbf{M}\ddot{\boldsymbol{u}} + \mathbf{C}\dot{\boldsymbol{u}} + \mathbf{K}\boldsymbol{u} = \mathbf{F}, \tag{1}$$

where $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{K}$ correspond with the mass, damping and stiffness matrices, respectively, and $\mathbf{F}$ contains the external force terms and $\boldsymbol{u}$ are the physical coordinates. In this work, it is considered a general case of damping, instead of a proportional or spectral definition.

2.1 KC method

The eigenvectors matrix is used as matrix transformation to move to modal coordinates. Only the symmetric part of the stiffness matrix, $\mathbf{K}_{sym}$, is taken in order to be able to use the transpose matrix (instead of the inverse one) to diagonalise the mass and stiffness matrices involved. N is the number of degrees of freedom of the system and m the truncation number selected.

$$eigs(\mathbf{K}_{sym}, \mathbf{M}, m) \rightarrow \mathbf{\Phi} = [\{\phi\}_1, \ldots, \{\phi\}_m]. \tag{2}$$

It is proposed a first variable transformation:

$$\boldsymbol{u} = \mathbf{\Phi}\boldsymbol{q}, \tag{3}$$

where $\boldsymbol{q}$ is the modal coordinates vector. Eq. (1) is reduced to dimension m:

$$\ddot{\boldsymbol{q}} + \widetilde{\mathbf{C}}\dot{\boldsymbol{q}} + \widetilde{\mathbf{K}}\boldsymbol{q} = \widetilde{\mathbf{F}}, \tag{4}$$

where the *eigs* function has normalised the mass matrix $\widetilde{\mathbf{M}} = \mathbf{\Phi}^T\mathbf{M}\mathbf{\Phi} = \mathbf{I}$ and $\widetilde{\mathbf{K}}_{sym} = \mathbf{\Phi}^T\mathbf{K}_{sym}\mathbf{\Phi} = [\omega_r^2]$ is a diagonal matrix that contains the square of the natural frequencies. Nevertheless, the stiffness matrix $\widetilde{\mathbf{K}} = \mathbf{\Phi}^T\mathbf{K}\mathbf{\Phi} = \mathbf{\Phi}^T(\mathbf{K}_{sym} + \mathbf{K}_{antisym})\mathbf{\Phi} = [\omega_r^2] + \widetilde{\mathbf{K}}_{antisym}$ and the damping one $\widetilde{\mathbf{C}} = \mathbf{\Phi}^T\mathbf{C}\mathbf{\Phi}$ are not diagonal. The generalised force is $\widetilde{\mathbf{F}} = \mathbf{\Phi}^T\mathbf{F}$, where $\widetilde{\mathbf{F}}_r = \sum_{k=1}^{N} \phi_{kr}F_k$.

Considering a harmonic excitation $\widetilde{\mathbf{F}} = \bar{\widetilde{\mathbf{F}}}e^{i\omega t}$, it is assumed a harmonic response $\boldsymbol{q} = \bar{\boldsymbol{q}}e^{i\omega t}$. Replacing in Eq. (4):

$$\bar{\boldsymbol{q}} = (-\omega^2\widetilde{\mathbf{I}} + i\omega\widetilde{\mathbf{C}} + \widetilde{\mathbf{K}})^{-1}\bar{\widetilde{\mathbf{F}}}. \tag{5}$$

Hence, the receptance can be defined through an inverse matrix:

$$H_{ij}(\omega) = \frac{\bar{\boldsymbol{q}}}{\bar{\widetilde{\mathbf{F}}}} = \mathbf{\Phi}_j(-\omega^2\widetilde{\mathbf{I}} + i\omega\widetilde{\mathbf{C}} + \widetilde{\mathbf{K}})^{-1}\mathbf{\Phi}_k^T. \tag{6}$$

2.2 AB-decoupling method

Considering the $2m$-extended system through

$$\mathbf{Q} = \begin{Bmatrix} \boldsymbol{q} \\ \dot{\boldsymbol{q}} \end{Bmatrix}, \tag{7}$$

the modal matrix equation results

$$\widetilde{\mathbf{A}}\dot{\mathbf{Q}} + \widetilde{\mathbf{B}}\mathbf{Q} = \begin{Bmatrix} \widetilde{\mathbf{F}} \\ \mathbf{0} \end{Bmatrix}, \tag{8}$$

where

$$\widetilde{\mathbf{A}} = \begin{pmatrix} \widetilde{\mathbf{C}} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}, \quad \widetilde{\mathbf{B}} = \begin{pmatrix} \widetilde{\mathbf{K}} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}. \tag{9}$$

At this stage, the eigenproblem associated with the linear first-order matrix equation of motion is solved without truncating, obtaining the $2m$-square matrix $\boldsymbol{\Theta}$,

$$eig(\widetilde{\mathbf{B}}, \widetilde{\mathbf{A}}) \rightarrow \boldsymbol{\Theta} = [\{\theta\}_1 \dots \{\theta\}_{2m}] \tag{10}$$

The eig function has normalised the first matrix $\widetilde{\mathbf{A}} = \boldsymbol{\Theta}^T\mathbf{A}\boldsymbol{\Theta} = \mathbf{I}$ and $\widetilde{\mathbf{B}} = \boldsymbol{\Theta}^T\mathbf{B}\boldsymbol{\Theta} = [\lambda_s]$ is diagonal, resulting a set of uncoupled first-order linear differential equations:

$$\dot{Q}_s + \lambda_s Q_s = \{\theta\}_s^T \begin{pmatrix} \widetilde{\mathbf{F}} \\ \mathbf{0} \end{pmatrix}. \tag{11}$$

The diagonalisation of the matrix equation of motion permits to compute the receptance for general damping using modal superposition. With just the first variable transformation, non-diagonal modal matrices were found and the application of the inverse in the resulting modal equation was needed to approach the calculation of the receptance. The inversion of a matrix of the dimension for common FE structures (hundreds of thousands of dofs) is not addressable for conventional PCs. The proposed method based on the second variable transformation from the extended $2m$-system drastically reduce the time consumption of the receptance computing through the following expression:

$$H_{jk}(\omega) = \sum_{s=1}^{m} \phi_{js} \sum_{r=1}^{2m} \theta_{sr} \frac{\sum_{l=1}^{2m} \left(\boldsymbol{\theta}^{-1}\right)_{rl} \sum_{t=1}^{m} \left(\tilde{\mathbf{A}}^{-1}\right)_{lt} \Phi_{kt}}{i\omega + \lambda_r} \tag{12}$$

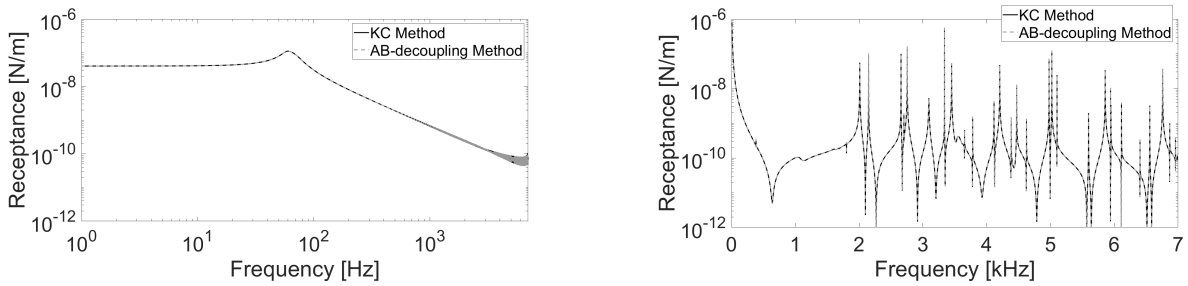
Being $\widetilde{\mathbf{A}}^{-1} = \begin{pmatrix} \widetilde{\mathbf{0}} & \widetilde{\mathbf{I}} \\ \widetilde{\mathbf{I}} & -\widetilde{\mathbf{C}} \end{pmatrix}$:

$$H_{jk}(\omega) = \boldsymbol{\Phi}_j \sum_{r=1}^{2m} \frac{\boldsymbol{\Theta}_{(1:m,r)} \left(\boldsymbol{\Theta}^{-1}\right)_{(r,m+1:2m)}}{i\omega + \lambda_r} \boldsymbol{\Phi}_k^{\mathrm{T}}. \tag{13}$$

The modal static correction [5] is implemented and applied to the previous expression in order to compensate the lack of contribution of the truncated vibration modes on the static response of the system.

3 Results

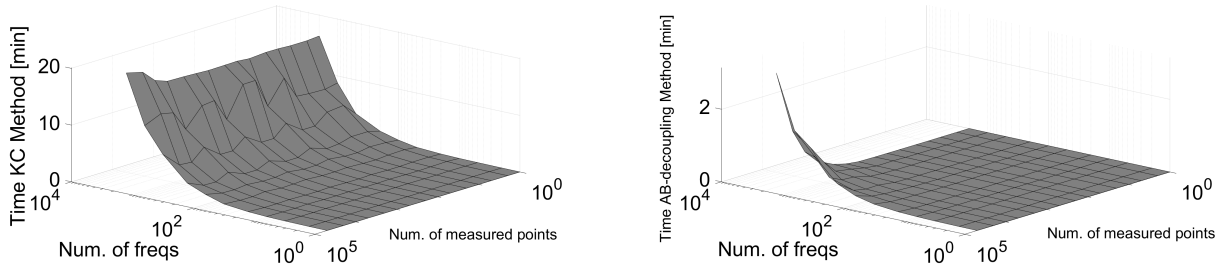
Using both KC and AB-decoupling methods sinthetised in Eqs. (6) and (13), respectively, the receptances for the track and the wheel have been evaluated. Fig. 1(a) shows that both methods give the same receptance for a rail modelled by the Moving Element Method [6] supported by a continuous viscoelastic Winkler bedding. A S1002 undamped Finite Element wheel model [1] has been also computed, obtaining again two overlapped curves for both methods.



(a) (b)

Figure 1: (a) Track receptance; (b) wheel receptance.

In terms of computational performance, the KC method requires the calculation of the $(-\omega^2\tilde{\mathbf{I}}+i\omega\tilde{\mathbf{C}}+\tilde{\mathbf{K}})^{-1}$ inverse, which is the most expensive operation. Hence, the time consumption exponentially grows with the number of frequencies selected to build the receptance. The pre- and post-multiplication of the modal transformation matrix $\mathbf{\Phi}$ barely increases the computational time, as reflected in Fig. 2(a), in which the influence of the number of physical measured points selected to calculate the receptance is almost negligible. Since there is not any inverse to compute for the AB-decoupling method, the influence of the number of frequencies and measured points can be clearly observed in Fig. 2(b).



(a) (b)

Figure 2: Computational time for the calculation of the wheel receptance.

The previous figures show that KC method requires higher computational time, especially when the receptance is evaluated for low number of measured points. The ratio between KC and AB-decoupling times plotted in Fig. 3 is in line with this observation since the first method needs to compute a very large matrix only for evaluating a few terms of the resulting matrix. When increasing the number of measured points, the ratio is reduced asymptotically but always above 1, showing that the AB-decoupling method is a more efficient one to compute the receptance.

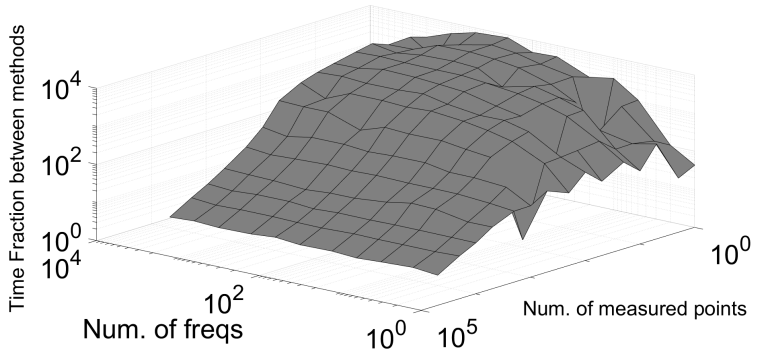


Figure 3: Ratio between the computational time for the KC and AB-decoupling methods for the calculation of the wheel receptance.

Acknowledgements

The authors gratefully acknowledge the financial support of FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación (project TRA2017-84701-R), as well as Conselleria d'Educació, Investigació, Cultura i Esport (project Prometeo/2016/007) and European Commission through the project "RUN2Rail - Innovative RUNning gear soluTiOns for new dependable, sustainable, intelligent and comfortable RAIL vehicles" (Horizon 2020 Shift2Rail JU call 2017, grant number 777564).

References

[1] Martínez-Casas, J., Fayos, J., Denia, F. D. and Baeza, L., Dynamics of damped rotating solids of revolution through an Eulerian modal approach, *Journal of Sound and Vibration*, 331: 868-882, 2012.

[2] Baeza, L. and Ouyang, H., A railway track dynamics model based on modal substructuring and cyclic boundary condition, *Journal of Sound and Vibration*, 330: 75-86, 2011.

[3] Baeza, L., Giner-Navarro, J., Thompson, D.J. and Monterde, J., Eulerian models of the rotating flexible wheelset for high frequency railway dynamics, *Journal of Sound and Vibration*, 449:300-314, 2019.

[4] Martínez-Casas, J., Mazzola, L., Baeza, L. and Bruni, S., Numerical estimation of stresses in railway axles using a train-track interaction model, *International Journal of Fatigue*, 47: 18-30, 2013.

[5] Giner-Navarro, J., Martínez-Casas, J., Denia, F.D. and Baeza, L., Study of railway curve squeal in the time domain using a high-frequency vehicle/track interaction model, *Journal of Sound and Vibration*, 431: 177-191, 2018.

[6] Martínez-Casas, J., Giner-Navarro, J., Baeza, L. and Denia, F.D., Improved railway wheelset–track interaction model in the high-frequency domain, *Journal of Computational and Applied Mathematics*, 309: 642-653, 2017.

20

Predicting healthcare cost of diabetes using machine learning models

Javier-Leonardo González-Rodríguez[♭1], Javier Díaz Carnicero[♮], David Vivas-Consuelo[♮], Silvia González de Julian[♮] and Olga Lucia Pinzon Espitia[♯]

(♭) Business Management School,
Universidad del Rosario, Bogotá,
(♮) INECO,
Universitat Politécnica de València,
(♯) Universidad Nacional de Colombia.

1 Introduction

Diabetes mellitus (DM) describes a group of metabolic disorders characterised by high blood glucose levels. People with diabetes have an increased risk of developing several serious life-threatening health problems resulting in higher medical care costs, reduced quality of life and increased mortality [1] DM, like the majority of non-contagious chronic diseases, is associated with multimorbidity, defined in the growing literature as the existence of two or more chronic conditions [2,3]. Multimorbidity causes a negative impact on both clinical and health indicators and primary health care costs [1,4]. While true that the analysis of multimorbidity in this type of population is relatively new, the tendency towards this approach to the study of chronic diseases is ever increasing [5, 6].

This co-occurrence of diseases has implications from a disease management point of view, as the features of comorbid diseases can be much more complicated than a simple aggregation of individual illnesses [7, 8]. Previous studies have related DM to a set of diseases such as cardio-vascular, renal, obesity and the metabolic syndrome.

Diabetes mellitus Type II (DM2) [9] is among the chronic diseases that generate the most health expenditure and clinical risk, due to the comorbidities that it frequently deals with. For this reason, it is very important to determine a total risk index calculated based on the variability determined by the number and severity of the associated morbidities.

Based on this risk index, a predictive model of pharmaceutical expenditure can be developed, applicable not only to DM2, but also to other chronic diseases.

[1]e-mail: jagonro1@upvnet.upv.es

2 Materials and methods

Objective
To design a predictive model of the pharmaceutical expenditure of DM2 patients, derived from the risk index determined by the associated comorbidities, in a health district of Valencian Community Spain.

- Cross-sectional descriptive and analytical study, and predictive models of total healthcare expenditure for application in clinical management.
    - **Population**: 28.345 DM2 patients in a public health district from Valencian Community.
    - **Sample**: 13.820 patients, equivalent to 40% who had complete data to asses, according to de defined variables.
    - **Variables**: Age, sex, Primary and secondary diagnosis, (Comorbidity and Multimorbidity), Clinical Risk Groups - CRG, Glycosylated haemoglobin – HbA1c, Average Glycemia, Creatinine, Microalbuminuria, Lipid profile, (total cholesterol, Triglycerides, Glomerular Filtering.

3 Modelling

Prediction of events and complexities related to DM2 based on clinical information using logistic regression models:

Main Risk:

- Acute Myocardial Infarction.
- Brain Vascular Stroke.

Complications:

- Chronic Kidney Disease - Kidney Failure - Transplant.
- Diabetic Retinopathy - Secondary Blindness.

Prediction of the pharmaceutical expenditure using the calculated risk of events and complexities, comparing between classical linear regression a machine learning models.

4 Results

Descriptive Analysis The descriptive analysis is presented in the following tables, highlighting the most relevant aspects, such as the distribution of patients by age and hospital stay, on the one hand, and on the other hand, the distribution of patients from the perspective of the most significant events or comorbidities related to diabetes, these are: retinopathy, chronic kidney disease - CKD, myocardial infarction and stroke BV (Table 3).

The variables used for the design of the predictive model respond to explicative aspects, both patients themselves and of the conditions of their illness and allow us to assume both the behaviour in the variation of the risk and its impact on the costs of care.

Thus, age and sex can be related to variations in days of hospital stay and therefore in cost. The existence of a primary diagnosis, in this case diabetes or some secondary diagnoses or comorbidities, affect the Clinical Risk Group CRG index, glycosylated haemoglobin - HbA1c, and average Glycemia, account for the state of diabetes (controlled or not controlled), Creatinine and Microalbuminuria allow the calculation of GFR, with which the risk classification KDIGO is obtained; and finally the lipidic profile (total cholesterol, Triglycerides, gives an idea of the state of cardiovascular risk, (total cholesterol, Triglycerides, gives an idea of the state of cardiovascular risk), (total cholesterol, Triglycerides, gives an idea of the state of cardiovascular risk), and finally the lipid profile, (total cholesterol, Triglycerides, gives an idea of the state of cardiovascular risk).

In the table 1, the predominance of male patients can be appreciated, with 52.7% of the cases as opposed to 47.3% of female sex. This higher frequency of male patients is also evident in the distribution by hospital stay, which is shown in (table 1, 2).

AGE / **G age**	**SEX**	**SIP_ RECOD** / **YS_RECOD**	
< 40 years	M	185	
	F	224	
	Total	**409**	**3,0%**
> 85 years	M	413	
	F	749	
	Total	**1162**	**8,4%**
$40 - 55$ years	M	978	
	F	555	
	Total	**1533**	**11,1%**
$55 - 70$ years	M	2997	
	F	2013	
	Total	**5010**	**36,3%**
$70 - 85$ years	M	2715	
	F	2991	
	Total	**5706**	**41,3%**
Total	**M**	**7288**	**52,7%**
	F	**6532**	**47,3%**
	Total	**13820**	

Table 1: Age distribution of patients.

It is striking that 86% of the cases correspond to patients over 55 years of age, so it is worth reflecting on whether age is a decisive factor in the presentation of greater association with comorbidities. As well as these patients over 55 years of age, they explain 80% of the 16831 days of hospital stay.

On the other hand, it was found that the distribution of the data has a non-parametric character, which is why, to evaluate the level of significance of the distributions, the binomial test was used, which yields a highly significant result, (table 4).

SEX	G. AGE	N	MEDIA	SUM	PERCENTAGE
M	< 40 years	185	0,23	43	0,3%
	> 85 years	413	2,09	864	5,1%
	$40 - 55$ years	978	0,62	606	3,6%
	$55 - 70$ years	2997	0,96	2887	17,1%
	$70 - 85$ years	2715	1,74	4711	27,9%
	Total	**7288**	**1,25**	**9111**	**54,0%**
F	< 40 years	224	1,06	237	1,4%
	> 85 years	749	1,82	1361	8,1%
	$40 - 55$ years	555	0,5	279	1,7%
	$55 - 70$ years	2013	0,69	1395	8,3%
	$70 - 85$ years	2991	1,50	4498	26,6%
	Total	**6532**	**1,19**	**7770**	**46,0%**
Total	< 40 years	409	0,68	280	
	> 85 years	1162	1,91	2225	
	$40 - 55$ years	1533	0,58	885	
	$55 - 70$ years	5010	0,85	4282	
	$70 - 85$ years	5706	1,61	9209	
	Total	**13820**	**1,22**	**16881**	

Table 2: Distribution of patients by stay.

SEX	N	RETINOPATHY	CDK	INFARCT	BVS	PIELONEFRITHYS
MALE	**7288 (54%)**	239	168	329	106	66
FEMALE	**6532 (46%)**	211	88	117	85	149
TOTAL	**16881 (100%)**	450	256	446	191	215
STAY	16881	918	1181	1132	4722	353

Table 3: Event summary morbidity.

Event Prediction. Logistic Regression Results. In order to the prediction of events and complexities related to DM2 (Infarction, Stroke, Retinopathy, Renal failure) we propose different logistic regression models, using available clinical information. A linear combination of the variables with their corresponding coefficients is transformed via the logistic function, presented below, in order to obtain the probability of an event occurring.

$$P(y = 1|x) = \frac{\exp(x)}{1 + \exp(x)}$$

For instance, the results obtained for the prediction of an infarction event occurring are shown hereafter. The results were obtained for the rest of the variables in a similar way, varying the correspondent coefficients in order to maximize the predictive value of the model.

$$x = -341,93 + 0,37\ State\,of\,health + 0,29\ Severity - 20,53\ Filtr - 97,25\ Album - 1579,92\ HbA1C - 2,95\ Cholestherol$$

The resulting ROC for our example curve can be seen in graph 2, and its corresponding area under the curve is 0.767, which determines a satisfactory predictive power of the model. Additionally, the calibrations of the model allow to have a high negative predictive value, over 75%, without trading off the overall results of the model. The results are similar for all the logistic regression prepared.

		Category	N	Observed Prop	Test Prop	Bilateral Exact Sign
SEX	Group1	1	3150	0,60	0,50	0,000
	Group 2	0	2083	0,40		
	Total		5233	1,00		

Table 4: Event summary morbidity.

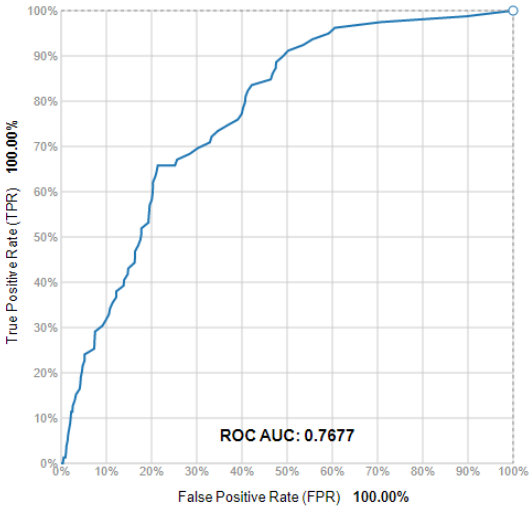


Figure 1: Predictive Power of the Regression Model.

Pharmaceutical expenditure prediction models. Now we try to create the model for the prediction of the pharmaceutical expenditure using the calculated risk of events and complexities. Firstly, we calculated the linear regression, and the equation proposed is as follows:

$$\begin{aligned} Expenditure = & -1,50 \times 10^5 - 172,41\ Age - 6537,10\ Gender + 41001\ State\,of\,Health \\ & + 2468,60\ Severity + 46377\ Retinopathy + 13946,3\ Renal\,failure \\ & + 34143\ Infarction + 22452,9\ Stroke \end{aligned}$$

This R^2 value obtained was $0,32$, it is not significant enough for a practical use, but values close to 50% would be expected according to the studies published.

Secondly, we prepare for the prediction model a machine learning approach. For this purpose, we prepare a neural network with 1 hidden layer and the ADAM algorithm for training. In order to compare the results with the classical linear regression, we selected the same variables. In this case the R^2 resultant is $0,35$. This is slightly higher than that obtained by linear regression, but there is no noticeable difference in practice.

5 Conclusions

The risk management model is based on the study of expenditure on health services caused by DM and its comorbidities, which have a significant impact on the health services budget, with pharmaceutical expenditure being the most relevant.

It is shown how expenditure increases significantly as the number of associated diseases increases, so it can be deduced that the financial risk index is definitively associated with the comorbid-based risk class. These elements provide some basis for the design of the prescriptive spending model. The risk prediction models are particularly valid for their negative predictive value.

While Machine learning models lightly improve the result, their computational cost is significantly higher, so the linear regression is globally a more effective alternative.

It would be necessary to collect more variables in order to improve the predictive outcome of our model.

References

[1] Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H. and Ghali, WA., New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *J Clin Epidemiol*, 57(12): 1288–94, 2004.

[2] Fox, R. and Fletcher, J., Alarm symptoms in primary care. *Br Med J.*, 334(7602): 1013–4, 2007.

[3] Valderas, JM., Starfield, B., Sibbald, B., Salisbuty, C. and Roland, M., Understanding Health and Health Services. *Ann Fam Med.*, 7(4): 357–63, 2009.

[4] Glynn, LG., Valderas, JM., Healy, P., Burke, E., Newell, J., Gillespie, P., et al. The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Fam Pract.*, 28(5): 516–23, 2011.

[5] Barnett, K., Mercer, SW., Norbury, M., Watt, G., Wyke, S. and Guthrie, B., Epidemiology of multimorbidity and implications for health care, research, and medical education: A cross-sectional study. *Lancet* [Internet], 380(9836):37–43, 2012. Available from: `http://dx.doi.org/10.1016/S0140-6736(12)60240-2`.

[6] Holden, L., Scuffham, PA., Hilton, MF., Muspratt, A., Ng, S. and Whiteford, HA., Patterns of multimorbidity in working Australians. 1–5, 2011.

[7] Stavem, K., Hoel, H., Skjaker, SA. and Haagensen, R., Charlson comorbidity index derived from chart review or administrative data: Agreement and prediction of mortality in intensive care patients. *Clin Epidemiol* [Internet], 9:311–20, 2017. Available from: `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020468676&doi=10.2147%2FCLEP.S133624&partnerID=40&md5=9b93e44559138d782f578b59f99cafc4`.

[8] Fritzen, K., Heinemann, L. and Schnell, O,. Modeling of Diabetes and Its Clinical Impact. *J Diabetes Sci Technol*, 12(5):976–84, 2018.

[9] Caballer-Tarazona, V., Guadalajara-Olmeda, N. and Vivas-Consuelo, D., Predicting healthcare expenditure by multimorbidity groups. *Health Policy* (New York) [Internet], 123(4):427–34, 2019. Available from: `https://doi.org/10.1016/j.healthpol.2019.02.002`

21

Sampling of pairwise comparisons in decision-making

J. Benítez[♭1], S. Carpitella[♮], A. Certa[♯] and J. Izquierdo[♮]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València,
(♮) FluIng - Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València,
(♯) Università degli Studi di Palermo.

1 Introduction

Various decision-making techniques rely on pairwise comparisons (PCs) between the involved elements. Traditionally, PCs are provided by experts or relevant actors, and compiled into pairwise comparison matrices (PCMs).

In highly complex problems, the number of elements to be compared may be very large. One of the issues limiting PC applicability to large-scale decision problems is the so-called curse of dimensionality, that is, many PCs need to be elicited from an actor, or built from a body of information.

In general, when applied to a set of n elements to be compared, the number of PCs that have to be made is $n(n-1)/2$. When the information in the comparison matrix is complete, the priorities can be obtained. This is the case of decision-making with complete information. However, if there are missing entries due to uncertainty or lack of information, decision-making must be performed from the available incomplete information. The authors have addressed the issue of incomplete information in [5, 6], and have characterized the consistent completion of a PCM using graph theory in [6].

In this contribution, we claim that less than that number of comparisons may be suitable to develop sound decision-making. There is a trivial solution providing a lower bound for the sample size: just produce $n-1$ PCs, for example comparing one element with the others. It can be shown that this is equivalent to give directly the priority vector. Here we reduce the number of pairwise comparisons in a decision-making problem by selecting just a sample of n PCs that are able to provide balanced and unbiased (incomplete) information that still produces consistent and robust decisions. Both the size of the sample and its distribution within the PCM are of interest.

[1]e-mail: jbenitez@mat.upv.es

We address this research within a linearization theory developed by the authors [2] based on optimizing the consistency of reciprocal matrices.

2 Problem statement and solution

If there are n alternatives, the expert must build an $n \times n$ reciprocal matrix, and therefore, produce $n(n-1)/2$ PCs. If n is large, $n(n-1)/2$ is also large and the expert can be easily tired and lose the necessary concentration. For example, if $n = 10$ (which is not very large), then $n(n-1)/2 = 45$, and a survey of 45 questions may be tedious, strenous and time-consuming. In contrast, if the expert is asked to fill fewer entries, the survey will become more friendly and, arguably, more reliable.

2.1 Problem

Here we focus on the incomplete $n \times n$ reciprocal matrix B, where only entries $b_{12}, b_{23}, \ldots, b_{n-1,n}$, b_{n1} are known. For example, for size 6×6,

$$B = \begin{bmatrix} 1 & b_{12} & \star & \star & \star & b_{n1}^{-1} \\ b_{12}^{-1} & 1 & b_{23} & \star & \star & \star \\ \star & b_{23}^{-1} & 1 & b_{34} & \star & \star \\ \star & \star & b_{34}^{-1} & 1 & b_{45} & \star \\ \star & \star & \star & b_{45}^{-1} & 1 & b_{56} \\ b_{n1} & \star & \star & \star & b_{56}^{-1} & 1 \end{bmatrix}. \tag{1}$$

The next result characterizes when B can be completed to be consistent.

Theorem 1 *Let $B \in M_n$ be a reciprocal incomplete matrix with known entries b_{12}, b_{23}, ..., $b_{n-1,n}$, b_{n1}.*

(i) Matrix B admits a consistent completion if and only if

$$b_{12}\ b_{23} \cdots b_{n-1,n}\ b_{n1} = 1. \tag{2}$$

(ii) If B admits a consistent completion, then it is unique, say C, and C satisfies the following condition: if $(B)_{ij}$ is unspecified and $i < j$, then $(C)_{ij} = b_{i,i+1} b_{i+1,i+2} \cdots b_{j-1,j}$.

Let's check the performance of this approach. Let A be a (fully known) reciprocal matrix. We can find X_A, the closest consistent matrix to A by using the formula given in [4]. Also, from the incomplete matrix B defined as in the statement of Theorem 1, supposing that B satisfies the criterion given in this theorem, we can easily compute C. The next example compares matrices X_A and C and calculates the distance between both.

Example 1 *Let*

$$A = \begin{bmatrix} 1 & 2 & 2 & 8 \\ 1/2 & 1 & 4 & 1/2 \\ 1/2 & 1/4 & 1 & 1 \\ 1/8 & 2 & 1 & 1 \end{bmatrix}. \tag{3}$$

This matrix A is not consistent (e.g., rank$(A) > 1$, see [3, Theorem 1]). The Perron eigenvalue is $\lambda_{\max} \simeq 4.84$, and (see [7]) $\mathrm{CI}(A) = (\lambda_{\max} - 4)/(4-1) \simeq 0.279$ and $\mathrm{CI}(A)/\mathrm{RI}_4 \simeq 0.314 > 0.1$; the consistency of A is not acceptable (Saaty's criterion); here $\mathrm{RI}_4 = 0.89$ is the random index for 4×4 matrices.

Using the formula given in [4], we have

$$X_A \simeq \begin{bmatrix} 1 & 2.38 & 4 & 3.36 \\ 0.42 & 1 & 1.68 & 1.41 \\ 0.25 & 0.59 & 1 & 0.84 \\ 0.29 & 0.71 & 1.19 & 1 \end{bmatrix}.$$

To apply Theorem 1, let us consider the following incomplete reciprocal matrix

$$B = \begin{bmatrix} 1 & 2 & \star & 8 \\ 1/2 & 1 & 4 & \star \\ \star & 1/4 & 1 & 1 \\ 1/8 & \star & 1 & 1 \end{bmatrix}.$$

Since (2) holds, then there exists a unique consistent completion, namely

$$C = \begin{bmatrix} 1 & 2 & b_{12}b_{23} & 8 \\ 1/2 & 1 & 4 & b_{23}b_{34} \\ (b_{12}b_{23})^{-1} & 1/4 & 1 & 1 \\ 1/8 & (b_{23}a_{34})^{-1} & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 8 & 8 \\ 1/2 & 1 & 4 & 4 \\ 1/8 & 1/4 & 1 & 1 \\ 1/8 & 1/4 & 1 & 1 \end{bmatrix}.$$

Now the distance between X_A and C is $d(X_A - C) = \|X_A - C\|_F = tr(X_A C^T) \simeq 2.4992$, $\mathrm{tr}(\cdot)$ being the trace operator.

If (2) does not hold, the matrix B defined in Theorem 1 has no consistent completion. If we denote by $\mathcal{C}_n$ the set of $n \times n$ consistent matrices, then we must find $D \in M_n$, a reciprocal completion of B, such that

$$\mathrm{d}(D, \mathcal{C}_n) \leq \mathrm{d}(D', \mathcal{C}_n)$$

for any $D' \in M_n$ reciprocal completion of B.

We summarize the obtained results in the following theorem.

Theorem 2 *Let $B \in M_n$ be a reciprocal incomplete matrix with known entries b_{12}, b_{23}, ..., $b_{n-1,n}$, b_{n1}.*

(i) There is a unique reciprocal completion of B, say D, such that $\mathrm{d}(D, \mathcal{C}_n) \leq \mathrm{d}(D', \mathcal{C}_n)$ for all $D' \in M_n$ reciprocal completion of B.

(ii) There is a unique $Z \in \mathcal{C}_n$ such that $\mathrm{d}(D, Z) = \mathrm{d}(D, \mathcal{C}_n)$.

(iii) $Z = E[\phi_n(\mathcal{L}^\dagger \mathcal{Q} \boldsymbol{\rho})]$, where $\boldsymbol{\rho} = (\log b_{12}, \ldots, \log b_{n-1,n}, \log b_{n1})^T$, and matrices $\mathcal{Q}, \mathcal{L}$ are the Laplacian matrix and the incidence matrix, respectively, of the graph associated to B.

(iv) If (i, j) is an unknown entry of B, then the (i, j) entry of D and Z are equal.

Note that the oriented graph with n vertices and edges associated to B is

$$\{1 \to 2, 2 \to 3, \ldots, n-1 \to n, n \to 1\}.$$

Let's define the following matrix J directly associated to the structure of B

$$J = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{4}$$

A number of considerations enable us to prove the following result.

Theorem 3 *Let $B \in M_n$ be a reciprocal incomplete matrix with known entries b_{12}, b_{23}, ..., $b_{n-1,n}$, b_{n1}. Under the notation of Theorem 2, one has*

$$Z = E\left[\phi_n\left(\frac{1}{2n}\sum_{k=0}^{n-1}(n-2k-1)J^k\boldsymbol{\rho}\right)\right],$$

where the matrix J is given in (4) and ϕ_n is the linear mapping $\phi_n : \mathbb{R}^n \to M_n$ given by $(\phi_n(\mathbf{v}))_{ij} = v_i - v_j$.

This expression shows that neither inverses nor pseudo-inverses have to be computed. Also, $\boldsymbol{\rho}, J\boldsymbol{\rho}, J^2\boldsymbol{\rho}, \ldots, J^{n-1}\boldsymbol{\rho}$ are trivial to compute. For example, for $n = 4$, one has $J^0 = J^4 = I_4$,

$$J^1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad J^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad J^3 = J^{-1} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

If $\boldsymbol{\rho} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$ and $\widehat{\boldsymbol{\rho}} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$, then $J^0\boldsymbol{\rho}$ are the $1, 2, 3, 4$ entries of $\widehat{\boldsymbol{\rho}}$; $J^1\boldsymbol{\rho}$ are the $2, 3, 4, 5$ entries of $\widehat{\boldsymbol{\rho}}$; $J^2\boldsymbol{\rho}$ are the $3, 4, 5, 6$ entries of $\widehat{\boldsymbol{\rho}}$; and $J^3\boldsymbol{\rho}$ are the $4, 5, 6, 7$ entries of $\widehat{\boldsymbol{\rho}}$.

3 Conclusions

Making too many comparisons may be strenuous and time-consuming, and lead to wrong and harmful conclusions. It is indispensable to focus on its reduction [1]. There is not a general solution to the problem of finding an optimal sample of PCs to be issued so that card(sample) $< n(n-1) = 2$ and still produce sound DM. We have given a solution in which one compares just the elements of a balanced and unbiased subset of items. The solution, according to Theorem 3, is obtained through elementary, simple calculations.

References

[1] Abel, E., Mikhailov, L. and Keane, J., Inconsistency reduction in decision making via multi-objective optimisation. *EJOR*, 267(1): 212-226, 2018.

[2] Benítez, J., Delgado-Galván, X., Izquierdo, J. and Pérez-García, R., Achieving Matrix Consistency in AHP through Linearization. *Applied Mathematical Modelling*, 35: 4449-4457, 2011.

[3] Benítez, J., Delgado-Galván, X., Izquierdo, J. and Pérez-García, R., Improving consistency in AHP decision-making processes. *Applied Mathematics and Computation*, 219: 2432-2441, 2012.

[4] Benítez, J., Izquierdo, J., Pérez-García, R. and Ramos-Martínez, E., A simple formula to find the closest consistent matrix to a reciprocal matrix. *Applied Mathematical Modelling*, 38: 3968-3974, 2014.

[5] Benítez, J., Delgado-Galván, X., Izquierdo, J. and Pérez-García, R. Consistent completion of incomplete judgments in decision making using AHP. *Journal of Computational and Applied Mathematics*, 290: 412-422, 2015.

[6] Benítez, J., Carpitella, S., Certa, A. and Izquierdo, J., Characterization of the consistent completion of analytic hierarchy process comparison matrices using graph theory. *J. of Multi-Criteria Decision Analysis*, 26: 3-15, 2019.

[7] Saaty, T., Relative measurement and its generalization in decision making: Why pairwise comparisons are central in mathematics for the measurement of intangible factors-The Analytic Hierarchy/Network Process. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas*, 102: 251-318, 2008.

22

A multi-objective and multi-criteria approach for district metered area design: water operation and quality analysis

Bruno Brentan[♭], Silvia Carpitella[♮], Joaquín Izquierdo[♮1], Edevar Luvizotto Jr[♯] and Gustavo Meirelles[♭]

(♭) Engineering School,
Federal University of Minas Gerais,
(♮) FluIng-IMM,
Universitat Politècnica de València,
(♮) Laboratory of Computational Hydraulics,
University of Campinas.

1 Introduction

Water distribution network (WDN) operation may be improved by district metered area (DMA) design [1]. A first step to create DMAs uses graph theory and non-supervised learning, where physical features of the WDN, such as node coordinates, elevation and demand, are used for clustering purposes [2]. A second step is related to the necessary isolation of the clustered elements. For isolation purposes, it is important to determine the DMA entrances and, consequently, the needed cut-off valves. Closure of pipes and definition of DMA entrances can be set as an optimization problem with the costs associated to the valves, which are linked to pipe diameters, as a primary objective. However, placement and operation of pressure reducing valves (PRVs) change the hydraulic conditions, and the optimization process should respect operation limits, such as minimum and maximum pressure and minimum and maximum tank levels. The optimization process can be written as:

$$\sum_{i=1}^{N_v} c(D_i) \quad \text{s.t. } P_{\min} \leq P_{t,j} \leq P_{\max} \text{ and } T_{\min,k} \leq T_{t,k} \leq T_{\max,k}, \tag{1}$$

where $c(D_i)$ is the cost of a valve placed in a pipe with diameter D_i; N_v is the number of valves; $P_{\min}$ and $P_{\max}$ are the limit pressures allowed; $P_{t,j}$ is the operational pressure at time step t in pipe j; $T_{\min,k}$ and $T_{\max,k}$ are the minimum and maximum tank levels allowed for tank k; and $T_{t,k}$ is the operational level of tank k at time step t.

Constrained problems are frequently handled by using penalty functions. However, as discussed in [3], penalty approaches modify the search space, impairing the search process by the creation

[1]e-mail: jizquier@upv.es

of new local minima. To solve this problem, a bio-inspired algorithm widely applied in water distribution problems [4], adapted for a multi-objective approach, is applied. In this context, constraints become objectives to be reached, which turns the problem unconstrained.

Moreover, as observed in [5], such crucial water distribution parameters as resilience, pressure uniformity and water quality strongly depend on DMA configurations. These parameters are known to depend on pressures and water tank levels and, together with cost, will be the other objectives of our optimization.

A multi-objective approach gives a set of solutions, the so-called Pareto front. To select, within that front, which non-dominated solution will be implemented may be hard task. To help this process, this work presents: a) multi-level optimization process for entrance location and set point definition of PRVs, and b) a post-processing based on a multi-criteria method, which ranks the non-dominated solutions based on the relative importance of the said four main objectives: implementation cost, resilience, pressure uniformity and water quality.

Among the wide range of MCDM (multi-criteria decision-making) methods used in the literature, the *Technique for Order of Preference by Similarity to Ideal Solution* (TOPSIS) effectively works across various application areas [6]. Such a technique was developed by Hwang and Yoon [7] as a simple way to solve decision-making problems by means of the ranking of various decision alternatives [8, 9]. In this context, the objective of the TOPSIS application to the multi-objective problem consists in selecting the solution representing the best trade-off (among the set of optimal solutions belonging to the Pareto front) under the perspective of the considered evaluation criteria.

2 Clustering process based on a k-means algorithm

The first step for DMA design is to cluster the nodes of the network. Among the various methods suitable for this step, a simple and effective one is the k-means algorithm. The method uses the Euclidean distance between samples and centroids, and clusters are defined according to the smallest distances. For a simple explanation, let's take a set with m data points $\chi = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$ where each point $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,n}]$. Taking a pre-defined number of clusters k, the method starts distributing randomly the k centres in the data space. The Euclidean distance $d_{i,j}$ between each center j and each data point $\boldsymbol{x}_i$ is computed. The data points are classified as belonging to cluster j if the distance $d_{i,j}$ is minimum when compared for all other centres. After the classification step, the centres are replaced to the mean value of all points belonging to a cluster. The process is repeated (distance calculation, point classification, and centre replacement) until the distance between the centres at iteration $t-1$ and t is smaller than a tolerance value.

3 The non-dominated sorting genetic algorithm (NSGA-II)

Different from single objective optimization algorithms, multi-objective approaches do not find just one optimal solution, but a set of compromise solutions, so-called Pareto front. Among several algorithms proposed for multi-objective optimization, population based algorithms, such

as NSGA-II [10], are widely applied for engineering problems, highlighting their applications in the water distribution domain [11, 12].

NSGA-II evaluates all the N possible solutions composing a population. The solutions are evaluated for all the objectives finding non-dominated solutions. Considering two solutions $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, it is said that $\boldsymbol{x}_a$ dominates $\boldsymbol{x}_b$ if and only if both conditions a) and b) below are satisfied.

a) $\boldsymbol{x}_a$ is no worse than $\boldsymbol{x}_b$ for all objectives, and

b) $\boldsymbol{x}_a$ is strictly better than $\boldsymbol{x}_b$ at least for one objective.

For each single solution $\boldsymbol{x}_p$ it is possible to know the number n_p of solutions dominating $\boldsymbol{x}_p$ and the set of solutions S_p dominated by $\boldsymbol{x}_p$. By definition, non-dominated solutions have $n_p = 0$, and integrate the so-called primary Pareto front. Now, for all solutions q in the sets S_p included in that primary Pareto front, the value of n_q is reduced by one and those solutions q with new $n_q = 0$ for all p are collected into a new set, so-called secondary Pareto front. This procedure is repeated until finding new high-level fronts.

The algorithm starts with a random population P_0 of size N. Individuals are evaluated for all the objectives and the non-dominated front is found as described next. Genetic operations (binary tournament selection, recombination, mutation) are used to create a new set of solutions. The new population, now sizing $2N$, is sorted according to non-domination, and the Pareto fronts for all the levels are found; if the number of solutions belonging to the primary Pareto front is smaller than N, all solutions are preserved and the new population is completed with the higher-level Pareto fronts, according to the ranking. Otherwise, the N first solutions of the primary Pareto front are selected. One important feature of population-based algorithms is the maintenance of the solution spread. This is fundamental for good convergence to a Pareto-optimal set [10]. To this purpose, NSGA-II uses a crowded-comparison approach, based on crowding distances (see [10]).

The process re-starts with the new population by re-evaluating each solution under all objectives and re-ranking solutions based on the non-domination criterion. The algorithm stops when reaching some termination criteria, such as maximum number of iterations, or no improvements in the Pareto front. The method results in a set of non-dominated solutions with an optimal compromise relation for all the objectives. However, for practical problems, the evaluation of the Pareto front by experts could be hard task. To help, a post-processing step, based on MCDM is proposed.

4 The TOPSIS to rank solutions

As mentioned, TOPSIS is a MCDM method aimed at ranking various alternatives, such as the solutions of the decision-making problem under analysis. The method calculates distances from each solution to a positive ideal solution and to a negative ideal solution. The solution representing the best trade-off under the considered criteria is the one characterised by the shortest distance to the positive ideal solution, and the farthest to the negative one.

First of all, the TOPSIS technique needs the preliminary collection of the following input data to be applied: a decision matrix (collecting the evaluations g_{ij} of each alternative i under each criterion j), the weights of criteria (representing their mutual importance), and their preference directions (to establish if criteria have to be minimised or maximised).

The implementation of the procedure is led by following five main steps:

- Building the weighted normalized decision matrix, for which the generic element u_{ij} is calculated as:

$$u_{ij} = w_j \cdot z_{ij}, \ \forall i, \forall j; \tag{2}$$

  where w_j is the weight of criterion j and z_{ij} is the score of the generic solution i under the criterion j, normalized by means of the equation:

$$z_{ij} = \frac{g_{ij}}{\sqrt{\sum_{i=1}^{n} g_{ij}^2}}, \ \forall i, \forall j. \tag{3}$$

- Identifying the positive ideal solution A^* and the negative ideal solution A^-, calculated through the following equations:

$$A^* = (u_1^*, \ldots, u_k^*) = \{(u_{ij}|j \in I'), (u_{ij}|j \in I'')\}; \tag{4}$$

$$A^- = (u_1^-, \ldots, u_k^-) = \{(u_{ij}|j \in I'), (u_{ij}|j \in I'')\}; \tag{5}$$

  I' and I'' being the sets of criteria to be, respectively, maximized and minimized.

- Computing the distance from each alternative i to the positive ideal solution A^* and to the negative ideal solution A^- as follows:

$$S_i^* = \sqrt{\sum_{j=1}^{k} (u_{ij} - u_j^*)^2}, \ i = 1, \ldots, n; \tag{6}$$

$$S_i^- = \sqrt{\sum_{j=1}^{k} (u_{ij} - u_j^-)^2}, \ i = 1, \ldots, n. \tag{7}$$

- Calculating, for each alternative i, the closeness coefficient C_i^* which represents how the solution i performs with respect to the ideal positive and negative solutions:

$$C_i^* = \frac{S_i^-}{S_i^- + S_i^*}, \ 0 \leq C_i^* \leq 1, \ \forall i. \tag{8}$$

- Obtaining the final ranking of alternatives on the basis of the closeness coefficients calculated above. In particular, with relation to two generic solutions i and z, solution i must be preferred to solution z when $C_i^* \geq C_z^*$.

5 Case study

The methodology proposed is applied to the literature water network called Exnet [13]. The system supplies 400,000 consumers approximately, requiring a delivery minimum pressure of 20m. The network is composed by 1,891 nodes, and 2,465 pipes. Two reservoirs and five injection nodes (wells) feed the network. For clustering analysis, each node is used as a data point, endowed with its topological features, namely geographical position, elevation and base demand. The number of clusters is defined using the Davies-Bouldin (BD) criterion [14], which evaluates the final clustered data, considering the distances among data points in a cluster and the corresponding centre (intra-criterion), and the distances among centres (inter-criterion). The best cluster number minimizes the intra-criterion and maximizes the inter-criterion. Varying from two to 15 clusters, the best DB criterion is found to be nine clusters. Fig. 1 shows the clustered network.

Once clustered, the network should pass by the optimization step in order to define the entrances and, consequently, those pipes where PRVs will be installed. The application of NSGA-II at this step results in a Pareto front with 115 non-dominated solutions, as shown in Fig. 2.

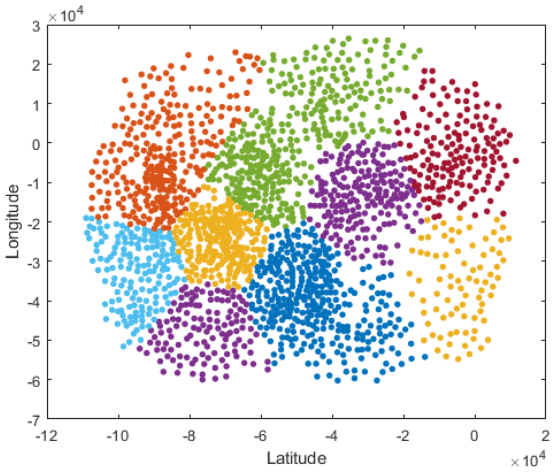


Figure 1: Clustered Exnet with optimal BD criterion, resulting in 9 clusters.

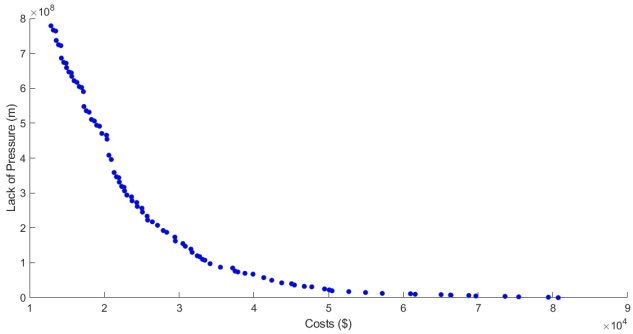


Figure 2: Pareto's front showing bi-objective problem of PRV placement in clustered networks.

The TOPSIS method described in the previous section has been applied to rank the 115 solutions belonging to the Pareto front. Obviously, resiliency is maximized whereas the other four criteria (pressure uniformity, dissipated energy, lack of pressure, and cost) are minimised. Moreover, at this stage of analysis, all the criteria have been considered as having the same importance. It means a weight equal to 20% has been assigned to each criterion. Results of TOPSIS application are reported in Table 1. Just the first five positions are given for sake of brevity.

Ranking position	**Pareto Solution**	**Resilience**	**Pressure Uniformity**	**Energy dissipated**	**Lack of Pressure**	**Costs**	**Closeness Coefficient value**
1	5	2,12E-01	8,30E+01	6,16E+03	7,54E+04	2,20E+06	0,758828
2	82	2,12E-01	8,30E+01	6,16E+03	7,54E+04	2,20E+06	0,758828
3	43	2,21E-01	8,38E+01	5,98E+03	7,94E+04	1,20E+06	0,754085
4	90	2,21E-01	8,38E+01	5,98E+03	7,94E+04	1,20E+06	0,754085
5	37	2,22E-01	8,38E+01	5,96E+03	8,07E+04	0,00E+00	0,750996

Table 1: TOPSIS results.

The solutions in the first positions present higher values of closeness coefficient, since they have large distance to the negative ideal solution and small distance to the positive ideal solution. Similar solutions appear in Table 1, such as 5 and 82, or 43 and 90. This happens by the closeness of those solutions in the Pareto front, with identical rounded values. Solution 37 exhibits an implementation cost equal to zero. This is a solution for the multi-objective problem from the mathematical point of view. From the engineering point of view, this means that all the boundary pipes remain open, thus resulting in a non-segregated network. Despite solution 37 is the last of the five top solutions, it still can help decision makers to find the benefits of DMA creation on that network.

To provide readers with an effective comparison of the results in terms of the values of the considered parameters, we also provide in Table 2 the last five positions of the ranking, those with the lowest closeness coefficient. It is possible to note as these last positions present higher associated costs (an objective to be minimised) and lower values of operation parameters (objectives to be maximise, instead).

Ranking position	**Pareto Solution**	**Resilience**	**Pressure Uniformity**	**Energy dissipated**	**Lack of Pressure**	**Costs**	**Closeness Coefficient value**
111	11	0,00E+00	1,92E+06	1,71E+04	1,42E+04	7,22E+08	0,249465
112	10	0,00E+00	1,92E+06	1,82E+04	1,29E+04	7,79E+08	0,249008
113	113	0,00E+00	1,92E+06	1,82E+04	1,29E+04	7,79E+08	0,249008
114	93	0,00E+00	1,92E+06	1,80E+04	1,32E+04	7,67E+08	0,248811
115	13	0,00E+00	1,92E+06	1,80E+04	1,35E+04	7,64E+08	0,248037

Table 2: TOPSIS results: last five positions of the ranking.

This ranking approach shows the interest of MCDMs to select trade-off scenarios under the considered criteria. The first solution shows the best pressure uniformity and lack of pressure, but the highest cost and lowest resilience. That means, the best hydraulic and operation

conditions will appear in the most expensive scenario. The relation of resilience and pressure uniformity can also be highlighted. Scenarios with lower pressure uniformity present lower resilience, since resilience is calculated based on overpressure, and pressure uniformity tries to minimize overpressure.

6 Conclusions and future developments

The present work proposes a fully automated algorithm for DMA design based on clustering analysis, multi-objective optimization and multi-criteria analysis. The clustering analysis is done by a k-means algorithm evaluated under the Davies-Bouldin criterion, resulting in nine DMAs. Multi-level optimization for entrance location and set point definition of pressure reducing valves achieve network clustering. NSGA-II finds 115 non-dominated solutions in a trade-off between various objectives. In addition, a MCDM is applied to rank the non-dominated solutions, to identify the one representing the best trade-off in fulfilling the objectives to be matched. Operational and hydraulic criteria are used to evaluate the solutions.

Regarding MCDM, the TOPSIS method has been applied to obtain the final ranking of non-dominated solutions. In particular, this application has been carried out under the evaluation of four criteria: implementation cost, resilience, pressure uniformity and water quality. In the presented case study, we assumed these criteria as having the same weight, in other terms, the same degree of mutual importance.

Results point to solution number 5 as the best trade-off among all the 115 non-dominated solutions, since it is the first in the ranking. While this solution embodies the best operational criteria, is the most expensive and least resilient.

Future developments of the present work may regard a further integration between the multi-objective and the multi-criteria perspectives, though with a different purpose: for example, instead of getting just a rank of the non-dominated solutions, the application of a MCDM method to classify them into proper clusters will be worth it.

References

[1] Campbell, E., Izquierdo, J., Montalvo, I., Ilaya-Ayza, A., Pérez-García, R. and Tavera, M., A flexible methodology to sectorize water supply networks based on social network theory concepts and multi-objective optimization. *Journal of Hydroinformatics*, 18(1): 62-76, 2016.

[2] Brentan, B., Campbell, E., Goulart, T., Manzi, D., Meirelles, G., Herrera, M., ... and Luvizotto Jr, E., Social network community detection and hybrid optimization for dividing water supply into district metered areas. *Journal of Water Resources Planning and Management*, 144(5), 04018020, 2018.

[3] Lima, G. M., Luvizotto Jr, E., Brentan, B. M. and Ramos, H. M., Leakage control and energy recovery using variable speed pumps as turbines. *Journal of Water Resources Planning and Management*, 144(1), 04017077, 2017.

[4] Montalvo, I., Izquierdo, J., Schwarze, S. and Pérez-García, R., Multi-objective particle swarm optimization applied to water distribution systems design: an approach with human interaction. *Mathematical and Computer Modelling*, 52(7-8): 1219-1227, 2010.

[5] Brentan, B. M., Campbell, E., Meirelles, G. L., Luvizotto, E. and Izquierdo, J., Social network community detection for DMA creation: criteria analysis through multilevel optimization. *Mathematical Problems in Engineering*, 2017.

[6] Behzadian, M., Otaghsara, S.K. Yazdani, M. and Ignatius, J., A state-of the-art survey of TOPSIS applications. *Expert Systems with Applications*, 39(17): 13051-13069, 2012.

[7] Huang, C.-L. and Yung, K., *Multiple-Attribute Decision Making. Methods and Applications. A state-of-the-Art Survey.* Springer-Verlag, Berlin, Heidelberg, New York, 1981.

[8] Carpitella, S., Certa, A., Enea, M., Galante, G., Izquierdo, J., La Fata C.M. and Vella, F., Combined HACCP and TOPSIS-based approach to prioritize risks in the salmon manufacturing process: an applicative case. Proceedings of the 22th Summer School "Francesco Turco", Palermo, Italy, September 13-15, 150-156, 2017.

[9] Carpitella, S., Certa, A., Izquierdo, J. and La Fata, C.M., k-out-of-n systems: an exact formula for stationary availability and multi-objective configuration design based on mathematical programming and TOPSIS. *Journal of Computational and Applied Mathematics*, 330: 1007-1015, 2018.

[10] Deb, K., Agrawal, S., Pratap, A. and Meyarivan, T., A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In International conference on parallel problem solving from nature (pp. 849-858). Springer, Berlin, Heidelberg, 2000.

[11] Atiquzzaman, M., Liong, S. Y. and Yu, X., Alternative decision making in water distribution network with NSGA-II. *Journal of water resources planning and management*, 132(2): 122-126, 2006.

[12] Preis, A. and Ostfeld, A., Multiobjective contaminant sensor network design for water distribution systems. *Journal of Water Resources Planning and Management*, 134(4): 366-377, 2008.

[13] Farmani, R., Savic, D. A. and Walters, G. A., Exnet benchmark problem for multi-objective optimization of large water systems. *Modelling and control for participatory planning and managing water systems*, 2004.

[14] Davies, D. L. and Bouldin, D. W., A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2): 224-227, 1979.

23

Updating the OSPF routing protocol for communication networks by optimal decision-making over the k-shortest path algorithm

Silvia Carpitella [♭♯], Manuel Herrera[♮], Antonella Certa[♭] and Joaquín Izquierdo[♯1]

(♭) Dipartimento di Ingegneria,
Università degli Studi di Palermo,
(♮) Institute for Manufacturing, Dept. of Engineering,
University of Cambridge,
(♯) Instituto de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

Internet routing protocols such as Routing Information Protocol (RIP) pre-compute all the shortest paths by Dijkstra's algorithm (shortest path first, SPF) based on the number of hops between one node and another. Every time any communication is intended, RIP looks-up for the optimal choice in a routing table. This is a high speed method in the decision-making process but not necessary fast for data traffic as it does not take into account any real-time measure of route congestion. Open Shortest Path First (OSPF) presents a dynamic version of this problem by computing the shortest paths taking into account network features such as bandwidth, delay and load. OSPF thereby maintains link-state databases updated at near real-time at every router. Although OSPF protocol is widely used, the Enhanced Interior Gateway Routing Protocol (EIGRP) presents another option for taking the optimal routing. At EIGRP, each node independently runs an algorithm to determine the shortest path (Dijkstra's algorithm) from itself to every other node in the network. In addition to the routing table, EIGRP uses neighbour information (neighbour nodes on node j are those having directly connection node$_j$) and a table of favourite (commonly used) routes.

In this context, a multi-criteria decision-making (MCDM) approach may represent an alternative perspective for the automatic selection of an optimal path among the k-shortest paths.

MCDM methods effectively support a plethora of decision problems, their crucial role being widely acknowledged [8]. With respect to routing protocols in communication networks, a final decision on selecting the optimal path depends on various evaluation criteria. These sometimes are mutually dependent and conflicting with each other. This is the case of criteria such as traffic density, path length, data type, and key performance indicators. Under such criteria,

[1]e-mail: jizquier@upv.es

the objective is to ultimately select one path among the k-shortest paths. MCDM methods have the ability of going towards the solution that represents a best trade-off for the selected network intends and satisfies their multiple aspects regarding their mutual importance. MCDM are capable of managing both qualitative and quantitative aspects when it is required an evaluation concerning a set of alternatives [10]. Several MCDM methods have been proposed in the existing literature, each one being characterised by specific procedures and objectives. However, common points for MCDM methods are that they can be mainly aimed at: selecting the best option among various alternatives, ranking alternatives to establish their weights and/or to draw up a list of priorities [11], and clustering alternatives into different groups on the basis of their common features [3].

Among MCDM methods, the fuzzy evolution of the Technique for Order Preference by Similarity to Ideal Solutions (TOPSIS), that is the FTOPSIS [4] method, is herein proposed to get the ranking of the possible k-paths related to a communication network according to the evaluation of suitable criteria and to simultaneously take into account uncertainty affecting input data as it is, for instance, the traffic density.

2 Existing approaches

The literature on routing protocols is commonly focused on protective strategies for communication networks [1]. This is the case of the creation of dedicated or shared backup networks among multiple connections [6] that provide spare capacity to a communication network and that allow to put in practice either reactive and proactive restoration schemes [7]. Using routing protocols based on k-shortest paths [5] instead of the single Dijkstra's algorithm (like in OSPF) is an alternate way to protect the communication network from disruptions [9]. Working with the k-shortest paths also enhances the adaptation capability of the network with respect to variations in the traffic density, not necessary facing disruption but with the objective of higher speed of data packets travelling through the network.

An interesting alternative to the Eppstein k-shortest paths algorithm is the loop-less approach which uses $k-1$ deviations of the main Dijkstra's shortest path. This was firstly developed by Yen [12] and has been successfully adapted further to communication networks [2]. The algorithm encompasses two main steps: determining the first of the k-shortest path, and then determining all other k-shortest paths. It is used an auxiliary matrix working as a container for the candidates to k-shortest paths and from which it is selected the minimum length path in an iterative process.

3 The FTOPSIS to rank the possible k-paths

As already expressed in the introduction, final decisions about the shortest path depend on evaluation criteria such as traffic density, path length, data type and key performance indicators. These often are interdependent and sometimes conflicting with each other.

Being the FTOPSIS technique the fuzzy evolution of the TOPSIS method, it allows to better represent practical real-life situations, since eliciting exact crisp numerical values may be diffi-

cult. Indeed, the TOPSIS ranks alternatives just on the basis of their crisp ratings on various qualitative and/or qualitative criteria, opportunely weighted.

The steps required to apply the FTOPSIS method are synthesised in Figure 1.

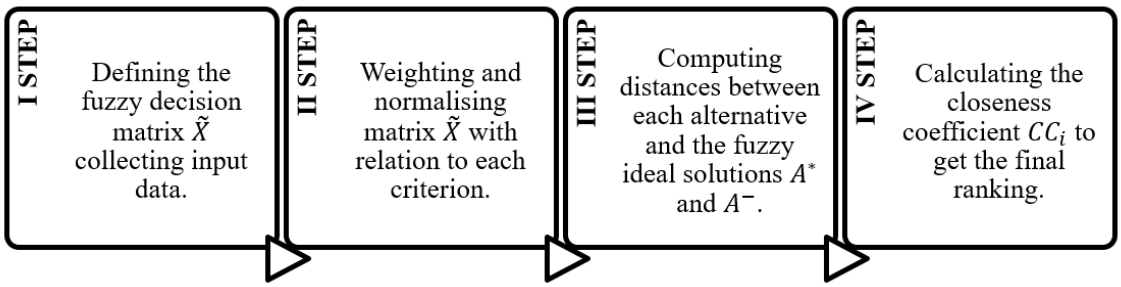


Figure 1: Steps representing the FTOPSIS procedure.

I STEP: defining the fuzzy decision matrix $\tilde{X}$ collecting input data.
It is firstly necessary to collect all the evaluations of the alternatives in the fuzzy decision matrix (Table 1):

$$\tilde{X} = \begin{bmatrix} \tilde{X}_{11} & \dots & \tilde{X}_{1n} \\ \vdots & \ddots & \vdots \\ \tilde{X}_{m1} & \dots & \tilde{X}_{mn} \end{bmatrix}, \tag{1}$$

The generic fuzzy number $\tilde{x}_{ij}$ represents the rating of alternative i under criterion j. In the present case, we take into account triangular fuzzy numbers (TFNs), characterized by ordered triples:

$$\tilde{x}_{ij} = (a_{ij}, b_{ij}, c_{ij}) \tag{2}$$

	DECISION MATRIX											
	Traffic Density (min)			**Path Lenght (min)**			**Data Type (max)**			**Network Features (max)**		
	25%			**25%**			**25%**			**25%**		
Path 1	a_{11}	b_{11}	c_{11}	a_{12}	b_{12}	c_{12}	a_{13}	b_{13}	c_{13}	a_{14}	b_{14}	c_{14}
Path 2	a_{21}	b_{21}	c_{21}	a_{22}	b_{22}	c_{22}	a_{23}	b_{23}	c_{23}	a_{24}	b_{24}	c_{24}
Path 3	a_{31}	b_{31}	c_{31}	a_{32}	b_{32}	c_{32}	a_{33}	b_{33}	c_{33}	a_{34}	b_{34}	c_{34}
Path 4	a_{41}	b_{41}	c_{41}	a_{42}	b_{42}	c_{42}	a_{43}	b_{43}	c_{43}	a_{44}	b_{44}	c_{44}
Path 5	a_{51}	b_{51}	c_{51}	a_{52}	b_{52}	c_{52}	a_{53}	b_{53}	c_{53}	a_{54}	b_{54}	c_{54}
Path 6	a_{61}	b_{61}	c_{61}	a_{62}	b_{62}	c_{62}	a_{63}	b_{63}	c_{63}	a_{64}	b_{64}	c_{64}
Path 7	a_{71}	b_{71}	c_{71}	a_{72}	b_{72}	c_{72}	a_{73}	b_{73}	c_{73}	a_{74}	b_{74}	c_{74}
Path 8	a_{81}	b_{81}	c_{81}	a_{82}	b_{82}	c_{82}	a_{83}	b_{83}	c_{83}	a_{84}	b_{84}	c_{84}
Path 9	a_{91}	b_{91}	c_{91}	a_{92}	b_{92}	c_{92}	a_{93}	b_{93}	c_{93}	a_{94}	b_{94}	c_{94}

Table 1: Fuzzy Decision Matrix $\tilde{X}$.

II STEP: weighting and normalising the previously defined matrix with relation to each criterion. The second step of the procedure consists in obtaining a matrix $\tilde{U}$ by weighting and

normalising matrix $\tilde{X}$. In particular, elements of matrix $\tilde{U}$ are defined as follows:

$$\tilde{u}_{ij} = \left(\frac{a_{ij}}{c_j^*}, \frac{b_{ij}}{c_j^*}, \frac{c_{ij}}{c_j^*}\right) \cdot w_j, \;\; j \in I', \tag{3}$$

$$\tilde{u}_{ij} = \left(\frac{a_j^-}{c_{ij}}, \frac{a_j^-}{b_{ij}}, \frac{a_j^-}{a_{ij}}\right) \cdot w_j, \;\; j \in I'', \tag{4}$$

where I' is the subset of criteria to be maximized, I'' is subset of criteria to be minimized, w_j represents the relative importance weight of criterion j, c_j^* and a_j^- are calculated as:

$$c_j^* = \max_i c_{ij} \;\; if \;\; j \in I', \tag{5}$$

$$a_j^- = \min_i a_{ij} \;\; if \;\; j \in I''. \tag{6}$$

III STEP: computing distances between each alternative and the fuzzy ideal solutions A^* and A^-. At the present stage, each fuzzy alternative has to be compared with both a fuzzy positive ideal solution A^* and a fuzzy negative ideal solution A^-, namely:

$$A^* = (\tilde{u}_1^*, \tilde{u}_2^*, \ldots, \tilde{u}_n^*), \tag{7}$$

$$A^- = (\tilde{u}_1^-, \tilde{u}_2^-, \ldots, \tilde{u}_n^-), \tag{8}$$

where $\tilde{u}_j^* = (1,1,1)$ and $\tilde{u}_j^- = (0,0,0)$, $j = 1 \ldots n$. In detail, distances between each alternative and these points are computed through the vertex method [4], for which the distance $d(\tilde{m}, \tilde{n})$ between two TFNs $\tilde{m} = (m_1, m_2, m_3)$ and $\tilde{n} = (n_1, n_2, n_3)$ is the crisp value:

$$d(\tilde{m}, \tilde{n}) = \sqrt{\frac{1}{3}[(m_1 - n_1)^2 + (m_2 - n_2)^2 + (m_3 - n_3)^2]}. \tag{9}$$

Then, aggregating with respect to the whole set of criteria, the related distances of each alternative i from A^* and A^- are:

$$d_i^* = \sum_{j=1}^{n} d(\tilde{u}_{ij}, \tilde{u}_j^*), \;\; i = 1 \ldots n, \tag{10}$$

$$d_i^- = \sum_{j=1}^{n} d(\tilde{u}_{ij}, \tilde{u}_j^-), \;\; i = 1 \ldots n. \tag{11}$$

IV STEP: calculating the closeness coefficient CC_i to get the final ranking. The mentioned closeness coefficient CC_i is calculated as:

$$CC_i = \frac{d_i^-}{d_i^- + d_i^*} \tag{12}$$

To get the final ranking it is necessary to sort the values of the closeness coefficient related to each alternatives in a decreasing way. That means the path with a higher value of CC_i will be selected.

4 Conclusions

The research deals with the topic of internet communication networks and it is focused on the problem of selecting the shortest path among the possible k-paths. Various existing protocols are capable to compute this kind of selection by taking into account networks features and a multi-criteria decision making approach has been herein proposed as alternative way to sort such problem out. The FTOPSIS technique appears to be suitable because of its capability to rank even a huge number of options. The FTOPSIS makes use of fuzzy numbers instead of crisp ones, so that uncertainty affecting input data can effectively be managed, and the final decision circa the shortest paths is taken by assigning different degrees of importance to those criteria of interest for the problem under analysis. The immediate benefits of the FTOPSIS over the k-shortest paths are the improvement on network protection strategies for abnormal scenarios as well as the speed up of network data-traffic under regular conditions.

References

[1] Awoyemi, B. S., Alfa, A. S. and Maharaj, B. T., Network restoration for next-generation communication and computing networks. *Journal of Computer Networks and Communications*, 2018.

[2] Bouillet, E., Ellinas, G., Labourdette, J. F. and Ramamurthy, R., *Path routing in mesh optical networks*. Wiley, 2007.

[3] Certa, A., Enea, M., Galante, G. and La Fata, C.M., ELECTRE TRI-based approach to the failure modes classification on the basis of risk parameters: An alternative to the risk priority number. *Computers & Industrial Engineering*, 108: 100–110, 2017.

[4] Chen, C.T., Extensions of the TOPSIS for group decision-making under fuzzy environment. *Fuzzy Sets and Systems*, 114(1): 1–9, 2000.

[5] Eppstein, D., Finding the k shortest paths. *SIAM Journal on computing*, 28(2): 652-673, 1998.

[6] Hegde, S., Koolagudi, S. G. and Bhattacharya, S., Path restoration in source routed software defined networks. In 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN) (pp. 720-725). IEEE, 2017.

[7] Kuipers, F. A., An overview of algorithms for network survivability. *ISRN Communications and Networking*, 2012.

[8] Kumar, A., Sah, B., Singh, A.R., Deng, Y., He, X., Kumar, P. and Bansal, R.C., A review of multi criteria decision making (MC DM) towards sustainable renewable energy development. *Renewable and Sustainable Energy Reviews*, 69: 596–609, 2017.

[9] Lian, J., Zhang, Y. and Li, C. J., An efficient k-shortest paths based routing algorithm. In *Advanced Materials Research. Trans Tech Publications*, 532: 1775-1779, 2012.

[10] Mulliner, E., Malys, N. and Maliene, V., Comparative analysis of MCDM methods for the assessment of sustainable housing affordability. *Omega*, 59: 146–156, 2016.

[11] Vargas, L.G., De Felice, F. and Petrillo, A., Editorial journal of multicriteria decision analysis special issue on "Industrial and Manufacturing Engineering: Theory and Application using AHP/ANP". *Journal of Multi-Criteria Decision Analysis*, 24(5–6): 201–202, 2017.

[12] Yen, J. Y., An algorithm for finding shortest routes from all source nodes to a given destination in general networks. *Quarterly of Applied Mathematics*, 27(4): 526-530, 1970.

24

Optimal placement of quality sensors in water distribution systems

Jorge Francés-Chust [♭], Silvia Carpitella[♮*], Manuel Herrera[♯], Joaquín Izquierdo[*1] and Idel Montalvo[⊥]

(♭) Aguas Bixquert, S.L.,
c/ José Chaix 7, 46800 Xátiva, Valencia,
(♮) Dipartamento di Ingegneria,
Università degli Studi di Palermo,
(♯) Institute for Manufacturing, Dept. of Engineering,
University of Cambridge,
(∗) FluIng-IMM,
Universitat Politècnica de València,
(⊥) Ingeniousware GmbH,
Jollystraße 11, 76137 Karlsruhe, Germany.

1 Introduction

Water supply infrastructures are crucial for the sustainable existence and development of modern cities [1, 2]. Water distribution systems (WDSs) are complex structures formed by many elements designed and erected to transport water of sufficient quality from water sources to consumers. The amount of the above elements, which can reach up to tens of thousands of links and junctions, their frequently wide spatial dispersion and the WDS characteristic of being very dynamic structures make the management of real WDSs a complex problem [3–5]. Moreover, although the main objective is to supply water in the quantity and quality required, other requirements are essential, namely maintaining conditions far from failure scenarios [6, 7], ability to quickly detect sources of contamination intrusion [8,9], minimization of leaks [10–12], etc.

Most of these objectives may be achieved through suitable location of sensors along the network and, currently, an increasing number of efforts are carried out in this direction [12–14]. The identification of potential contaminant intrusion in water networks is a crucial point to fully guarantee water quality in WDSs. As a consequence, water utilities are bound to measure water quality parameters continuously, so that quality can be adequately monitored. To this end, an optimal lattice of sensors should be designed that covers strategical points of the water network [15]. It is a matter of safety and security arrangement in WDS management, and sensors cannot be randomly placed along the network. Placing sensors may seem simple at the beginning, but considering sensor station costs and the extension of the network that should

[1]e-mail: jizquier@upv.es

be covered, it turns out to be a challenging problem.

The plurality of potential contaminants, the identification of the contaminant sources in the network, and the reaction time of the utilities to deal with a contamination event are also important elements to consider. This work is not intended to cover all the aspects related to network protection against potential contaminant intrusion. It will rather concentrate on proposing a solution just for the sensor placement problem, namely, optimally determining the number of sensors and their locations. And we address this optimization problem from a multi-objective perspective.

Several goals should be taken into account when placing water quality sensors. Optimal sensor placement aims to achieve early contaminant detection and seclusion of affected areas so that the public exposure to contamination be minimum. First, it is desired to identify quality problems as soon as possible, it means, to minimize the detection time. Second, irrespective of the location of the contaminant source, at least one sensor should always be able to identify a quality problem; this amounts to maximizing the coverage of protection. Additionally, the bulk of poor or bad quality water consumed should be minimized; this, specifically, involves that high population density areas have to receive special attention compared to other areas with much lower consumption rate. And, importantly, the cost, which is directly proportional to the number of installed sensors, should be kept to a minimum.

These objectives are mutually conflicting and improving one of them will probably result in a detriment for another. The rationale is clear. For example, maximizing the protection coverage in the network will require either to increase the number of sensors (it means the cost) or to probably be bound to accept larger detection times. Consequently, the final solution will result from a compromise among objectives rather than from a unique "best alternative". Suitably solving problems of this nature requires the use of a multi-objective approach. Such an approach is able, for example, to answer marginal cost questions, such as if it is worth buying an additional sensor to get a reasonable improvement in another objective, because there is no way to know how much improvement in protection coverage and detection time will bring that additional sensor. Those are the kinds of questions that a multi-objective approach helps to answer. We claim that those are the kind of questions and answers needed to eventually find a sensor placement solution that represents a good trade-off among all the objectives involved.

In this contribution we present the necessary materials and methods. Then, we develop contaminations scenarios and evaluate the considered objectives based on the so-called contamination matrix concept. Next, we develop a multi-objective solution using a well-known multi-objective optimization algorithm [16]. A use case corresponding to a medium-size water distribution network is presented together with the obtained results and a thorough discussion.

2 Contamination scenarios and evaluation of objectives

WDSs are vulnerable against various sources of accidental and intentional contaminations. The US EPA [17] considers three protocol steps: (i) detection of contaminant presence, (ii) source identification and (iii) consequence management. To develop suitable Early Warning Systems (EWSs) for alerting the consumers and isolating contaminated areas, optimal location

of measurement devices is paramount to accurately identify the source of contamination. Hart and Murray [18] describe EWSs and conclude that sensor placement is one of the critical aspects of the design of EWSs.

2.1 The objectives

The objectives we consider to solve the sensor location problem are: detection time, coverage of protection, affected population and implementation costs.

- Detection time: First we consider the time elapsed since the contamination is introduced through one node till one sensor is reached by the contaminant. The detection time is the average of those times calculated for all the nodes. For the case that no sensor detects the contaminant we use a null detection time. This circumstance will heavily penalized by other objectives in charge of evaluating the amount of contaminated water and the detection failure.

- Detection Failure: It is an index related to the amount of contamination cases happening downstream of all sensor locations, and where no detection is possible considering the current sensor placement solution.

- Contaminated water consumption: It refers to the amount of contaminated water consumed in the network before the contaminant has reached at least one of the sensor locations.

- Implementation costs: It is the cost of the solution expressed as a function of the number of sensors to be installed in the network multiplied by an estimated global cost per sensor.

2.2 The contamination matrix

The first step for solving a sensor placement problem is the generation of a contamination matrix. This matrix (of size number of nodes times number of nodes) stores, for every single contamination alternative in a given node, how long it takes to reach each of the other network nodes. Once all the contamination alternatives have been calculated, the search of Pareto dominant solution can be started.

3 Algorithm and software for calculations

Many approaches may be used to find the Pareto front in a multi-objective optimization problem. Here, we use Agent Swarm Optimization (ASO) [19]. ASO combines multi-objective evolutionary algorithms, rule-based agents and data analytics, intelligently integrating problem-domain knowledge within the optimization process and learning engineer's preferences to achieve more real results.

In this research we introduce basic rules for reducing the decision space. A "normal" agent could locate a sensor at virtually any node of the network. However, based on the experience of the authors on solving several use cases it was found that: (i) locating sensors too downstream of the network will probably guarantee a good coverage of the network but will result in big

detection time; (ii) locating sensors too much upstream of the network will help to detect events faster but the coverage of the network will be compromised.

These two ideas suggest drawing boundaries that help the algorithm delimit nodes of bigger interest to host sensors. These nodes should be neither too close to the water sources nor at the very end of the piping network. One issue is the reaction, which is the time until operation actions are enforced. This is used by the rule-based agents to define a border so that nodes downstream of that border cannot host a water quality sensor. Another issue is the distance to the water sources. Another boundary should be drawn to discard too upstream nodes as eligible for sensor location. Incidentally, these boundaries help reduce the search space.

Another key aspect from the computational point of view is the size of the contamination matrix. Such a matrix cannot be fully hosted in RAM for large WDSs. An MS SQL database is used to hold that matrix in this research. Then, calculations have been suitable encoded.

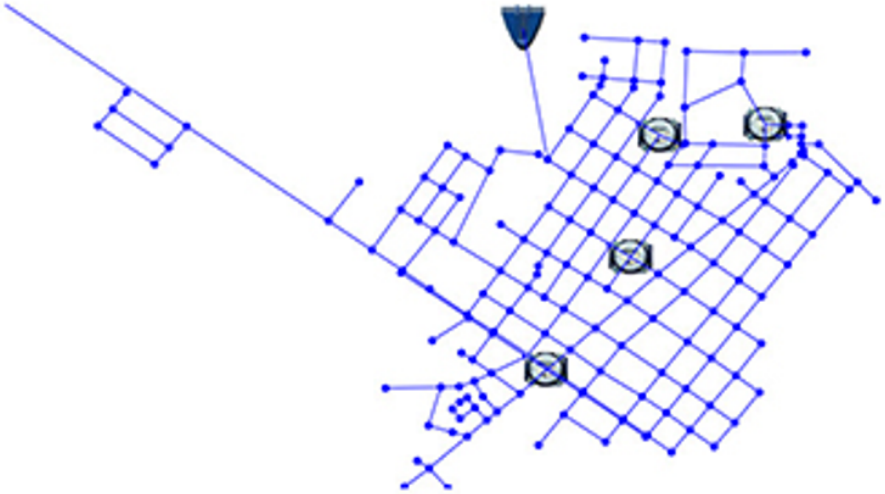


Figure 1: Network model of San José with 4 water quality sensors.

4 Case study

We consider a modified version of the water network of San José de las Lajas, a small town in Cuba, closed to Havana, with more than 24 km of pipes and one single entry point. Fig. 1 represents the network with a solution for placing 4 water quality sensors. This solution will be specially marked in red in Fig. 2-4 for a better interpretation of results. The execution of sensor placement results in the charts represented in fig. 2 to 4.

In Fig. 2 it can be seen what happens with the contaminated water that is consumed if the average detection time changes. For very low detection times we are not standing at solutions that can detect a significant number of contamination event. Note that the detection time is assumed equal to zero for those non-detected events.

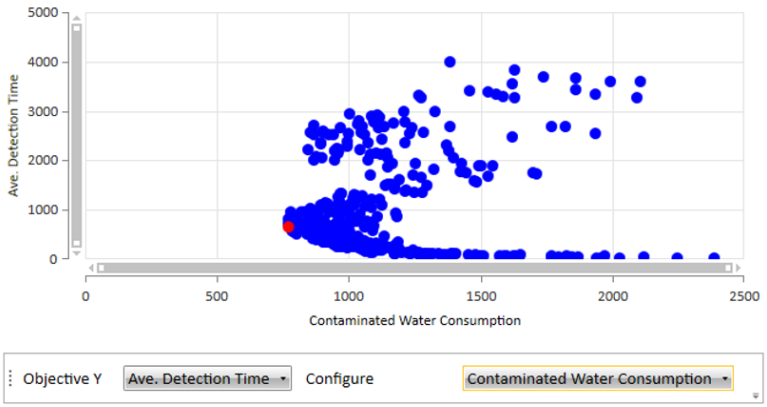


Figure 2: Average detection time vs contaminated water consumption.

Fig. 3 relates the amount of detection failure with the contaminant detection time. Using solutions with very high detection time means that sensors will be located at nodes very downstream in the network. In these cases, it takes a little longer to detect a contaminant (as average considering all possible contamination) but the detection failure is much lower. Again, from fig. 3 it can be seen that for higher values of detection time, the detection failure is relative lower.

Fig. 3 relates the amount of detection failure with the contaminant detection time. Using solutions with very high detection time means that sensors will be located at nodes very downstream in the network. In these cases, it takes a little longer to detect a contaminant (as average considering all possible contamination) but the detection failure is much lower. Again, from fig. 3 it can be seen that for higher values of detection time, the detection failure is relative lower.

Fig. 4, on the other hand, shows that the average volume consumed of contaminated water can be increased because of two main reasons: either we are standing at solutions with higher detection failure in average (sensors located too close to the sources that cannot detect contamination downstream) or we are standing at solutions where sensors are located at nodes in very downstream positions, which requires longer in average to receive the contamination effects. The relation between detection time and detection failure was previously mentioned and can be seen in Fig. 3.

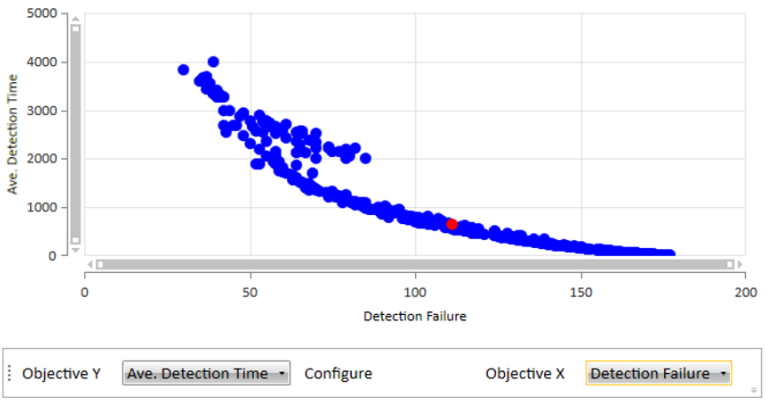


Figure 3: Average detection time vs detection failure.

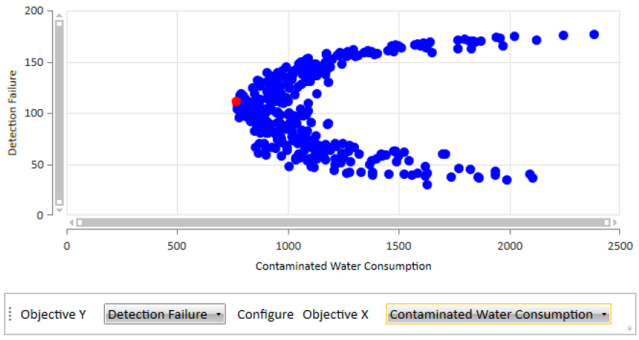


Figure 4: Average detection failure vs average contaminated water consumed.

5 Conclusions

Two important questions have to be answered in order to properly protect a water network against accidental or provoked contamination events and water quality problems: how many sensors are needed and where to place them. Answering these questions requires a decision about the criteria and requirements to be considered for achieving a good solution combined with a multi-objective approach for solving the problem. The final solution should be based on a trade-off among the objectives involved and the tolerance to "fail" that we could have in each of them. An improvement in all the objectives analyzed can be done by adding new sensors but this, of course, has the consequence of increasing the costs which can be a constraint for the implementation of the solution.

References

[1] Gandy, M., Rethinking urban metabolism: Water, space and the modern city. *City*, 8(3), 2004.

[2] Hoekstra, A.Y. and Mekonnen, M.M., The water footprint of humanity, *PNAS*, 109(9): 3232-3237, 2012.

[3] Perelman, L. and Ostfeld, A., Water distribution systems simplifications through clustering, *Journal of Water Resources Planning and Management*, ASCE, 138(3): 218-229, 2012.

[4] Izquierdo, J., Montalvo, I., Pérez-García, R. and Matías, A., On the Complexities of the Design of Water Distribution Networks. *Mathematical Problems in Engineering*, 2012: 1-25, 2012.

[5] Diao, K., Fu, G., Farmani, R., Guidolin, M. and Butler, D., Twin-hierarchy decomposition for optimal design of water distribution systems. *Journal of Water Resources Planning and Management*, 142(5), p.C4015008, 2015.

[6] Ostfeld, A., Oliker, N. and Salomons, E. (2014) "Multi-objective optimization for least cost design and resiliency of water distribution systems", *Journal of Water Resources Planning and Management Division*, ASCE, 140(12), 04014037.

[7] Herrera, M., Abraham, E. and Stoianov, I., A graph-theoretic framework for assessing the resilience of sectorised water distribution networks. *Water Resources Management*, 30(5): 1685-1699, 2016.

[8] Islam, N., Farahat, A., Al-Zahrani, M.A.M., Rodríguez, M.J. and Sadiq, R., Contaminant intrusion in water distribution networks: review and proposal of an integrated model for decision making. *Environ Rev*, 23(3):337-352, 2015.

[9] Nafi, A., Crastes, E., Sadiq, R. et al. Intentional contamination of water distribution networks: developing indicators for sensitivity and vulnerability assessments Stoch Environ Res Risk Assess, 32: 527, 2018.

[10] Covas, D. and Ramos, H., Practical methods for leakage control, detection, and location in pressurized systems. In BHR Group Conference Series Publication, Bury St. Edmunds; Professional Engineering Publishing, 37: 135-152, 1999.

[11] Candelieri, A., Conti, D. and Archetti, F., A graph based analysis of leak localization in urban water networks. *Procedia Engineering*, 70: 228-237, 2014.

[12] Zhao, Y., Schwartz, R., Salomons, E., Ostfeld, A. and Poor, H.V., New formulation and optimization methods for water sensor placement. *Environmental Modelling & Software*, 76: 128-136, 2016.

[13] Rathi, S. and Gupta, R., Optimal sensor locations for contamination detection in pressure-deficient water distribution networks using genetic algorithm. *Urban Water Journal*, 14(2): 160-172, 2017.

[14] Antunes, C.H. and Dolores, M., Sensor location in water distribution networks to detect contamination events—A multiobjective approach based on NSGA-II. In Evolutionary Computation (CEC), 2016 IEEE Congress, 1093-1099, 2016.

[15] Oliker, N. and Ostfeld, A., Network hydraulics inclusion in water quality event detection using multiple sensor stations data, *Water Research*, 80: 47-58, 2015.

[16] Montalvo, I., Izquierdo, J., Herrera, M. and Pérez-García, R., Water Distribution System Computer-aided Design by Agent Swarm Optimization. *Computer-Aided Civil and Infrastructure Engineering*, 29(6): 433-448, 2014.

[17] US EPA (2003) Response protocol toolbox: planning for and responding to drinking water contamination threats and incidents. `http://www.epa.gov/safewater/watersecurity/pubs/guide_response_overview.pdf`.

[18] Hart, WE and Murray, R., Review of sensor placement strategies for contamination warning systems in drinking water distribution systems. *J Water Resour Plan Manag*, 136(6): 611–619, 2010.

[19] Montalvo, I., Izquierdo, J., Herrera, M. and Pérez-García, R., Water Distribution System Computer-aided Design by Agent Swarm Optimization. *Computer-Aided Civil and Infrastructure Engineering*, 29(6): 433-448, 2014.

25

Mapping musical notes to socio-political events

C.-H. Lai[♭1], N. Kokulan[♭] and S. Kenndy[♮]

(♭) School of Computing and Mathematical Sciences,
University of Greenwich,
(♮) School of Design,
University of Greenwich.

1 Introduction

Musical genres are labels created to characterize different types of music. In the past, categorising using musical genres was carried out manually by humans. Nowadays these may be replaced by automatic musical genre classification replacing the manual procedure. This is a topic which has seen an increased interest recently as one of the cornerstones of the general area of Music Information Retrieval. It is also certain that music has a significant commercial, cultural and political impact on real-world events bringing positive change and unity into the commercial, cultural and political world.

In this paper, a neural and fuzzy technique is investigated for two main aims: (1) automatic musical genre classification and (2) mapping music notes to socio-political events. The validation of the algorithm is studied by using historical data.

Section 2 provides some basic methods for audio features extraction. Section 3 introduces briefly the neuro-fuzzy modelling of this study. Section 4 examines the genre classification and section 5 examines the mapping of musical pieces to one socio-political event. Some early conclusions are drawn.

2 Audio features extraction

Audio data are time series where the vertical axis corresponds to the current amplitude of a loudspeaker?s membrane and the horizontal axis corresponds to time. In order to obtain high accuracy for classification and segmentation, it is important to select specific features of audio files. Generally, audio file analysis is based on the nature of the waveform. Therefore, the features are selected on the basis of their numerical values. In this paper volume and aero-crossing rate are the two main features extracted in the numerical studies.The definitions for these two

[1]e-mail: C.H.Lai@gre.ac.uk

features are provided below.

Volume is one feature which represents the level of sound of the audio signal. This is represented by the amplitude and is also referred to as energy or intensity of audio signals.

$$V = \sum_{i=1}^{n} |S_i| \tag{1}$$

where V is the volume and S_i is the amplitude of frame i assuming the signal is subdivided into n frames.

The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or vice versa. This feature has been used in both speech recognition and music information retrieval.

The typical mathematical tool used in treating a signal is the Fast Fourier Transform of it from time domain to frequency domain.

3 Neuro-fuzzy modelling

Neural networks are good at recognising patterns but poor at explaining the decision making process. On the other hand fuzzy logic systems are good at explaining their decisions. However, they are unable to automatically acquire the rules and membership functions. A combination of these two systems generates a useful tool which overcomes the weaknesses.

A neuro-fuzzy system is a neural network which is functionally equivalent to a fuzzy inference system. Without any prior knowledge of rules and membership function it can be trained to develop fuzzy rules [1] and determine membership functions for the input and output variables of the system [2]. This modelling approach can be used to (1) build a model that can predict the behaviour of the underlying system and (2) control the system.

In this paper the Adaptive Neuro-Fuzzy Inference System (ANFIS) proposed by Jang [3] is adopted for the computational tests. This is one of the specific approaches of neuro-fuzzy systems in which a fuzzy system is implemented in the framework of an adaptive network. The ANFIS is similar to a multi-layer neural network with five layers. The first layer is an input layer implementing the first-order Takagi Sugeno inference system with two inputs and one output. It is referred to as the fuzzification layer and is used to determine the membership grades. The membership function used in the system can be any continuous and piecewise differentiable function such as the bell shaped trapezium, triangular, or Gaussian distribution. The second and third layers contain fixed nodes that provide the antecedent parts in each rule. The fourth layer contains nodes that are adaptive and computes the first-order Takagi-Sugeno rule output for each fuzzy rule. The fifth output layer computes the weighted global output of the system.

4 Music genre classification model

The authors implemented fuzzy neural techniques based on the ANFIS system described above in order to classify a song or a short sound clip into its corresponding music genre. Our algorithms have two phases: (1) feature extraction and (2) model implementations.

In feature extraction, six features have been used, including Short Time Energy (the energy of the signal in each analysis frame/window), Spectral Centroid (centre of gravity of the magnitude spectrum of the Fourier transform), Zero-crossing (Mean of zero crossings across time frames in the texture window), Spectral Flux (the squared difference between the normalized magnitudes of successive spectral distributions), and Spectral Rolloff (the frequency below 85% of the magnitude distribution).

In the validation process five types of music, namely, blues, classical, country, disco and pop were used to examine the automatic music genre classification methods described above. A total of 125 songs along the horizontal axis as shown in Figure 1 was used. The vertical axis uses the genres 1, 3, 5, 7 and 9 to represent blues, classical, country, disco and pop, respectively. For each genre there are 25 songs. The model is used to predict the genres provided by the songs. Note that the predictions (red plus) are in good agreement with the experimental data (blue star). As such the model is ready to be used to classify the genres of any song provided.

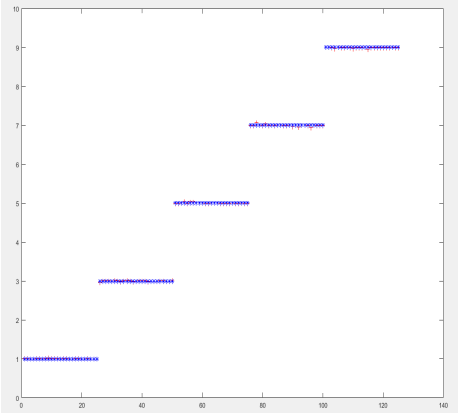


Figure 1: Genre classification results: Comparison between model prediction (red plus) and historical data (blue star).

5 Predication of one political event using popular music

A correlation between the popular music prior to election and the outcome of the election possibly exists. This indicates a dependency of the election results on the popular music. A model was developed for the purpose of validating the prediction. Training data which maps the election results to the popular music using the music features. This has been done using

the above fuzzy neural network techniques. Data on the popular music hit list one year prior to the election and also data on the election results from June 1970 to May 2010 were collected.

In the validation process 26 songs from the hit list were used to check the model. Figure 2 shows the prediction (red star) of the election results in comparison with the historical data (blue circle). Along the vertical axis 1 represents the labour party and 0 represents the conservative party winning an election and the horizontal axis represents the hit songs. It can be seen the model predictions show good agreement with the historical data. Only a limited amount of data was available on the songs and larger data set would produce better and reliable prediction.

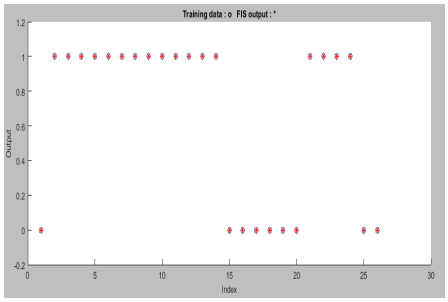


Figure 2: General election results: Comparison between model prediction (red plus) and historical data (blue star).

6 Conclusion

First, five types of music, including, blues, classical, country, disco and pop were used to automate music genre classification. A total of 125 songs were used in the experiments. The model developed is used to predict the genres provided the songs. Computational results demonstrated that the predictions are in good agreement with the experimental data. The model may now be used to classify the genres of any song provided.

Second, a correlation between the popular music prior to election and the outcome of the election is assumed. This could mean a dependency of the election results on the popular music. Therefore, a model is developed using a set of training data to map the election results to the popular music. This is done using neural and fuzzy network techniques. Data related to the popular music one year prior to the election and also data on the election results were collected for the study. A total of 26 songs from the hit list was used to check the model. The prediction of the election results in comparison with the historical data is well with the party winning the election matching the prediction. The model predictions show good agreement with the historical data.

Only a limited amount of data was available on the songs used in the above computational experiments. More data would allow for more reliable prediction. Future work includes refining

the model for the political event prediction to become more sophisticated by adapting other machine learning techniques with a large amount of data. This technique described is suitable for understanding the correlation of music and political events.

References

[1] Zadeh, L. A., Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. on Systems, Man, and Cybernetics*, 3(1): 28–44, 1973.

[2] Denai, M. A., Palis, F. and Zeghbib, A., Modeling and control of non-linear systems using soft computing techniques. *Applied Soft Computing*. 7(3): 728–738, 2007.

[3] Jang, J.S.R., ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans Systems, Man, Cybernetics*, 23(3): 665–685, 1993.

26

Comparison between DKGA optimization algorithm and Grammar Swarm surrogated model applied to CEC2005 optimization benchmark

Gabriela Bracho[♭], David Martínez-Rodríguez[♮1], Ricardo Novella[♭], Cassio Spohr[♭] and R.-J. Villanueva [♮]

(♭) CMT – Motores Térmicos,
Universitat Politècnica de València,
(♮) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

In the last years, Computer Fluid Dynamics (CFD) models have been implemented in the design chain of the automotive sector in order to decrease costs and increase the control during the creation of thermodynamic engines. CFD models are based on the resolution of a high number of differential equations which have not easy analytical resolution. This is the reason why numerical methods are used on these kind of problems.

Nevertheless, numerical methods involve high number of computations and not every computer device can handle them. Execution time of this simulations might take days or even weeks until the results are obtained. The next step in the automotive sector related to the design of thermodynamic engines is the use of the CFD codes in order to optimize those parameters that are required by the regulators and the global market.

For this purpose, evolutionary optimization algorithms (EOA) are used. In order to optimize functions, EOA perform a high number of fitness function evaluations. This fact becomes a problem when the fitness function evaluation requires a high computational cost, even more if the evaluation time is in the order of days or weeks.

In this work we propose the use of surrogate models, which are simpler models that provide an approximation of the real value of the fitness function from its input parameters, in order to use them as objective function instead of CFD models on EOA algorithms.

The aim of this work is to compare the results of an optimization problem. In one hand, the execution of DKGA optimization algorithm with a small number of fitness function evaluations.

[1]e-mail: damarro3@upv.es

In the other hand, the optimization of a simplified model from the original function using Grammar Swarm, where the set of data has the same number of elements as the DKGA number of fitness function evaluations. Then, the results of the optimum found by both techniques are compared.

2 DKGA Optimization Algorithm

DKGA is an optimization algorithm, based on optimization genetic algorithm (GA) but with accelerated convergence. Chromosomes are represented in decimal format, and mutations are reduced as the algorithm progresses. It uses Punnett square technique in order to cross breed the parents [1].

To increase convergence, the mutation rate varies according to:

$$\tau_{it} = \tau_0 * exp(-\sigma \frac{iteration}{generations})$$

Algorithm 1 DKGA Algorithm - Minimization.

Require: $parents, generations, \tau_0, \sigma$
Ensure: $Min\ of\ f$

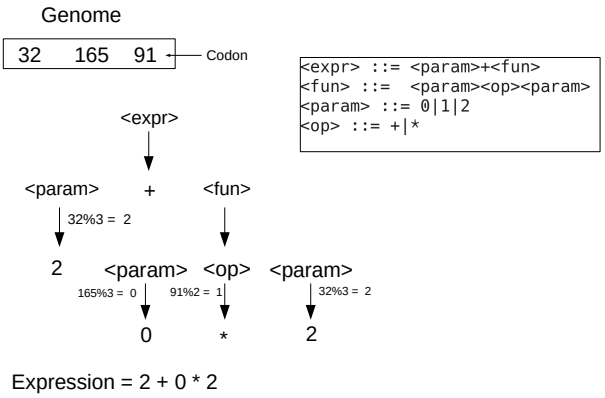
```
while Number of generations is not reached do
    Select best parents for next generation. (First generation generated randomly).
    Cross breed the parents using Punnett Square Technique.
    for Each children do
        if rand() ≤ τ_it then
            Mutate chromosomes of the children.
        end if
    end for
    Evaluate objective function for each children.
    Penalize children out of range.
    Sort population.
end while
```

The selection of this optimization technique is based on a previous work, where a combustion system model was optimized using this algorithm [2]. The results of this study are promising, but still leaves some doubts about the premature convergence of the solution because of the stack of the solution in possible local minima.

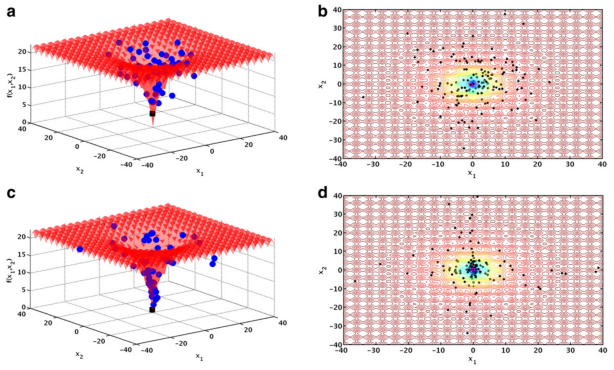
3 Grammar Swarm Algorithm

Grammar Swarm (GS) is a technique based on genetic programming that uses Particle Swarm Optimization algorithm (PSO) as search engine. This technique automatically obtains models from a set of functions in the form of Backus-Naur (BNF) and a set of data from the process to be modeled.

Grammar Swarm is based on two different tools interconnected. The first one is the interpreter, which produces a model from BNF grammar and a specific genome. The second tool is the search engine (PSO), which obtains the genome that once it has been translated by the interpreter, is able to reproduce the behavior of the data provided from the process that wants to be modeled [3].



(a) Grammar interpreter procedure. On the box, the BNF grammar.



(b) PSO evolution [4]. The dimensions of the search space are the codons of the genome, and the value of the objective function is the difference between the output of the model obtained with the genome and the real data.

4 Comparison between DKGA ans GS

In this work, we are comparing the performance of optimizing a model with Grammar Swarm and DKGA with the same number of function evaluations. Due to the high cost of the evaluation of a thermal engine CFD code, we are going to use models obtained from CEC2005 benchmark [5].
The comparison process is the following:

- Making a sample of the model parameters with the Latin Hypercube Sampling (LHS) technique and calculate its outputs.
- Obtaining a surrogated model of the original CEC2005 function with GS using the sampled points.
- Optimize the simplified (algebraic) model obtained previously.
- Calculate the value of the CEC2005 function on the optimum point obtained with the simplified model.
- Optimize the CEC2005 function with DKGA, forcing the maximum number of evaluations as the number of sampled points with LHS.

With this process, as it can be seen, the total number of evaluations of the original model (which is supposed to be computationally expensive) is the same in both techniques.

5 Results

The first four CEC2005 functions have been optimized with this procedure. 96, 220, 360 and 588 points of dimension ten are sampled with LHS for each CEC2005 function selected. We optimize 32 times with GS and DKGA each function in order to decrease the randomness of the optimization. In order to check the results with a more common optimization algorithm, PSO [4] is going to be used with the same procedure as DKGA to compare the results of both of them.

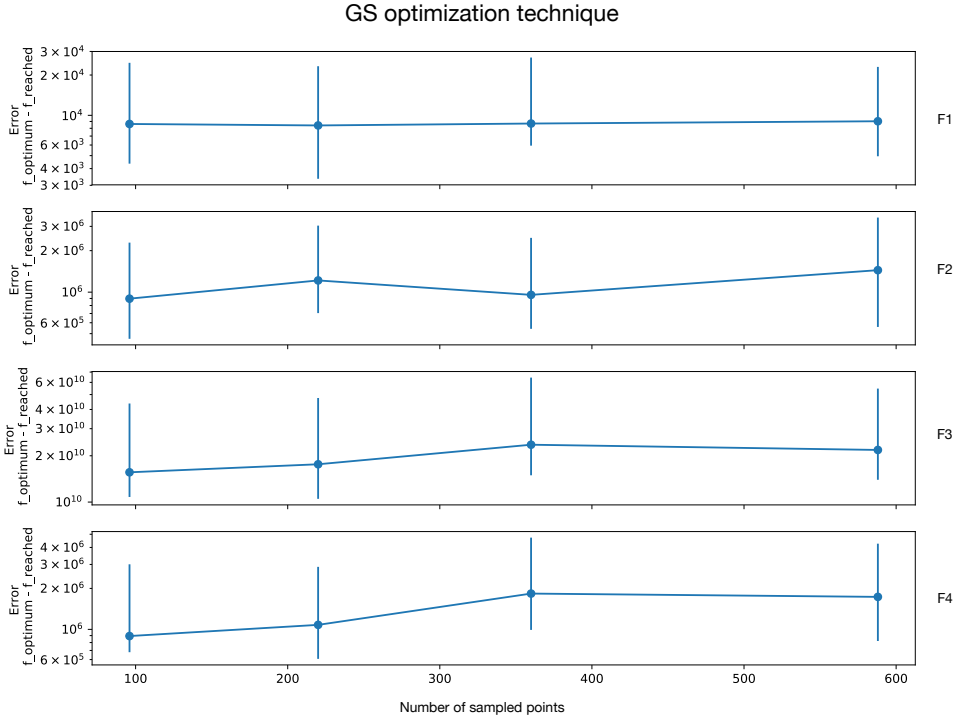


Figure 2: On the X axis, the number of points used as data to obtain the model. On the Y axis, the difference between the optimum point reached by GS and the real optimum.

6 Conclusions

As it can be seen comparing Figure 2 and Figure 3, both techniques may be considered of similar performance. A thing to be pointed out is the lack of improvement when the number of data is increased in GS. This might be caused because this technique is not able to provide any surrogated model with this few number of points. This fact also happens on DKGA due to the premature convergence of the method.

However, Grammar Swarm is not a good technique for obtaining simplified models with only a small set of points of a very complex model and Grammar Swarm is not a good technique for obtaining simplified models with only a small set of points of a more complex model. This can be confirmed when they are compared to PSO in Figure 4, where it can be seen that the error made with this few number of function evaluations if several orders of magnitude smaller than GS and DKGA.

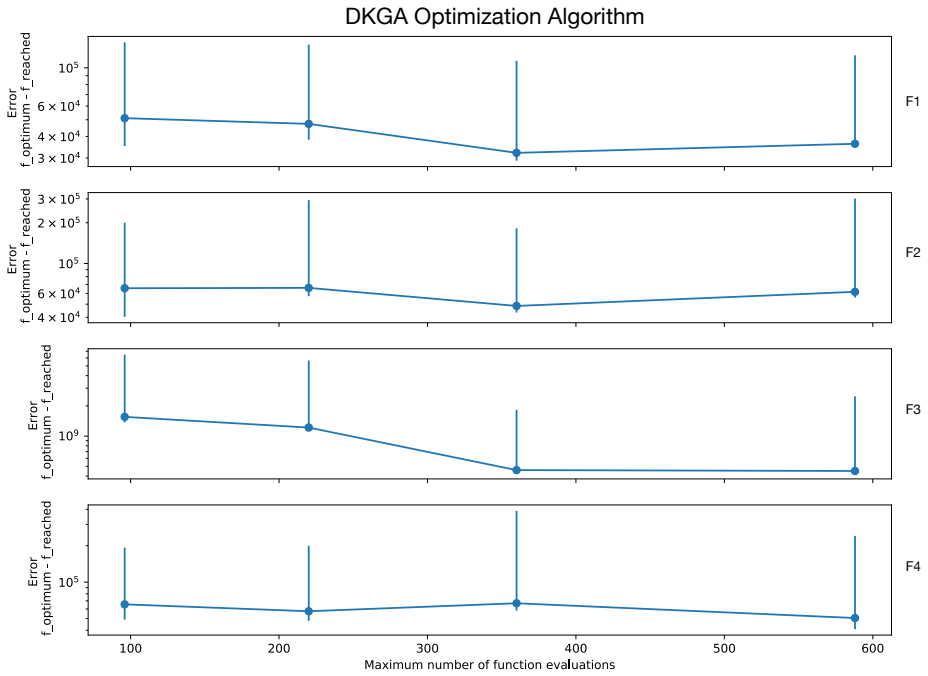


Figure 3: On the X axis, the number of iterations of the algorithm. On the Y axis, the difference between the optimum point reached by DKGA and the real optimum.

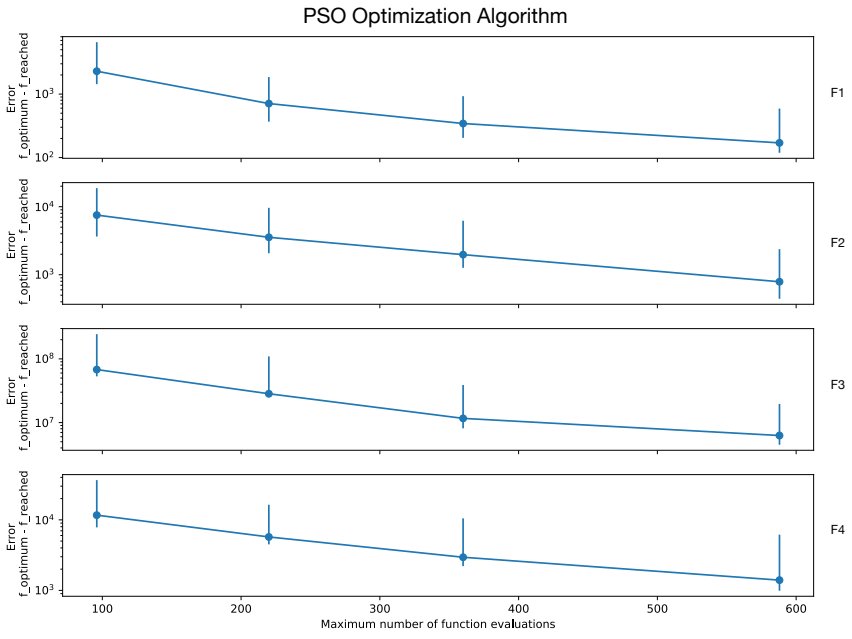


Figure 4: On the X axis, the number of iterations of the algorithm. On the Y axis, the difference between the optimum point reached by the algorithm and the real optimum.

References

[1] Klos D., "Investigations of low temperature combustion engine design and combustion stability", M.S. Thesis in Mechanical Engineering, University of Wisconsin-Madison, 2015.

[2] Benajes, J., Novella, R., Pastor, J.M., Hernández-López, A. and Kokjohn, S., Computational optimization of a combustion system for a stoichiometric DME fueled compression ignition engine, *Fuel*, Vol. 223, 2018.

[3] O'Neill, M. and Brabazon, A., "Grammatical Swarm, The generation of programs by social programming", *Natural Computing*, 5: 443-462, 2006.

[4] Marini, F. and Walczak, B., Particle swarm optimization (PSO). A tutorial, Chemometrics and Intelligent Laboratory Systems, Volume 149, Part B, 2015.

[5] Suganthan, P. N., Hansen, N., Liang, J. J., Deb, K., Chen, Y.-P., Auger, A. and Tiwari, S., "Problem Definitions and Evaluation Criteria for the CEC 2005 Special Session on Real-Parameter Optimization", Technical Report, Nanyang Technological University, Singapore, May 2005 AND KanGAL Report 2005005, IIT Kanpur, India.

27

The quantum brain model

Joan C. Micó [♭1], Antonio Caselles[♮], Salvador Amigó[♯], David Soler[♭] and Maria T. Sanz[*]

(♭) Institut Universitari de Matemàtica Multidisciplinar,
Universitat Politècnica de València,
(♮) IASCYS member (retired), Departament de Matemàtica Aplicada,
Universitat de València,
(♯) Departament de Personalitat, Avaluació i Tractaments Psicològics,
Universitat de València,
(∗) Departament de Didàctica de la Matemàtica,
Universitat de València.

1 Introduction

A macroscopic spatio-temporal model of the brain dynamics is presented. It is here called as the spatio-temporal response model (STRM), and also the *quantum brain model*, due to its eigenvalues and eigenfunctions quantization are quantized due due the boundary conditions. Thus, the quantized prediction of the spatio-temporal brain activity (STBA) from a known initial one is possible with this model. Its mathematical structure is a generalization of the temporal response model (TRM) that predicts the temporal brain activity (TBA) as a consequence of several stimuli [1]:

$$\left.\begin{array}{c} \frac{dy(t)}{dt} = \bar{a}(\bar{b} - y(t)) + \sum_i \bar{p}_i \cdot s_i(t) \cdot y(t) - \sum_i \bar{q}_i \cdot \int_{t_0}^{t} e^{\frac{x-t}{\bar{\tau}_i}} \cdot s_i(x) \cdot y(x) dx \\ y\left(t_0\right) = y_0 \end{array}\right\} \quad (1)$$

In (1), t is the time, and $y(t)$, $\bar{b}$ and y_0 are respectively the TBA, its tonic level and its initial value. The TBA is measured with the psychological variable called as General Factor of Personality (GFP) [1]. Besides, $s_i(t)$, $i = 1, 2, \ldots, n$, are the different stimuli, which can be of different natures: the amount of non-consumed drug by cells, a sound, a view, etc., which can hold different mathematical temporal functions. In addition, $\bar{a}(\bar{b} - y(t))$ is the *homeostatic control*, i.e., the cause of the fast recovering of the tonic level $\bar{b}$, being $\bar{a}$ the *homeostatic control power* of this control; $\bar{p}_i \cdot s_i(t) \cdot y(t)$ are the different *excitation effects*, which tend to increase the temporal brain activity, being $\bar{p}_i$ the *excitation effect powers*; $\bar{q}_i \cdot \int_0^t e^{\frac{x-t}{\bar{\tau}_i}} \cdot s_i(x) \cdot y(x) dx$ are the different *inhibitor effects*, which tend to decrease the temporal brain activity and are the cause of the its slow recovering, being $\bar{q}_i$ the *inhibitor effect powers* and being $\bar{\tau}_i$ the *inhibitor effect delays*.

[1]e-mail: jmico@mat.upv.es

2 The spatio-temporal response model or quantum brain

The STRM is obtained as a generalization of the TRM. To do this, consider in Eq. (1) that the TBA variable $y(t)$ must be substituted by a function that represents the STBA as a spatial-density depending on the time t and on the three spatial rectangular variables $\mathbf{r} = (x_1, x_2, x_3)$. Then, the time derivative in Eq. (1) must be a partial time derivative. The $\Psi(t, \mathbf{r})$ be the STBA variable, thus, the starting hypothesis is that:

$$y(t) = \left(\iiint_D \Psi^2(t, \mathbf{r}) d\mathbf{r} \right)^{1/2} = (\Psi(t, \mathbf{r}), \Psi(t, \mathbf{r}))^{1/2} \tag{2}$$

In (2), D is the integration domain that depends on the brain geometry considered, and "(,)" represents the inner product. In addition, the spatial dynamics in (1) is introduced as a diffusion term through a Laplacianne function of $\Psi(t, \mathbf{r})$:

$$\frac{\partial \Psi(t, \mathbf{r})}{\partial t} = a \cdot (\omega(\mathbf{r}) - \Psi(t, \mathbf{r})) + \sum_i p_i \cdot s_i(t) \cdot \Psi(t, \mathbf{r}) - \sum_i q_i \cdot \int_0^t \mathrm{e}^{\frac{x-t}{\tau_i}} \cdot s_i(x) \cdot \Psi(x, \mathbf{r}) dx + \sigma \nabla^2 \Psi(t, \mathbf{r}) \tag{3}$$

$$\Psi(t_0, \mathbf{r}) = \phi(\mathbf{r}) \tag{4}$$

Note in (3) that the tonic level $\bar{b}$ in (1) has been substituted by $\omega(\mathbf{r})$, i.e., a spatial function, unknown by the moment. In addition, σ is the diffusion coefficient, here considered positive-valued, while the other parameters are also positive-valued and conserve the same meanings than in (1), i.e., a instead $\bar{a}$, p_i instead $\bar{p_i}$, q_i instead $\bar{q_i}$ and τ_i instead $\bar{\tau_i}$. However, they are related in a way provided below. The initial condition Eq. (4) must be provided through the spatial distribution of brain activity in the instant $t = t_0$. In addition, the boundary conditions must be also provided, but they depend on the brain geometry considered. They are provided in Section 4 for an idealized box-brain geometry, which considers that the spatial flow through the brain walls cancels. Observe that the spatio-temporal response model provided by Eqs. (3) and (4) can be considered as a generalization of the cable model for a pulse translation on a neuron axon [2], from an only spatial direction to the three spatial dimensions of the brain.

3 Analytical solution of the spatio-temporal response model: the idealized box-brain

Eq. (3) is not separable due to it is a non-homogeneous equation as a consequence of the term $a \cdot \omega(\mathbf{r})$ (known as the non-homogeneous equation source). However, this problem can be overcome by the *method of eigenfunction expansions*. This method considers the solutions of the *associated homogeneous spatio-temporal response model* for a function $\Psi_h(t, \mathbf{r})$, which does not have the source:

$$\frac{\partial \Psi_h(t, \mathbf{r})}{\partial t} = -a \cdot \Psi_h(t, \mathbf{r}) + \sum_i p_i \cdot s_i(t) \cdot \Psi_h(t, \mathbf{r}) - \sum_i q_i \cdot \int_0^t \mathrm{e}^{\frac{x-t}{\tau_i}} \cdot s_i(x) \cdot \Psi_h(x, \mathbf{r}) dx + \sigma \cdot \nabla^2 \Psi_h(t, \mathbf{r}) \tag{5}$$

Then, Eq. (5) is so separable by a product:

$$\Psi_h(t, \mathbf{r}) = \rho(t) \cdot \Omega(\mathbf{r}) \tag{6}$$

whose substitution in (5) provides:

$$\frac{\rho'(t)}{\rho(t)} + a - \sum_i p_i \cdot s_i(t) + \frac{1}{\rho(t)} \sum_i q_i \int_0^t \mathrm{e}^{\frac{x-t}{\tau_i}} \cdot s_i(x) \cdot \rho(x) dx = \sigma \cdot \frac{\nabla^2 \Omega(r)}{\Omega(r)} \tag{7}$$

In order to Eq. (7) holds, both members of the equation must be a constant. Let λ be this constant. The temporal part of Eq. (7) does not play any role in the solution of the non-homogeneous Eq. (3). However, from the spatial part of (7):

$$\nabla^2 \Omega(\mathbf{r}) = \frac{\lambda}{\sigma} \Omega(\mathbf{r}) \tag{8}$$

Eq. (8) is the Helmoltz equation, which can be solved by separating variables for several coordinate systems, and it is fundamental in the solution of Eq. (3). The solution considered for Eq. (8) is the use of the rectangular coordinates for an idealized box-brain geometry of dimensions L_1 (length, from back to forebrain), L_2 (width, from side to side of brain) and L_3 (height, from down to up brain):

$$\mathbf{r} = (x_1, x_2, x_3) \in [0, L_1] \times [0, L_2] \times [0, L_3] \tag{9}$$

Thus, separating variables in Eq. (8) as: $\Omega(\mathbf{r}) = \Omega_1(x_1) \cdot \Omega_2(x_2) \cdot \Omega_3(x_3)$ and subsequently dividing by the product $\Omega_1(x_1) \cdot \Omega_2(x_2) \cdot \Omega_3(x_3)$:

$$\frac{1}{\Omega_1} \frac{d^2 \Omega_1}{dx_1^2} + \frac{1}{\Omega_2} \frac{d^2 \Omega_2}{dx_2^2} + \frac{1}{\Omega_3} \frac{d^2 \Omega_3}{dx_3^2} = \frac{\lambda}{\sigma} \tag{10}$$

In order to Eq. (10) holds, each member of the addition must be a constant. These constants must be negative-valued to obtain an oscillatory dynamics, thus let $-k_i^2$ be, $i = 1, 2, 3$, these constants:

$$\frac{1}{\Omega_i} \frac{d^2 \Omega_i}{dx_i^2} = -k_i^2, \qquad i = 1, 2, 3. \tag{11}$$

And from (10) and (11):

$$\lambda = -\sigma(k_1^2 + k_2^2 + k_3^2). \tag{12}$$

Also, from (11):

$$\Omega_i(x_i) = A_i \cos(k_i x_i) + B_i \sin(k_i x_i), \qquad i = 1, 2, 3. \tag{13}$$

being A_i and B_i ($i = 1, 2, 3$) arbitrary constants.
With the boundary conditions that the spatial flow through the brain walls cancels in Eq. (4):

$$\left.\frac{\partial \Psi_h(t, \mathbf{r})}{\partial x_i}\right|_{x_i = 0} = 0; \quad \left.\frac{\partial \Psi(t, \mathbf{r})}{\partial x_i}\right|_{x = L_i} = 0 \quad i = 1, 2, 3. \tag{14}$$

which provide, from Eq. (13):

$$B_i = 0; \quad \sin(k_i L_i) = 0 \rightarrow k_i L_i = n_i \pi \rightarrow k_i = \frac{\pi}{L_i} n_i; \quad n_i = 1, 2, \ldots, +\infty; \quad i = 1, 2, 3 \tag{15}$$

Eq. (15) represents the quantization of the eigenvalues of the associated homogeneous spatio-temporal response model, as function of three positive integers. Note that the integers are restricted to vary in the range $n_i = 1, 2, \ldots, +\infty$ ($i = 1, 2, 3$), thus, the constants k_i are as

well positive-valued. On a hand, $k_i = 0$ has not physical sense, and the integers varying in the range $n_i = -1, -2, \ldots, -\infty$ $(i = 1, 2, 3)$ will duplicate unnecessarily the solutions. In addition, the separating constant λ from Eq. (12) becomes quantized, which can be rewritten as $\lambda_{n_1 n_2 n_3}$ (from now onwards the expression $n_i = 1, 2, \ldots, +\infty$ is over understood):

$$\lambda_{n_1 n_2 n_3} = -\sigma\pi^2 \left(\left(\frac{n_1}{L_1}\right)^2 + \left(\frac{n_2}{L_2}\right)^2 + \left(\frac{n_3}{L_3}\right)^2 \right) \tag{16}$$

As a consequence of (16), the solution of Eq. (8) is a superposition of the eigenfunctions:

$$\bar{\Omega}_{n_1 n_2 n_3}(\boldsymbol{r}) = \prod_{i=1}^{3} \sin\left(\frac{\pi \cdot n_i}{L_i} x_i\right) \tag{17}$$

Note that the eigenfunction Eq. (17) define an orthogonal base, due to:

$$\begin{aligned} (\bar{\Omega}_{n_1 n_2 n_3}(\mathbf{r}), \bar{\Omega}_{m_1 m_2 m_3}(\mathbf{r})) &= \iiint_D \bar{\Omega}_{n_1 n_2 n_3}(\mathbf{r}) \cdot \bar{\Omega}_{m_1 m_2 m_3}(\mathbf{r}) d\mathbf{r} \\ &= \prod_{i=1}^{3} \int_0^{L_i} \sin(\frac{\pi n_i}{L_i} x_i) \cdot \sin(\frac{2\pi m_i}{L_i} x_i) dx_i \\ &= \frac{L_1 L_2 L_3}{8} \delta_{n_1 m_1} \delta_{n_2 m_2} \delta_{n_3 m_3}. \end{aligned} \tag{18}$$

Thus, the corresponding orthonormal base is given by the eigenfunctions:

$$\Omega_{n_1 n_2 n_3}(\mathbf{r}) = \frac{\bar{\Omega}_{n_1 n_2 n_3}(\mathbf{r})}{\left(\bar{\Omega}_{n_1 n_2 n_3}, \bar{\Omega}_{n_1 n_2 n_3}\right)^{1/2}} = \left(\frac{8}{L_1 \cdot L_2 \cdot L_3}\right)^{1/2} \prod_{i=1}^{3} \sin\left(\frac{\pi \cdot n_i}{L_i} x_i\right). \tag{19}$$

Such that, from Eqs. (8) and (19):

$$\nabla^2 \Omega_{n_1 n_2 n_3}(\mathbf{r}) = \frac{\lambda_{n_1 n_2 n_3}}{\sigma} \Omega_{n_1 n_2 n_3}(\mathbf{r}) \tag{20}$$

$$(\Omega_{n_1 n_2 n_3}(\mathbf{r}), \Omega_{m_1 m_2 m_3}(\mathbf{r})) = \delta_{n_1 m_1} \delta_{n_2 m_2} \delta_{n_3 m_3} \tag{21}$$

$$\Omega(\mathbf{r}) = \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \sum_{n_3=1}^{\infty} \Omega_{n_1 n_2 n_3}(\mathbf{r}) = \left(\frac{8}{L_1 \cdot L_2 \cdot L_3}\right)^{1/2} \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \sum_{n_3=1}^{\infty} \prod_{i=1}^{3} \sin\left(\frac{2\pi \cdot n_i}{L_i} x_i\right) \tag{22}$$

That is, $\frac{\lambda_{n_1 n_2 n_3}}{\sigma}$ are the eigenvalues of the operator ∇^2 with associated eigenfunctions $\Omega_{n_1 n_2 n_3}$. These eigenfunctions are fundamental to find the solutions of the non-homogeneous spatio-temporal response model given by Eq. (3) by the following expansions:

$$\begin{aligned} \Psi(t, \mathbf{r}) &= \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \sum_{n_3=1}^{\infty} \eta_{n_1 n_2 n_3}(t) \cdot \Omega_{n_1 n_2 n_3}(\mathbf{r}) \\ &= \left(\frac{8}{L_1 \cdot L_2 \cdot L_3}\right)^{1/2} \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \sum_{n_3=1}^{\infty} \eta_{n_1 n_2 n_3}(t) \cdot \prod_{i=1}^{3} \sin\left(\frac{\pi \cdot n_i}{L_i} x_i\right) \end{aligned} \tag{23}$$

$$\begin{aligned}\omega(\mathbf{r}) &= \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty} C_{n_1n_2n_3}\cdot\Omega_{n_1n_2n_3}(\mathbf{r}) \\ &= \left(\frac{8}{L_1\cdot L_2\cdot L_3}\right)^{1/2}\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty} C_{n_1n_2n_3}\cdot\prod_{i=1}^{3}\sin\left(\frac{\pi\cdot n_i}{L_i}x_i\right)\end{aligned} \tag{24}$$

Such that:

$$\eta_{n_1n_2n_3}(t) = (\Psi(t,\mathbf{r}),\Omega_{n_1n_2n_3}(\mathbf{r})) = \left(\frac{8}{L_1\cdot L_2\cdot L_3}\right)^{1/2}\prod_{i=1}^{3}\int_0^{L_i}\Psi(t,\mathbf{r})\cdot\sin\left(\frac{2\pi\cdot n_i}{L_i}x_i\right)dx_i \tag{25}$$

$$C_{n_1n_nn_3} = (\omega(\mathbf{r}),\Omega_{n_1n_2n_3}(\mathbf{r})) = \left(\frac{8}{L_1\cdot L_2\cdot L_3}\right)^{1/2}\prod_{i=1}^{3}\int_0^{L_i}\omega(\mathbf{r})\cdot\sin\left(\frac{2\pi\cdot n_i}{L_i}x_i\right)dx_i \tag{26}$$

In the beginning, the two sides of Eq. (3) are multiplied by $\Omega_{m_1m_2m_3}(\mathbf{r})$ and taken the inner product $(\,,)$:

$$\begin{aligned}\eta'_{n_1n_2n_3}(t) = &\left(-a+\lambda_{n_1n_2n_3}+\sum_i p_i\cdot s_i(t)\right)\eta_{n_1n_2n_3}(t) \\ &-\sum_i q_i\cdot\int_0^t \mathrm{e}^{\frac{x-t}{\tau_i}}\cdot s_i(x)\cdot\eta_{n_1n_2n_3}(x)dx + a\cdot C_{n_1n_2n_3}\end{aligned} \tag{27}$$

The initial conditions for the integro-differential Eq. (27) are given by Eq. (25) in $t=t_0$ and by Eq. (4):

$$\begin{aligned}\eta_{n_1n_2n_3}(t_0) &= (\Psi(t_0,\boldsymbol{r}),\Omega_{n_1n_2n_3}(\boldsymbol{r})) = (\phi(\mathbf{r}),\Omega_{n_1n_2n_3}(\mathbf{r})) \\ &= \left(\frac{8}{L_1\cdot L_2\cdot L_3}\right)^{1/2}\prod_{i=1}^{3}\int_0^{L_i}\phi(\mathbf{r})\cdot\sin\left(\frac{2\pi\cdot n_i}{L_i}x_i\right)dx_i\end{aligned} \tag{28}$$

In conclusions: the solution of the spatio-temporal response model given by Eqs. (3) and (4) is provided by the expansion Eq. (23), where $\eta_{n_1n_2n_3}(t)$ is given by Eq. (27), with initial conditions Eq. (28), and $C_{n_1n_2n_3}$ by Eq. (26). Note that the functions $\Omega_{n_1n_2n_3}(\mathbf{r})$ are given by Eq. (19), considering the geometric idealized case of a box-brain.

4 Steady solution of the spatio-temporal response model

The steady solution of the spatio-temporal response model $\Psi^{(s)}(\mathbf{r})$ is provided as $t\to+\infty$, which corresponds for the idealized case that no stimuli are influencing on brain, i.e., when $s_i(t)=0$, in Eq. (3):

$$\Psi^{(s)}(\mathbf{r}) = \lim_{t\to+\infty}\Psi(t,\mathbf{r}) = \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}\lim_{t\to+\infty}\eta_{n_1n_2n_3}(t)\cdot\Omega_{n_1n_2n_3}(\mathbf{r}) \tag{29}$$

In this case, when $s_i(t)=0$, Eq. (27) becomes:

$$\eta'_{n_1n_2n_3}(t) + (a-\lambda_{n_1n_2n_3})\,\eta_{n_1n_2n_3}(t) = a\cdot C_{n_1n_2n_3} \tag{30}$$

Eq. (30) has a simple analytical solution:

$$\begin{aligned}\eta_{n_1n_2n_3}(t) &= \mathrm{e}^{-(a-\lambda_{n_1n_2n_3})(t-t_0)}\left(\eta_{n_1n_2n_3}(t_0) + a\cdot k_{n_1n_2n_3}\int_{t_0}^{t}\mathrm{e}^{\left(a-\lambda_{n_1n_2n_3}\right)(x-t_0)}dx\right)\\ &= \eta_{n_1n_2n_3}(t_0)\,\mathrm{e}^{-\left(a-\lambda_{n_1n_2n_3}\right)(t-t_0)} + \frac{a\cdot k_{n_1n_2n_3}}{(a-\lambda_{n_1n_2n_3})}\left(1-\mathrm{e}^{-\left(a-\lambda_{n_1n_2n_3}\right)(t-t_0)}\right)\end{aligned} \tag{31}$$

Note in Eq. (31) that, from Eq. (16), $a-\lambda_{n_1n_2n_3}>0$ due to $\lambda_{n_1n_2n_3}<0$. In addition, it tends to the steady state $\eta^{(s)}_{n_1n_2n_3}(t)$ as $t\to+\infty$:

$$\eta^{(s)}_{n_1n_2n_3} = \frac{a\cdot C_{n_1n_2n_3}}{(a-\lambda_{n_1n_2n_3})}. \tag{32}$$

And from Eq. (29):

$$\begin{aligned}\Psi^{(s)}(\mathbf{r}) &= \lim_{t\to+\infty}\Psi(t,\mathbf{r}) = \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}\eta^{(s)}_{n_1n_2n_3}\cdot\Omega_{n_1n_2n_3}(\mathbf{r}) = a\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}\frac{C_{n_1n_2n_3}}{(a-\lambda_{n_1n_2n_3})}\Omega_{n_1n_2n_3}(\mathbf{r})\\ &= \left(\frac{8}{L_1\cdot L_2\cdot L_3}\right)^{1/2}\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}\frac{a\cdot C_{n_1n_2n_3}}{(a-\lambda_{n_1n_2n_3})}\prod_{i=1}^{3}\sin\left(\frac{\pi\cdot n_i}{L_i}x_i\right).\end{aligned} \tag{33}$$

5 Relationship between the spatio-temporal response model and the temporal response model

The relationship of the solutions of both models, taking into account Eqs. (2) and (23):

$$\begin{aligned}y^2(t) &= (\Psi(t,\mathbf{r}),\Psi(t,\mathbf{r})) = \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}\sum_{m_1=1}^{\infty}\sum_{m_2=1}^{\infty}\sum_{m_3=1}^{\infty}\left(\Omega_{n_1n_2n_3}(\mathbf{r})\right)\Omega_{m_1m_2m_3}(\boldsymbol{r}))\cdot\eta_{n_1n_2n_3}(t)\\ &\cdot\eta_{m_1m_2m_3}(t) = \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}\eta^2_{n_1n_2n_3}(t)\end{aligned} \tag{34}$$

That is:

$$y(t) = \left(\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}\eta^2_{n_1n_2n_3}(t)\right)^{1/2} \tag{35}$$

From Eq. (38) the hypothesis of the isolated existence of the time functions $y_{n_1n_2n_3}(t)$ can be state as the projection:

$$y_{n_1n_2n_3}(t) = \left(\Psi_{n_1n_2n_3}(t,\mathbf{r}),\Psi(t,\mathbf{r})\right)^{1/2} = \eta_{n_1n_2n_3}(t) \tag{36}$$

Such that:

$$\Psi_{n_1n_2n_3}(t,\mathbf{r}) = y_{n_1n_2n_3}(t)\cdot\Omega_{n_1n_2n_3}(\mathbf{r});\quad y(t) = \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}y_{n_1n_2n_3}(t) \tag{37}$$

Due to $\eta_{n_1n_2n_3}(t)$ hold the integro-differential Eq. (27), with initial conditions Eq. (28), the functions $y_{n_1n_2n_3}(t)$ hold the equations:

$$y'_{n_1n_2n_3}(t) = \left(-a + \lambda_{n_1n_2n_3} + \sum_i p_i \cdot s_i(t)\right) y_{n_1n_2n_3}(t) - \sum_i q_i \cdot \int_0^t e^{\frac{x-t}{\tau_i}} \cdot s_i(x) \cdot y_{n_1n_2n_3}(x)dx + a \cdot C_{n_1n_2n_3} \tag{38}$$

$$y_{n_1n_2n_3}(t_0) = \left(\frac{8}{L_1 \cdot L_2 \cdot L_3}\right)^{1/2} \prod_{i=1}^{3} \int_0^{L_i} \phi(\mathbf{r}) \cdot \sin\left(\frac{2\pi \cdot n_i}{L_i} x_i\right) dx_i \tag{39}$$

The corresponding temporal steady states of $y_{n_1n_2n_3}(t)$ are, from Eq. (32):

$$y^{(s)}_{n_1n_2n_3} = \frac{a \cdot C_{n_1n_2n_3}}{(a - \lambda_{n_1n_2n_3})} \tag{40}$$

In fact, reorganizing Eq. (38) to obtain the mathematical structure of Eq. (1):

$$\begin{aligned} y'_{n_1n_2n_3}(t) = (a - \lambda_{n_1n_2n_3}) \left(\frac{a \cdot C_{n_1n_2n_3}}{a - \lambda_{n_1n_2n_3}} - y_{n_1n_2n_3}(t)\right) + \sum_i p_i \cdot s_i(t) \cdot y_{n_1n_2n_3}(t) \\ - \sum_i q_i \cdot \int_0^t \mathrm{e}^{\frac{x-t}{\tau_i}} \cdot s_i(x) \cdot y_{n_1n_2n_3}(x)dx \end{aligned} \tag{41}$$

With initial conditions (28) in (41). Then, by comparing Eqs. (1) and (41), the following equivalences can be derived:

$$\bar{a} \to \bar{a}_{n_1n_2n_3} = a - \lambda_{n_1n_2n_3}; \quad \bar{b} \to \bar{b}_{n_1n_2n_3} = \frac{a \cdot C_{n_1n_2n_3}}{a - \lambda_{n_1n_2n_3}}; \quad \bar{p}_i = p_i; \quad \bar{q}_i = q_i; \quad \bar{\tau}_i = \tau_i \tag{42}$$

In Eq. (42) the parameter values $\bar{a}_{n_1n_2n_3}$ and $\bar{b}_{n_1n_2n_3}$ are quantized, such that, by Eq. (16), $\lambda_{n_1n_2n_3} = -\sigma\pi^2 \left(\left(\frac{n_1}{L_1}\right)^2 + \left(\frac{n_2}{L_2}\right)^2 + \left(\frac{n_3}{L_3}\right)^2\right)$; $n = 1, 2, \ldots, +\infty$. Particularly, note that the quantized $\bar{b}_{\lambda_{n_1n_2n_3}}$ parameter values coincide with the values provided by Eq. (32), such as it must be held by the theory coherence. Note that only these parameters, $\bar{a}_{n_1n_2n_3}$ and $\bar{b}_{n_1n_2n_3}$, which represent biological properties of the brain, are quantized, but not those that are related with the stimuli dynamics, such that $\bar{p}_i$, $\bar{q}_i$ and $\bar{\tau}_i$.

6 Calibration of the STRM

There are several ways to observe experimentally the STBA. One of the most important ways is Neuroimage, which has had to develop the brain mappings, by using the Talairach and MNI coordinates [3] to measure the STBA by measuring the change of some important biological indicators in the brain, such as oxygen, blood, etc. In fact, one of the crucial aims of the Neuroimage technic is the study of the brain resting state [4], which can be identified with the steady state $\Psi^{(s)}(\mathbf{r})$ of Eq. (33). The information that Neuroimage would provide about the mathematical structure of the brain resting state would be a first way to validate the quantum brain model presented. However, in general, to validate the STRM the Neuroimage technic needs the $\phi(\mathbf{r})$ function knowledge, in order to obtain the initial conditions $\eta_{n_1n_2n_3}(t_0)$ through Eq. (28). And the same problems happen with the EEG (electroencephalogram) technic [5],

which measures the STBA by the electrical potential.

However, a first result can be provided for the STBA by its relationship with the TBA through the identities Eq. (42). Due to the $\phi(\mathbf{r})$ is unknown, the calibration of Eq. (41) is done with the initial condition of the experimental design presented in the beginning.

One subject consumed 10 mg of methylphenidate, and the GFP was observed every 7.5 minutes during 180 minutes (3 hours), with the 5 adjectives scale, GFP-FAS [6], in the interval [0,50]. The initial condition was also observed before consumption, with value y_0, which is considered as initial condition of Eq. (41) instead the unknown one of Eq. (28). Assuming that no methylphenidate is present in the organism, the temporal function of the methylphenidate [1] is given by:

$$s(t) = \begin{cases} \frac{\alpha \cdot M}{\beta - \alpha} \left(e^{-\alpha \cdot t} - e^{-\beta \cdot t} \right) : \alpha \neq \beta \\ \alpha \cdot M \cdot t \cdot e^{-\alpha \cdot t} : \alpha = \beta \end{cases} \tag{43}$$

The calibration of Eq. (41) by generating random numbers is provided graphically in Figs. 1 and 2.

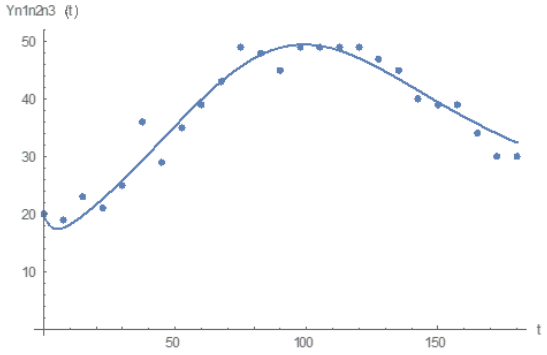


Figure 1: GFP response, $y_{n_1 n_2 n_3}(t)$, to the 10 mg of MF versus time. Experimental values (dots) and theoretical values (line). R2=0.94.

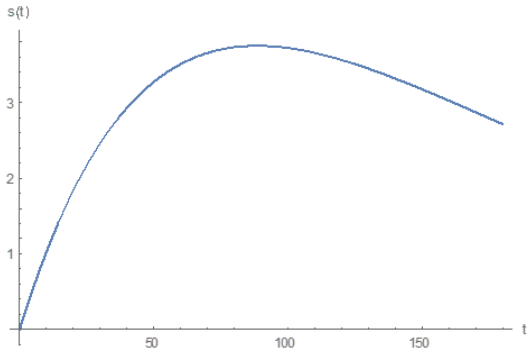


Figure 2: MF evolution s(t) of Eq. 44 in the organism.

The results of the calibration provide the following parameter values: $\alpha = 0.011510165332$; $\beta = 0.011069991532$; $a = 0.312844518371$; $C_{n_1 n_2 n_3} = 15.580863173952$; $p = 0.057535289406$; $q = 0.000000125055$; $\tau = 0.035782907172$; $M = 10.0$; $\sigma = 0.000432176762$; $n1 = 1$; $n2 = 1$; $n3 = 1$. These values permit to obtain the results of the corresponding eigenfunction $\Psi_{n_1 n_2 n_3}(t, \mathbf{r})(n1 = 1; n2 = 1; n3 = 1)$ of Eq. (37). Two graphical representations are provided in Figs. 3 and 4:

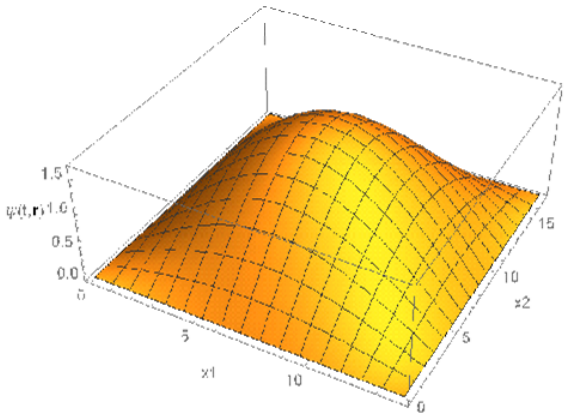


Figure 3: STBA result for $\Psi_{n_1 n_2 n_3}(t, \mathbf{r})$ with n1=1; n2=1; n3=1: t=88.5846 (instant of maximum TBA) for x3=L3/2.

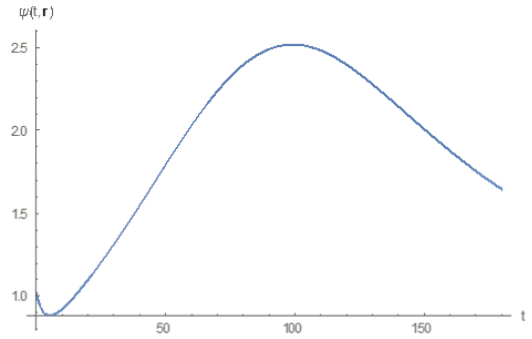


Figure 4: STBA result for $\Psi_{n_1 n_2 n_3}(t, \mathbf{r})$ with n1=1; n2=1; n3=1: $t \in [t_0, T]$ for $\mathbf{r} = (L_1/2, L_2/2, L_3/2)$.

References

[1] Micó, J.C., Amigó, S. and Caselles, A., From the Big Five to the General Factor of Personality: a Dynamic Approach, *Span. J. Psychol.*, 17 E74: 1-18, 2014.

[2] Scott, A., *Neuroscience. A Mathematical Primer*, Springer-Verlag, 2002.

[3] Laird, A. R. et al., Comparison of the disparity between Talairach and MNI coordinates in functional neuroimaging data: Validation of the Lancaster transform, *NeuroImage*, 51: 677–683, 2010.

[4] Smith, S. M. et al., Correspondence of the brain's functional architecture during activation and rest, *PNAS*, 106(31): 13040–13045, 2009.

[5] Goodfellow, M., *Spatio-temporal modelling and analysis of epileptiform EEG*, University of Manchester, 2011.

[6] Amigó, S., Micó, J.C. and Caselles, A., Five adjectives to explain the whole personality: a brief scale of personality, *Rev. Int. Sist.*, 16: 41–43, 2009.

28

Probabilistic solution of a randomized first order differential equation with discrete delay

Tomás Caraballo[♭], J.-C. Cortés[♮] and A. Navarro-Quiles [♮] [1]

(♭) Dpto. Ecuaciones Diferenciales y Análisis Numérico,
Universidad de Sevilla,
(♮) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

At the end of the seventeenth century the first attempts to solve physical problems were made through differential calculus. This situation gradually led to the creation of a new branch of mathematics, namely, differential equations. In the mid-eighteenth century differential equations became an independent branch and its resolution an end in itself. Therefore, since centuries, it has been demonstrated that differential equations are key tools for modeling different interesting problems in many areas. Being these models fundamental to understand the world around us. In many applications of differential equations, it is considered that the future state is independent of its past and it only depends on the current information. However, there exist some situations where it is more realistic to model the problem taking into account certain information from the past. A differential equation which incorporate the past in its formulation is called delayed differential equation.

On the other hand, the parameters that appear in this kind of formulations are generally fixed from experimental data. Therefore, since these values are obtained from certain measurements and samplings, they contain an intrinsic uncertainty. This situation allows us to consider the inputs as random variables or stochastic processes rather than constants or deterministic functions. Taking into account the randomness in the formulation, the solution of the initial value problem considered is a stochastic process. Then, as difference with the classical literature where the main objective is the computation of the exact solution, in this case we can calculate interesting probabilistic information. The main objective of this work is to obtain an expression of the first probability density function, $f(x,t)$, of the solution stochastic process. This function give us a full probabilistic description of the solution in every time instant t. In addition, from it we can calculate other interesting information, like the symmetry, kurtosis, as well as the mean and the variance

$$\mathbb{E}[X(t)] = \int_{\mathbb{R}} x f(x,t)\mathrm{d}x, \qquad \mathbb{V}[X(t)] = \int_{\mathbb{R}} x^2 f(x,t)\mathrm{d}x - \mathbb{E}[X(t)]^2. \tag{1}$$

[1] e-mail: annaqui@doctor.upv.es

From these probabilistic information we can compute confidence intervals, a fundamental tool to obtain predictions. In particular, we consider in this contribution the following situation.

As it can be observed in the existent literature, delayed differential equations can be classified in different classes depending on the type of delay involved in its formulation. In this work, we consider the simplest model in order to introduce ourself to the probabilistic study of randomized delayed differential equations. We study, from a probabilistic point of view, the linear differential equation with discrete delay, $\tau > 0$,

$$\left\{ \begin{array}{rcl} x'_\tau(t;\omega) & = & a(\omega)x_\tau(t;\omega) + b(\omega)x_\tau(t-\tau;\omega), \ \ t>0, \ \tau>0, \\ x_\tau(t;\omega) & = & g(t;\omega), \quad -\tau \leq t \leq 0, \end{array} \right. , \tag{2}$$

where the coefficients $a(\omega)$ and $b(\omega)$ are assumed to be continuous random variables and the initial condition $g(t;\omega)$ is a stochastic process defined on the interval $[-\tau, 0]$. We consider the initial value problem (2) since it has an unique solution which can be easily calculated, see Reference [1],

$$\begin{aligned} x_\tau(t;\omega) &= \mathrm{e}^{a(\omega)(t+\tau)}\, \mathrm{e}_\tau^{b_1(\omega),\, t}\, g(-\tau;\omega) \\ &\quad + \int_{-\tau}^{0} \mathrm{e}^{a(\omega)(t-s)}\, \mathrm{e}_\tau^{b_1(\omega),\, t-\tau-s}\, \left(g'(s;\omega) - a(\omega)g(s;\omega)\right) \mathrm{d}s, \end{aligned}$$

where $b_1(\omega) = \mathrm{e}^{-a(\omega)\tau}\, b(\omega)$.

2 Probabilistic Solution

To solve this problem we apply the Random Variable Transformation method (Reference [4]) to obtain the first probability density function of the solution stochastic process. To obtain the expression of the density function we must to consider particular expressions to the initial condition stochastic process, $g(t;\omega)$, in the initial value problem (2). Choosing, for example, $g(t;\omega) = \mathrm{e}^{a(\omega)t + c(\omega)}$, the solution is

$$x_\tau(t;\omega) = \mathrm{e}^{a(\omega)t + c(\omega)}\, \mathrm{e}_\tau^{b_1(\omega),\, t},$$

where $b_1(\omega) = \mathrm{e}^{-a(\omega)\tau}\, b(\omega)$. Then, considering the following bijective transformation, $r : \mathbb{R}^3 \to \mathbb{R}^3$

$$x_1 = r_1(c,a,b) = \mathrm{e}^{a(\omega)t + c(\omega)}\, \mathrm{e}_\tau^{b_1(\omega)}, \qquad x_2 = r_2(c,a,b) = a, \qquad x_3 = r_3(c,a,b) = b.$$

we apply the Random Variable Transformation method, obtaining

$$\boxed{f(x,t;\tau) = \int_{\mathbb{R}^2} f_{c,a,b}\left(\lambda^n\left(\frac{x\,\mathrm{e}^{-at}}{\mathrm{e}_\tau^{b_1,\,t}}\right), a, b\right) \frac{1}{|x|}\, \mathrm{d}a\, \mathrm{d}b, \ \text{where } b_1 = \mathrm{e}^{-a\tau}\, b.} \tag{3}$$

Finally, given the first probability density function of the corresponding problem with zero delay, $f(x,t)$, it can be proved that under some conditions

$$\lim_{\tau \to 0^+} f(x,t;\tau) = f(x,t), \quad \text{for each } (x,t) \in \mathcal{D} \times [(n-1)\tau, n\tau[\ \text{ fixed}, \tag{4}$$

being $\mathcal{D} = \mathcal{D}(x_\tau(t;\omega)) \cap \mathcal{D}(x(t;\omega))$, where $\mathcal{D}(x_\tau(t;\omega))$ and $\mathcal{D}(x(t;\omega))$ denote the codomains of SPs $x_\tau(t;\omega)$ and $x(t;\omega)$, respectively.

3 Numerical example

In this example, we consider that $a(\omega)$, $b(\omega)$ and $c(\omega)$ are independent random variables with the following distributions:

- $a(\omega)$ is a Gaussian distribution with 0 mean and 0.1 standard deviation, i.e., $a(\omega) \sim \mathrm{N}(0; 0.1)$.
- $b(\omega)$ follows a Beta distribution with positive parameters 2 and 3, i.e., $b(\omega) \sim \mathrm{Be}(2; 3)$.
- $c(\omega)$ follows an Exponential distribution with mean $1/20$, i.e., $c(\omega) \sim \mathrm{Exp}(20)$.

To show how $f(x, t; \tau)$ converges to $f(x, t)$ when τ tends to zero, in Figure (1) we have plotted $f(x, t)$ together $f(x, t; \tau)$ with different delays $\tau \in \{0.01, 0.05, 0.1, 0.5, 2\}$ at different time instants $t = 0.1$ and $t = 1$. We observe that the first probability function converges when τ vanishes. This situation can be also observed in Table 1 where the following error is computed

$$e_\tau^{\mathrm{PDF}}(t) = \int_{\mathbb{R}} |f(x, t; \tau) - f(x, t)| \, \mathrm{d}x. \tag{5}$$

As one would expect, we observe that for t fixed, the error $e_\tau^{\mathrm{PDF}}(t)$ decreases as $\tau \to 0^+$.

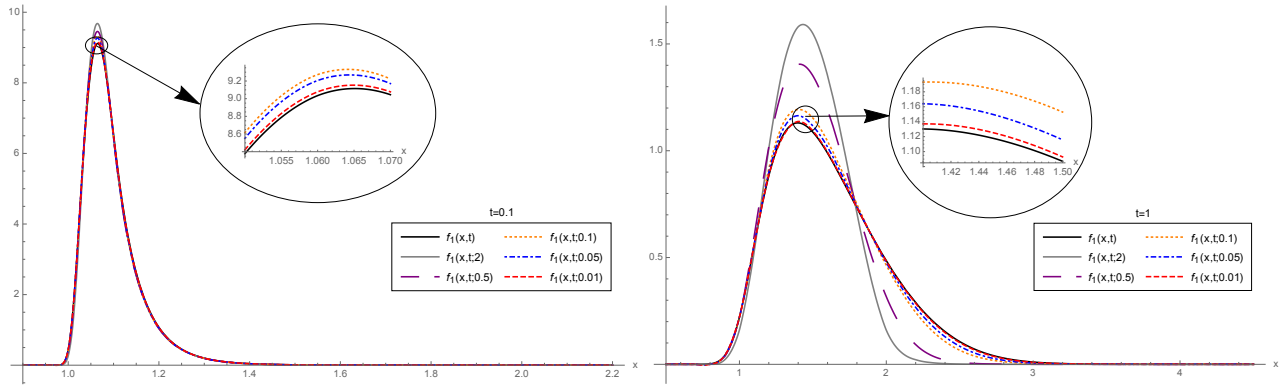


Figure 1: Graphical representation of the PDFs, $f(x, t)$ and $f(x, t; \tau)$, with different delays $\tau \in \{0.01, 0.05, 0.1, 0.5, 2\}$, at different time instants: $t = 0.1$ (left) and $t = 1$ (right).

$e_\tau^{\mathrm{PDF}}(t)$	$\tau = 2$	$\tau = 0.5$	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.01$
$t = 0.10$	0.040793	0.027599	0.021251	0.015753	0.003949
$t = 1.00$	0.382194	0.252025	0.069224	0.035766	0.007786

Table 1: Error measure $e_\tau^{\mathrm{PDF}}(t)$, defined by (5), with different delays $\tau \in \{0.01, 0.05, 0.1, 0.5, 2\}$, at different time instants, $t = 0.1$ and $t = 1$.

Finally, we use formulas (1) to calculate the mean and the variance of $x_\tau(t; \omega)$ and $x(t; \omega)$ with different delays $\tau \in \{0.01, 0.05, 0.1, 0.5, 2\}$ at the time interval $[0, 1]$. We can also observe in

Fig. 2 the converge as $\tau \to 0^+$. For sake of clarity, we have compute in Table 2 the following errors

$$e_\tau^{\mathbb{E}} = \int_0^T |\mathbb{E}\left[x_\tau(t;\omega)\right] - \mathbb{E}\left[x(t;\omega)\right]| \, \mathrm{d}t, \qquad e_\tau^{\mathbb{V}} = \int_0^T |\mathbb{V}\left[x_\tau(t;\omega)\right] - \mathbb{V}\left[x(t;\omega)\right]| \, \mathrm{d}t. \tag{6}$$

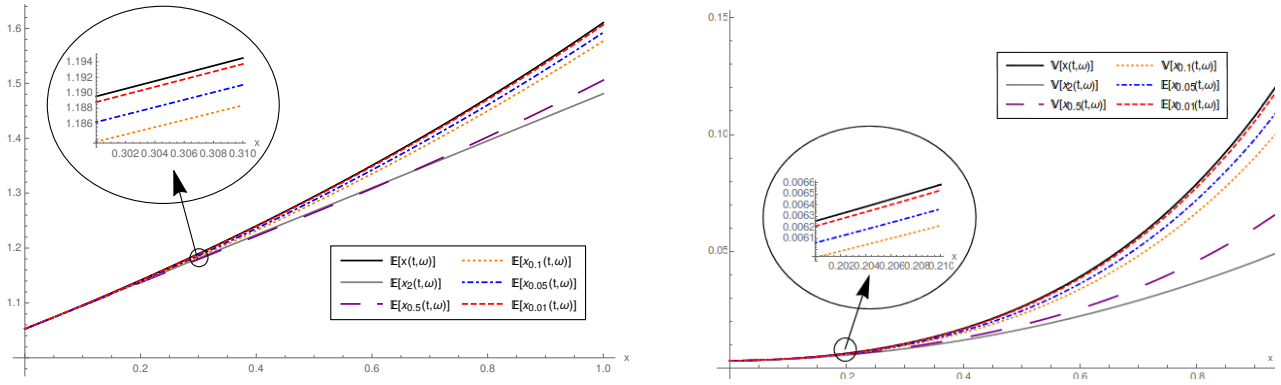


Figure 2: Graphical representation of the mean (left) and the variance (right) of the solutions SP $x(t;\omega)$ and $x_\tau(t;\omega)$, with different delays $\tau \in \{0.01, 0.05, 0.1, 0.5, 2\}$, at the time interval $[0, 0.5]$.

	$\tau = 2$	$\tau = 0.5$	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.01$
$e_\tau^{\mathbb{E}}$	0.039532	0.036982	0.013269	0.007252	0.001562
$e_\tau^{\mathbb{V}}$	0.02085	0.016288	0.006171	0.003446	0.000759

Table 2: Error measures $e_\tau^{\mathbb{E}}$ and $e_\tau^{\mathbb{V}}$, defined by (6), with different delays $\tau \in \{0.01, 0.05, 0.1, 0.5, 2\}$, taking the final time instant $T = 1$.

Acknowledgements

This work has been partially supported by the Ministerio de Economía y Competitividad grant MTM2017–89664–P. Ana Navarro Quiles acknowledges the funding received from Generalitat Valenciana through a postdoctoral contract (APOSTD/2019/128).

References

[1] Khusainov, D. Y. and Ivanov, A. F. and Kovarzh, I. V. Solution of one heat equation with delay, *Nonlinear Ocillations*, 12: 260–282,2009.

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

29

A predictive method for bridge health monitoring under operational conditions

Ana Sancho[♭], Fran Ribes-Llario[♭1], Adrián Zornoza[♭] and Teresa Real[♭]

(♭) Institute of Multidisciplinary Mathematics,
Universitat Politécnica de València.

1 Introduction

Approximately, the average number of bridge collapses per year is 1/4700 [1]. One of the most recent examples took place on August 14th 2018. The Morandi Bridge in Genoa collapsed killing 41 people and causing an economic damage that will take years to be repaired. Even Morandi Bridge was known to be in trouble long before collapse [2], experts affirmed that the collapse of the bridge was unexpected and sudden with respect to the monitoring that the bridge was subjected to [3]. In addition, common means and methods used for bridge health monitoring are intrusive and difficult the normal exploitation of the infrastructure – traffic deviations, incidences during maintenance operations, collisions, etc. -.

For this reason, a semi-empirical method based on strategical parameters and its measurement locations is exposed herein. In this method, the combination of numerical modelling with data registered in specific points of the structure allows to characterize the response of the structure under operational conditions in real time.

This strategy entails driving maintenance of fixed infrastructure assets from an observe and react approach failure towards a *predict and prevent* strategy [4].

2 Method

To achieve this goal, the purposed methodology is divided in following steps:

i) Numerical modelling. First of all, a **3D numerical representation of the structure** is carried out in a Finite Element Model software in accordance with 'as-built' information and maintenance history.

ii) Preliminary simulations. Instrumentation plan based on the identification of **critical parameters** depending on the structure typology, static and dynamic behaviour obtained

[1]e-mail: frarilla@cam.upv.es

from initial simulations performed on the 3D FEM model and **maximum response points**.

iii) Bridge instrumentation. Usual sensors required are: accelerometers for vibration analysis, strain gauges for stress / strain measurement, displacement sensors for relative movements, weather stations, cameras, inclinometers for element rotations and weather stations for thermal, humidity and wind intensity control.

iv) Calibration of the model. The static and dynamic response of the structure is adjusted by using load test data – static and dynamic tests' results – and measurements registered trough sensors installed.
Structural stiffness $[K_s]$ and boundary conditions are usually adjusted by using static load test results – deflections and displacements – while structural mass $[M]$ and support stiffness $[K_g]$ are fitted by a modal analysis and the dynamic load test results.

v) Structural damage simulations. Collapse and usability loss of the structure is determined by the simulation of several extreme loading scenarios. Results obtained in this stage allow to define admissible limits and tolerances to real time measured parameters.

vi) Prediction of the remaining lifespan and decision making assessment. By using degradation models – i.e. fatigue for metallic elements, rheological formulations for concrete, stiffness loss due to weather conditions – and real time measurements it is possible to obtain an updated prediction of the lifespan of the most relevant elements of the structure. The automation of this process combined with an alert system has strong potential for infrastructure managers.

In this paper, most relevant aspects about how steps i) to v) are applied to both bridges located in Chile is explained. And the step vi) is widely discussed.

Figure 1: Tolten bridge (left) and Seminario bridge (right).

3 Results

Steps i) and ii). A numerical 3D model with beam – for bar and truss modelling - and shell elements – for slabs and Tolten's embedded support walls – was carried out. Terrain was modelled as a set of springs $[K_g]$ attached to pier nodes. Self-weight plus dead loads were only

considered from preliminary static and modal analysis.

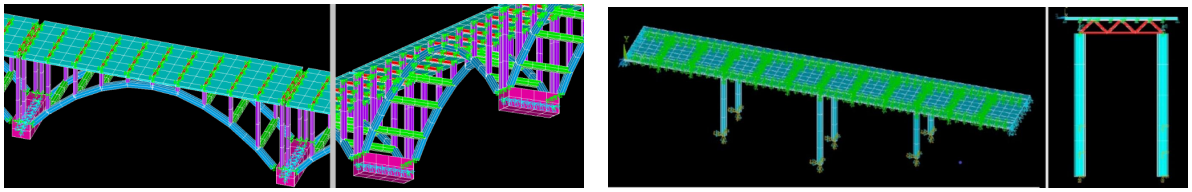


Figure 2: Tolten bridge numerical model (left) and Seminario bridge numerical model (right).

Step iii). By using accelerometers and inclinometers in intermediate span length points, inclinometers and scour sensors in piers and displacement sensors in joints it was possible to control the geometrical distortion of the structure in real time and under normal operation conditions.

Step iv). Static calibration was performed by locating a 20 ton truck in $\frac{1}{2}$ length of each span. Deflection deviation obtained after calibration between measured data and model prediction was lower than 5% in both bridges. Regarding dynamic calibration, it was carried out by achieving the first two vertical bending modes deviation did not exceed simultaneously a 0.5 Hz threshold. Accelerations registered with installed sensors were used for this purpose by applying Modal Operational Analysis technique. First two deck bending modes were used to calibrate the dynamic behaviour of the model in both structures. Results obtained, in terms of frequency deviation, after the calibration procedure were:

	Seminario Bridge		**Tolten Bridge**	
	1^{st} mode freq. $[Hz]$	2^{nd} mode freq. $[Hz]$	1^{st} mode freq. $[Hz]$	2^{nd} mode freq. $[Hz]$
Measured	4.458	5.223	1.728	10.096
Calculated	4.238	5.177	1.673	10.069
Deviation	0.220	0.046	0.055	0.027

Table 1: Results of dynamic calibration in both bridges using first two deck bending modes.

Step v). Ten load combinations were considered by taking into account maximum traffic in both lanes - Traffic 1 and Traffic 2 -, AASHTO lateral wind forces, maximum admissible scour and self-weight actions. Cross sectional failure of each structural element was considered as the limiting criteria. To characterize it, the following overstress ratio (OR%) is used:

$$\boldsymbol{OR}\% = \frac{\sigma_{vm}}{\sigma_{adm}}$$

where σ_{vm} and σ_{adm} are the maximum Von Mises stress obtained by the FEM analysis and the maximum strength of the corresponding material.

		Slab	Pier	Beam	Truss
Current situation					
	OR%.	13,57%	8,13%	44,23%	43,79%
No scour					
Traffic 1	**OR%.**	24,52%	12,33%	72,66%	70,23%
Traffic 2	**OR%.**	31,41%	16,24%	97,71%	99,31%
Wind	**OR%.**	15,27%	19,84%	47,19%	44,14%
Earthquake (long Axis)	**OR%.**	13,20%	63,84%	43,27%	77,50%
Earthquake (trans Axis)	**OR%.**	35,69%	49,93%	69,20%	39,46%

Slab	Pier	Beam	Truss
Max scour			
25,84%	12,57%	72,73%	70,28%
31,76%	16,38%	97,53%	99,46%
15,86%	21,81%	47,44%	44,15%
14,00%	69,26%	43,29%	84,85%
35,83%	93,78%	69,21%	46,93%

		Slab	Pier	Diagonal tube	Trans Beam	Diag Beam	Girder	Truss
Current situation								
	OR%.	43%	12%	7%	18%	5%	22%	5%
No scour								
Traffic 1	**OR%.**	105%	12%	7%	30%	5%	22%	5%
Traffic 2	**OR%.**	102%	30%	16%	44%	6%	59%	15%
Wind	**OR%.**	43%	12%	17%	18%	6%	23%	5%
Earthquake (long Axis)	**OR%.**	9%	89%	36%	7%	25%	8%	4%
Earthquake (trans Axis)	**OR%.**	60%	87%	46%	33%	8%	18%	9%
Max scour								
Traffic 1	**OR%.**	104%	20%	11%	30%	5%	60%	10%
Traffic 2	**OR%.**	102%	30%	17%	44%	10%	59%	15%
Wind	**OR%.**	35%	12%	7%	16%	6%	21%	5%
Earthquake (long Axis)	**OR%.**	8%	81%	39%	5%	11%	8%	4%
Earthquake (trans Axis)	**OR%.**	50%	80%	44%	26%	19%	14%	9%

Table 2: Maximum Von Mises stress and OR% ratio obtained for Seminario bridge (above) and Tolten (below) simulations.

Step vi). Finally, the individual parameter evolution trends were obtained by minimizing deviation with measured data. This was carried out considering three months of continued hourly registers per each parameter. In addition, an Artificial Neural Network was also implemented. The objective of this ANN is to automate the detection of trend patterns in monitored parameters. From previous experiences, it is expected that this procedure, when becomes fully developed, allows to achieve less than a 9% of deviation in pathologies identification. For each bridge, following ANN parameters are defined:

	Seminario	**Tolten**
Inputs	**26** [Numerical results of 13 health parameters x 2 monitoring points]	**39** [Numerical results of 13 health parameters x 3 monitoring points]
Outputs	**36** [Potential pathologies]	**58** [Potential pathologies]
Training	**108** FEM simulations	**174** FEM simulations
Verification	>**2200** registers / health parameter	>**2200** registers / health parameter

Table 3: ANN parameters for Seminario bridge (left) and Tolten bridge (right).

4 Conclusions

In this paper, a non-intrusive method for continuous bridge health monitoring has been defined. In addition, an ANN has been implemented according with real field data measurements and numerical simulation results. Conclusions obtained are mentioned as follows:

- Combination of static and dynamic calibration allows to obtain high fidelity models. This procedure has been successfully applied to two different bridges in Chile: Tolten and Seminario.

- Seminario, which is a modern structure, shown a great performance in any analysed scenario. However, Tolten bridge, which was built in 1910's, shown its integrity damaged under severe seismic actions.

- Predictive techniques based on the study of individual evolution of each monitored parameter could not be adequate in structures with high variability in its registers. For this reason, AI techniques as ANN combined with real data collection and numerical modelling could improve substantially the assistance in infrastructure management in real time.

References

[1] Cook, W., *Bridge Failure Rates, Consequences and Predictive Trends*. Utah State University. Logan, USA, 2014.

[2] Piangiani, G., *Italy Bridge Was Known to Be in Trouble Long Before Collapse*. The New York Times. Genoa, Italy, 2018.

[3] Piangiani, G., *Italy Bridge Collapse Leaves 37 Dead*. The New York Times. Genoa, Italy, 2018.

[4] Saadé, L., *Enhanced points condition monitoring on Network Rail Infrastructure*. Network Rail. University of Birmingham. Birmingam, UK, 2016.

30

Comparison of a new maximum power point tracking based on neural network with conventional methodologies

Ernesto A. Colomer Rosell[♭], Laura Andrés López[♭], Juan Ramón Sánchez Vicedo[♭] and Jorge del Pozo[♮]

(♭) Institute of Multidisciplinary Mathematics,
Universitat Politécnica de València,
(♮) Universidad Europea de Madrid.

1 Introduction

Over the last two decades, solar energy has undergone an extraordinary evolution. becoming in the last years one of the main sources of renewable electricity generation. As far as photovoltaic solar energy is concerned, at the end of 2015, about 230 GW of photovoltaic power were installed around the world, making it the third most important renewable energy source in terms of installed global level capacity (after hydro and wind energy) [1]. Also, based on the forecasts made in 2018 by the association SolarPower Europe [2], the global demand for solar photovoltaic energy will be increased year after year. Considering a medium scenario, it is expected that total global installed photovoltaic generation capacity will pass 1.1 TW in 2022.

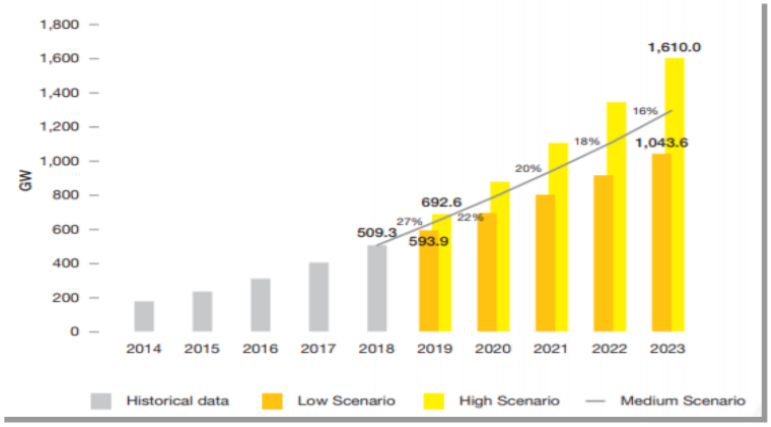


Figure 1: Evolution of installed solar power capacity [2].

One of the main problems of photovoltaic energy is that the production depends on several

e-mail: frarilla@cam.upv.es

factors. Like solar radiation, with daily and annual variations, panel temperature and operating point. For that reason, it is so important to develop algorithms that allow the maximum power point tracking.

2 Conventional method

Traditionally, the use of irradiation and temperature has been tried to avoid from the MPP algorithms due to the difficulty of measuring solar irradiation (the high cost of sensors: Pyranometers and Pyrheliometer and its difficult calibration).

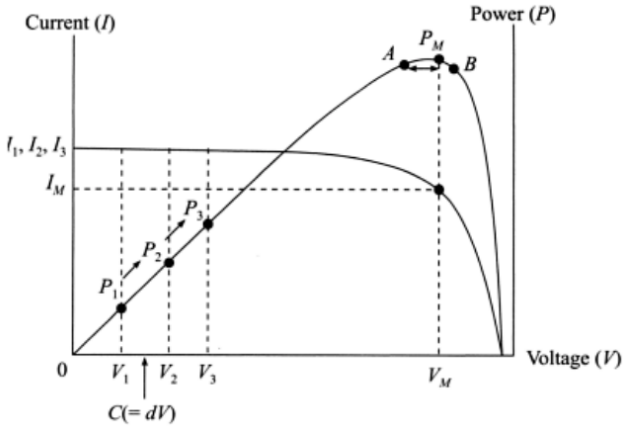


Figure 2: Algorithm of Perturbe and Observe. [3]

The most used methodology in these systems is based on the application of an algorithm known as *Perturbe and Observe* (P&O), which consists in altering the work point, increasing or decreasing the working voltage, and check whether the power produced has been increased or decreased. Depending on the response, the work point is still varying in the same direction or change in the opposite direction.

This algorithm causes important oscillations around the maximum power point. Besides, partial shadows can generate local maximums, which avoid the correct operation of tracker P&O.

In view of the existing problem, which affects all photovoltaic installations, and the problems of the current systems of maximum power point tracking (MPPT's), a viable alternative is required. This document presents a new methodology based on the measurement of irradiance (indirectly) [4] and temperature of the panel, using two low-cost sensors and the use of neural networks, which allow to reach the maximum power point quicker and more efficiently.

3 New proposed method

It can be differentiated three stages. The first consists of data collection phase for neural networks training. After the sensors installation, it is necessary to collect data for at least one year to cover all the seasons of the year. Note that a recurrent neural network has been used

so the algorithm remembers previous outputs to use as inputs.

After this first stage, the algorithm will be validated by means of verification that generated power is greater than the generated with traditional methods. It is important to mention that this stage will be carried out only until validating the proposed new method.

After this validation it will be possible to carry out the definitive installation. Here, from short-circuit current and temperature of the panel measurement, the maximum power point can be obtained.

To carry out the method it is necessary to have two hardware types, that is, one for training: more expensive but reusable for all installations, and hardware that will be installed for continuous operation: very low cost circuit.

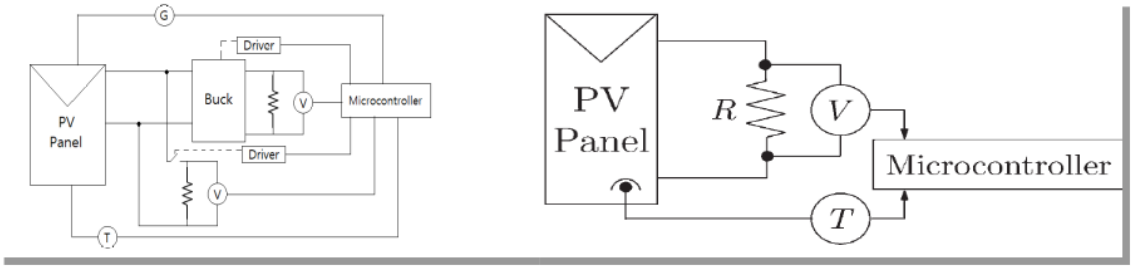


Figure 3: Training hardware (left) and permanent hardware (right).

On the one hand, the hardware that will be installed permanently is formed by a temperature sensor and a MOSFET that is used to measure the short circuit current. In this way with neural networks the maximum power point is obtained. The processor will communicate this point to the investor and, following the BigData tendency, it will be recovered the instantaneous power and the accumulated energy.

On the other hand, training hardware, besides all the above, it has a Buck converter and a pyranometer. Thanks to the Buck converter it is possible to obtain the voltage that generates the real maximum power point. It is done by means of a sweep of duty cycle obtaining the point where the highest power is reached. The pyranometer allows knowing real solar radiation values to be able to train the neural network.

As mentioned, the software has two neural networks. The first of them has the function of calculating the solar irradiance from short-circuit current and panel temperature measurement. This first neural network therefore has two neurons in its first layer and an output neuron. It has been developed with eight neurons in its hidden layer.

The second neural network from the first network output data, together with temperature and short-circuit current, obtains the maximum power working point of the photovoltaic panel. This second network consists of three neurons in the first layer, and also a single output neuron. In this case it has been developed with seven neurons in its hidden layer.

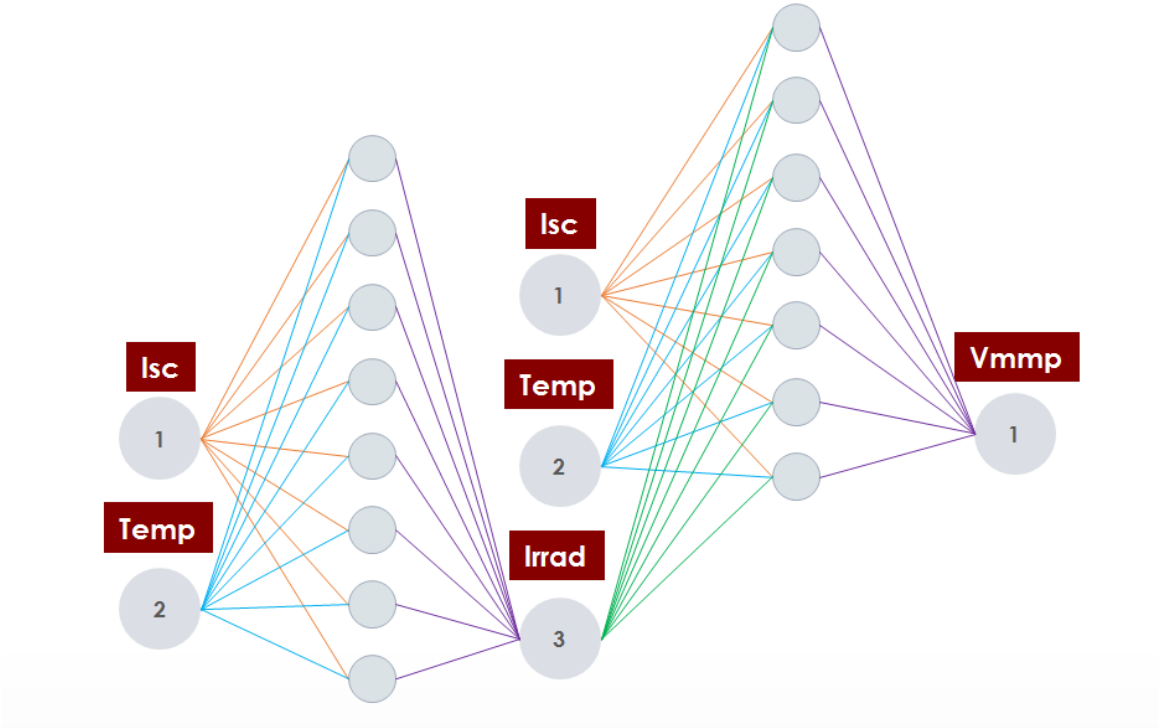


Figure 4: Structure of ANNs to provide maximum power voltage (Vmmp).

The performance of neural networks after the training process has been evaluated with validation data set. The real maximum power point obtained by the training hardware was compared with the results obtained through the neural network, as can be seen in the graphic. The quadratic error is less than 5%, and then the obtained neural network works correctly.

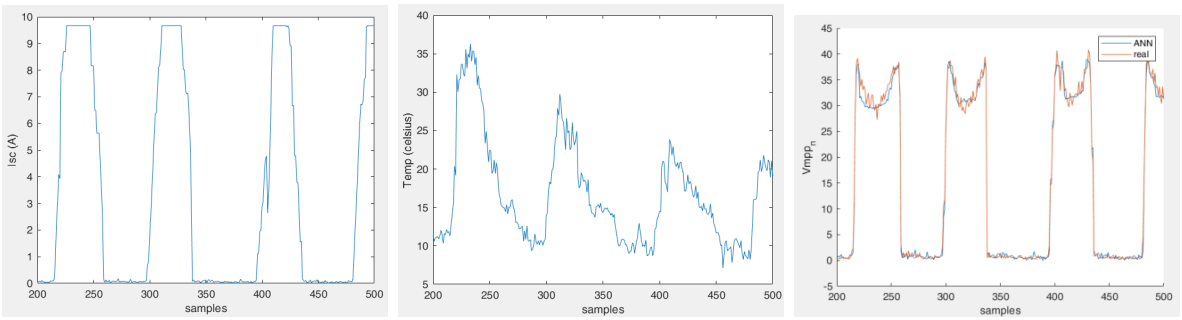
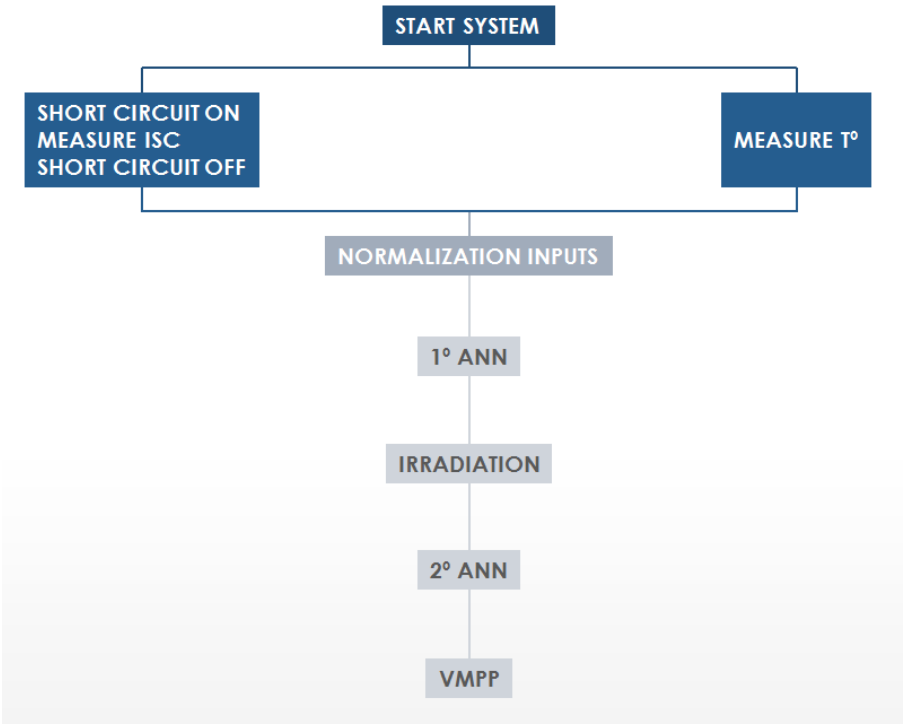


Figure 5: Performance of ANNS.

The operation of the global system is summarized in the following flowchart:



4 Results and conclusions

After the development of the method, a real scenario tests have been carried out. During these tests, the method has been validated by comparing the electric power generation with the new method and the conventional one. Both installations are equivalent with the same number and kind of panels. The conventional maximum power point tracking system used is the "perturb and observe" one.

Thus, tests started with the assembly of the two test facilities, one next to the other. Each of the installations consists of four high-power photovoltaic panels of 200 W each. In this way, the environmental conditions, temperature and irradiance are the same in both installations.

Figure 6: Installation of the system.

After training process, the facilities are operating under normal condition. Then, energy pro-

duction of both inverters has been recovered. The energy production has been monitored throughout a 6-month trial period, and the results of both installations are compared.

As a conclusion it can be said that there is a slightly higher value in the installation controlled with the new proposed method. Along the trial period, it has been seen that the improvements exceed 9%. And also the increase is much higher on clear days, exceeding 10% of improvements.

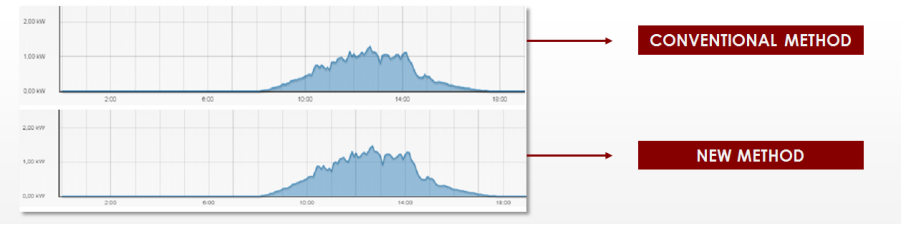


Figure 7: Comparison of methods over a day.

Additionally, according to the BigData philosophy, to carry out a remote control management software has been developed. It allows showing information in real time and historical data in a selectable time interval. Mainly it is shown: power, energy, irradiance and temperature.

References

[1] Renewables 2017 Global Status Report. REN21.

[2] Global Market Outlook 2017-2021. SolarPower Europe.

[3] Sharma, D.K., and Purohit, G., Maximum Power Angle (MPA) Based Maximum Power Point Tracking (MPPT) Technique for Efficiency Optimization of Solar PV System.

[4] Rodney & L. J. Tai, Priscilla & H. Mok, V., Solar irradiance estimation based on photovoltaic module short circuit current measurement. 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications, 2013.

[5] Hua, Ch. and Shen, Ch., Control of DC/DC Converters for Solar Energy System with Maximum Power Tracking. Department of Electrical Engineering National Yunlin University of Science & Technology, Taiwan.

31

Influence of different pathologies on the dynamic behaviour and against fatigue of railway steel bridges

Fran Ribes-Llario[♭1], Julián Santos[♭], Álvaro Poti[♭] and Julia Real Herráiz[♭]

(♭) Institute of Multidisciplinary Mathematics,
Universitat Politécnica de València.

1 Introduction

To fully understand track and vehicle dynamics is crucial when it comes to maintenance, safeness and operational matters. This importance is even bigger when analyzing singular sections of the infrastructure, as it is the case of railway bridges.

According to [2], the analysis of the response of a bridge under the effect of a moving train is not trivial, because the excitation forces exerted over the infrastructure involve not only the characteristics from a moving load but also repeated load pulses from consecutive axles, bogies, and carriages. It is influenced by numerous factors, such as structure natural frequencies and damping, train speed, or the length of both the structure and the train. Within this context, [2] studied the influence of different parameters, including the bridge-to-carriage length ratio, while [5] demonstrated that the primary frequencies in the bridge response might be caused by the driving frequencies, related to the time the train spent crossing the bridge, and by the dominant frequencies, which are caused by the repeated loads.

On the other hand, the actions generated by the trains on the railway bridges have this cyclic character necessary to generate the breakage by fatigue, with very important load oscillations, being the tensional variation the most influential parameter.

2 Methodology

The aim of the current research is to study how different pathologies in either the infrastructure, or the structure may affect the dynamic behavior of a vehicle-track-structure system. This affection is analyzed at different vehicle speeds, and it is presented in terms of displacements and structure accelerations. It is also intended to determine how these pathologies can reduce the fatigue life of the different structural elements that make up the bridge.

[1]e-mail: frarilla@cam.upv.es

In order to identify the affection of different pathologies to the dynamic response of the Bridge caused by the traffic of vehicles, and to the structural affection that it implies, a FEM is developed. A steel railway bridge built in the beginning of the XX century will be the focus of the research. Measurements obtained form a load test on the structure will be used for calibrating the FEM model.

Once the model has been calibrated, a series of scenarios with different pathologies will be simulated, both wheel-rail contact and structure. The aim is to know the dynamic behavior, on the one hand, and its affection by the fatigue resistance of the materials. It is analyzed the impact of vehicle speed on displacements and vibrations in an undamaged track and the influence of different pathologies is studied. The main objective is to assess to which extent the behavior shown for undamaged tracks is altered by the considered pathologies at the normal speed of the section (50 km/h).

For this analysis, two wheel-rail contact pathologies will be considered, rail corrugation and squat.

Carrying out an evaluation of the fatigue behavior of the structure would be a complex and imprecise task, since the history of loads and axes that the structure has suffered throughout its history is totally unknown. In this type of structure, the elements that tend to be more involved in front of fatigue are the stringers. This fact occurs because the number of cycles presented is much higher than that presented by the main elements of the structure. Fatigue analysis is based on Miner's cumulative damage rule. The analysis is reduced to simplify the efforts to which the piece is subjected and simplify it in a set of simple efforts and analyze the total damage as the sum of the various accumulated damages due to the different efforts:

$$\sum_{j=1}^{i} \frac{n_j}{N_j} = \frac{n_1}{N_1} + \frac{n_2}{N_2} + \frac{n_3}{N_3} + \cdots + \frac{n_i}{N_i} \geq 1$$

Once the tensions of the beam have been obtained, the remaining fatigue life is estimated. To achieve this, the number of cycles will be counted from the S-N curve, for the detail 71. From this table the value of N will be obtained for each load case.

Once the different scenarios have been planned, a comparison of these results will be conducted. The purpose is to quantify in which condition each of the pathologies presents with respect to the original pattern.

3 Results

The procedure followed is as follows, firstly, the load cycles that the bridge can support are calculated considering the track in perfect condition at a certain speed. Then, the same parameter is calculated for the different scenarios considered. From these data, we obtain the reduction of service life of the element.

In the following graph, the calculation has been made at a speed of 50 km/h. It can be observed that with the appearance of a 2 meter wave wear, which hardly had any implications from a

dynamic point of view, we find a 10% decrease in the useful life. Considering the rest of the defects we find values of decrease between 60 and 90%, which shows the great influence of these defects in the resistance to fatigue of these elements in metal bridges.

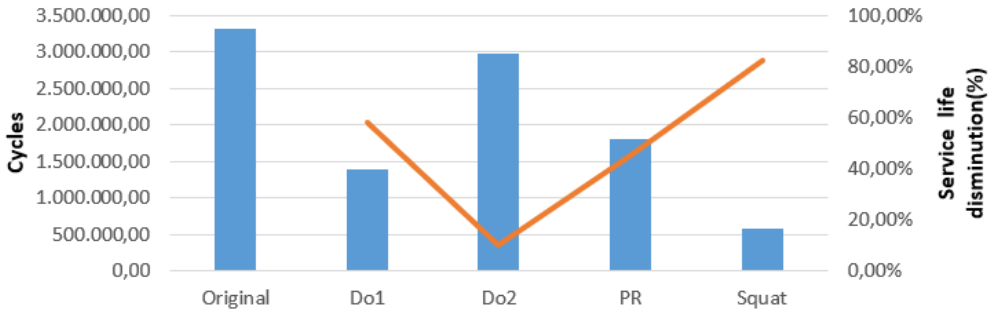


Figure 1: Cycles supported by the stringers with the passage of trains at 50 km/h.

When increasing the speed to 90 and 150 km/h we find the same trends as in the case of 50 km/h. There are variations in the percentages that decrease, but the trend continues in each of them. The data can be observed in the following figures.

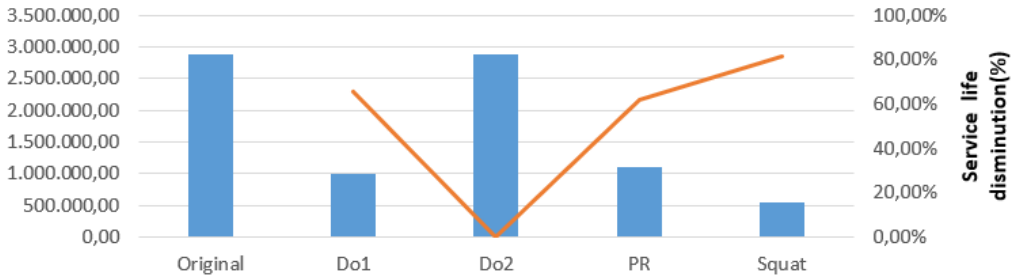


Figure 2: Cycles supported by the stringers with the passage of trains at 50 km/h.

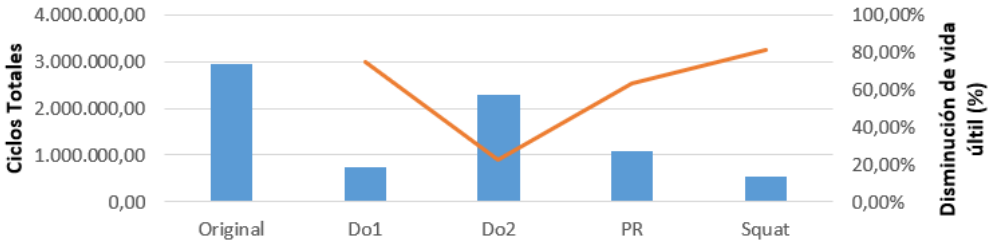


Figure 3: Cycles supported by the stringers with the passage of trains at 150 km/h.

In the following graph, the number of cycles for each defect is compared with the speed. It is observed that speed is not a factor that affects both the stresses that the elements are subject to as well as the defects.

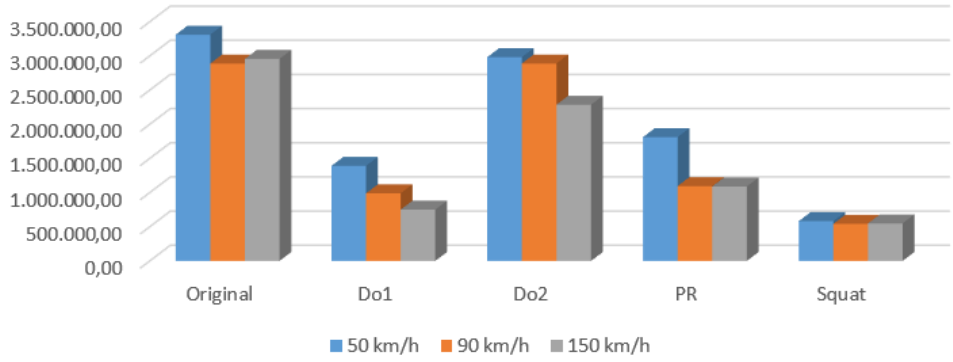


Figure 4: Cycles supported by the stringers with the passage of trains at different speeds and different wheel-rail pathologies. Original.

4 Conclusions

In the current research the dynamic performance of damaged and undamaged vehicle-track-systems built over a multi-span bridge is studied. This analysis is carried out in terms of displacements and accelerations, which are calculated at different checkpoints of the track and the structure.

Once studied the dynamic behavior of the structure, we proceed to study the affection of these pathologies to fatigue.The study of the fatigue life of the different elements is clearly influenced by any pathology that appears. In long-standing bridges, the need for periodic studies is obvious, since the different elements will be influenced by numerous pathologies due to their high age.

The great influence of the pathologies in the life to fatigue of the elements becomes noticeable, much more significant of what is the speed.

It is worth noting the great influence of short wave undulatory wear and squat. This fact shows that there is a great relationship between the increase of vibrations and the reduction of supported cycles.

References

[1] Dominguez, J., Railways bridges dynamics for high-speed railways: calculus methods and study of resonance, 2001.

[2] Mao, L., Lu, Y. and Woodward, P., Frequency characteristics of railway bridge response to moving trains with consideration of train mass. *Engineering Structures*, 42: 9-22, 2012.

[3] Museros, P., Vehicle-structure interaction and resonance effects in isostatic railways bridges for high-speed lines, 2002.

[4] Real, J., Zamorano, C., Comendador, R. and Real, T., Computational considerations on 3-D finite element method models of railway vibration prediction in ballasted tracks. *Journal of Vibroengineering*, 16(4): 1709-1722, 2014.

[5] Yang, Y. and Lin, C., Vehicle bridge interaction dynamics and potential applications. *Journal of Sound and Vibration*, 247-259, 2005.

32

Statistical-vibratory analysis of wind turbine multipliers under different working conditions

Miriam Labrado Palomo[♭], Teresa Real Herráiz[♭], Rubén Sancho McGill[♭], Carlos Canales Guerola[♭] and Bárbara Carreras Peris[♭]

(♭) Institute of Multidisciplinary Mathematics, Universitat Politécnica de València.

The aim of this research is to analyse the effect of different working conditions of wind turbines and sensor placement in fault diagnosis of a wind turbine multiplier.

The Hilbert-Huang transform has been selected to decompose and analyse the vibratory signals obtained from the gearbox. This method is divided in two fundamental steps [2]: first, the original signal is decomposed into Intrinsic Mode Functions (or IMF for short) through Empirical Mode Decomposition, then, the IMF which contain the frequencies of interest are transformed to obtain the instantaneous frequency of the signal and its energy. This methodology has been proven as the most reliable to extract the necessary information of the vibratory signals [3].

Four wind turbines were selected to be analysed from a wind farm located in the Canary Islands (Spain). These wind turbines contain a three-step multiplier gearbox with a ratio of 1 to 59.5, and meshing frequencies located at 31, 122 and 455 Hz.

The vibratory signal for each sensor was decomposed into 16 IMF, and the meshing frequencies were found at IMFs 4, 6 and 7. No defect presence was detected directly, since these units are well maintained. A posterior visual inspection performed on each gearbox confirmed this.

Different load cases were analysed to assess the performance of this methodology, by comparing the results obtained from a minimum power output state (wind speed < 3.5 m/s) and peak power output state (wind speed > 15 m/s). The results obtained show that the algorithm was able to correctly detect the meshing frequencies in both scenarios, and no single defects were found, with the only difference being the magnitude of the acceleration signals and energy spectrum, as it would be expected.

e-mail: frarilla@cam.upv.es

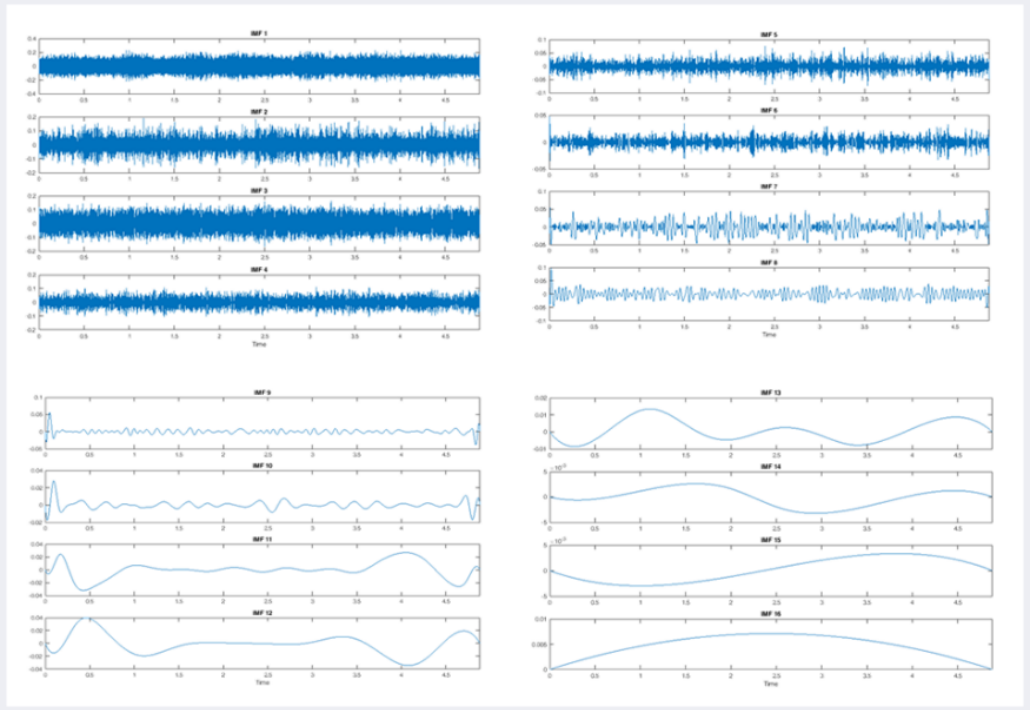


Figure 1: Acceleration signal decomposition into 16 IMF.

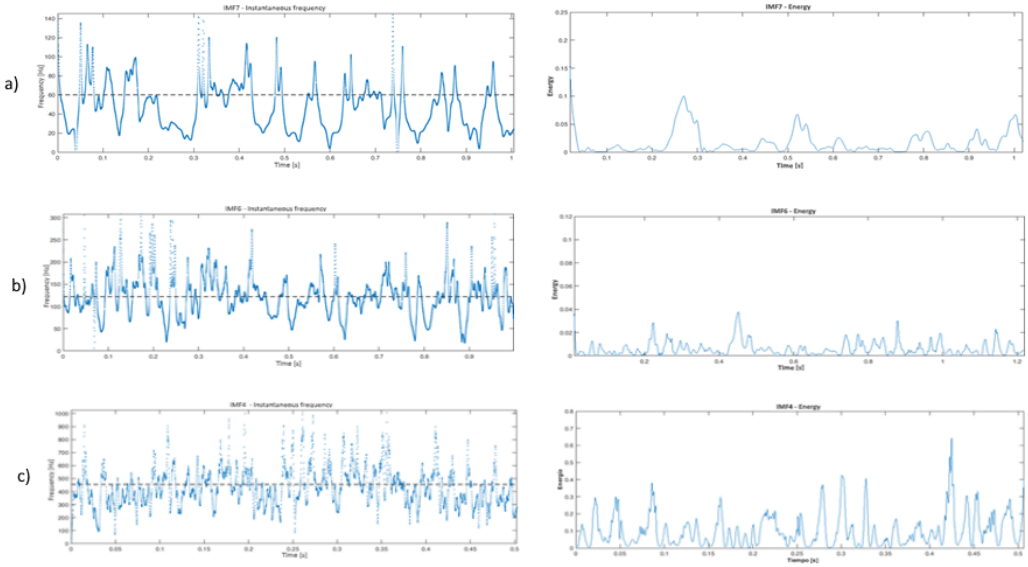


Figure 2: Instantaneous frequency (left) and signal energy (right) plots for three IMF: a) IMF7 [455 Hz], b) IMF6 [122 Hz], and c) IMF4 [31 Hz].

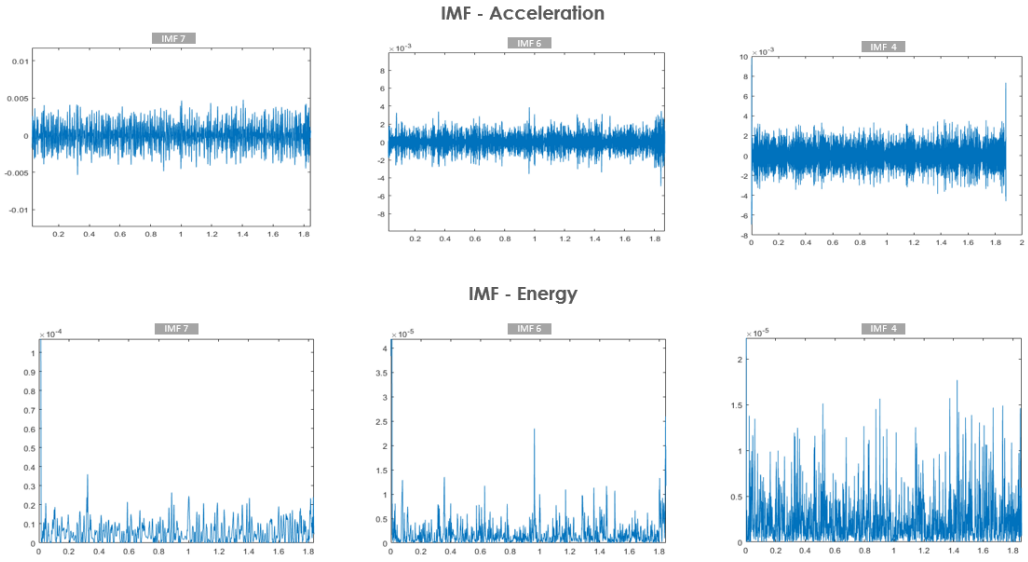


Figure 3: Acceleration data and signal energy for IMF 7, 6, and 4 (left to right) in a low power working condition (< 3 m/s wind speed).

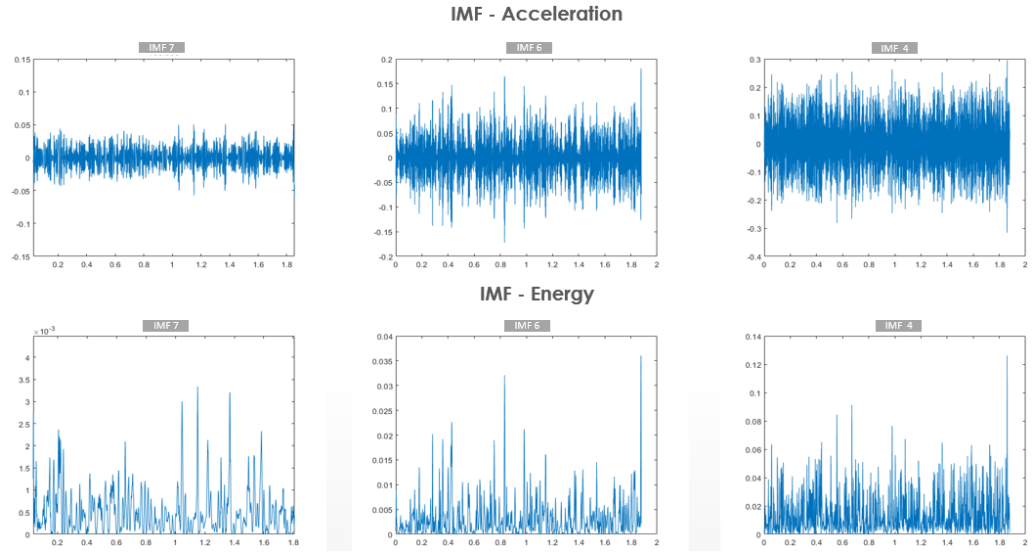


Figure 4: Acceleration data and signal energy for IMF 7, 6, and 4 (left to right) in a peak power working condition (> 15 m/s wind speed).

Different locations of the accelerometer sensors were also reviewed. The gearbox casing allows to install sensors in three different directions: radial, tangential and axial. Therefore, these three cases were tested in order to obtain the best location for the sensor location.

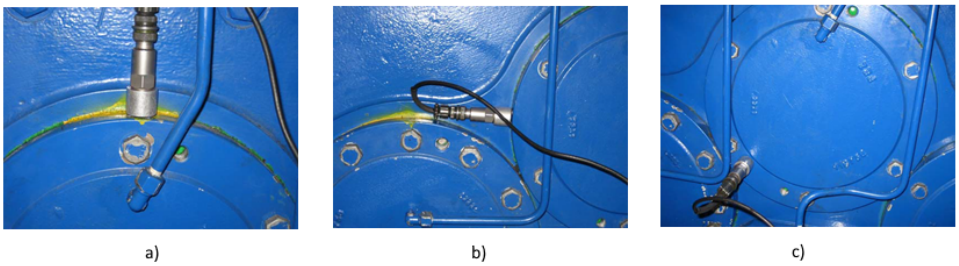


Figure 5: Sensor placement for each possible configuration: a) radial, b) tangential, and c) axial.

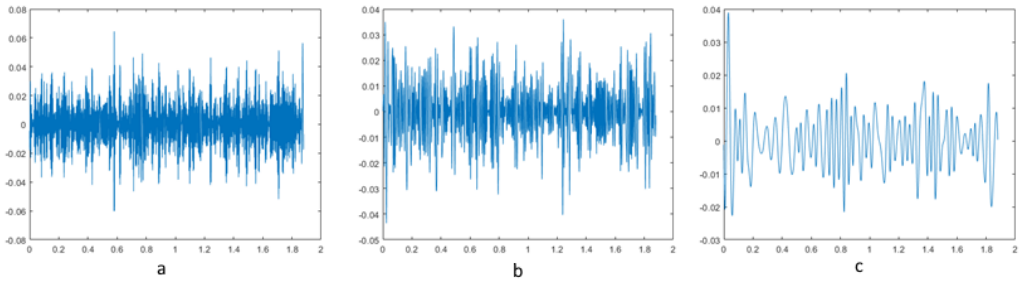


Figure 6: Acceleration plot in a 2 second time window for three different sensor directions: a) radial, b) tangential, and c) axial.

A first review of the signals obtained from each position reveals that both radial and tangential directions offer better results at detecting and obtaining the frequencies of interest, this is due to the higher intensity of the signal since most vibrations produced by the gears are transferred to the structure in these directions.

Finally, all different positions and directions were tested for each gear stage in all possible configurations, as shown in Figure 7.

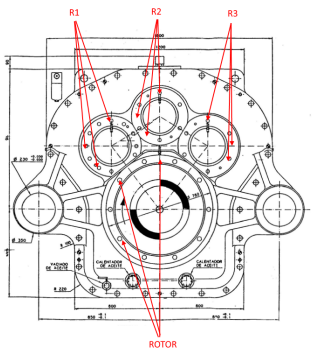


Figure 7: Sensor location in the gear box for all different possible configurations.

Statistical analysis of all collected data shows that the success rate for detecting and obtaining these vibratory signals of each step is higher for the radial and tangential sensor positions.

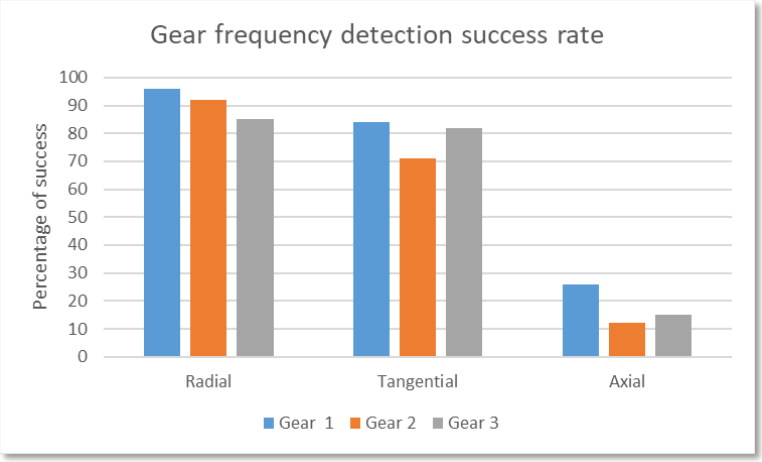


Figure 8: Percentage of success in meshing frequency detection for all three possible directions: radial, tangential, and axial.

In summary, in this work a methodology for analysing vibratory data of wind turbine gearboxes has been selected, and the necessary algorithms to process acceleration signals from the sensors installed, which can be run autonomously to process large sets of data were developed.

Data was collected and analysed from three different wind turbines during a period of two months, which, after extensive review, lead to the conclusion that the best location for obtaining the vibratory signals of interest is directly on the gearbox casing, at radial or tangential directions.

It was also obtained that different wind turbine regimes affect the obtained data by modifying its magnitude, but the necessary information can still be extracted for all possible working conditions.

For future work, different wind turbine models will be analysed, and the described methodology will be further validated by testing it in a known problematic gearbox.

References

[1] Martin, N., KAStrion porject: a new concept for the condición monitoring of wind turbines. Twelve International Conference on Condition Monitoring and Machinery Failure Prevention Technologies CM2015, 2015, Oxford UK, United Kingdom.

[2] Lei, Y., Lin, J., He, Z. and Zuo, M. J., A review on empirical mode decomposition in fault diagnosis of rotating machinery, *Mechanical Systems and Signal Processing*, 35(1-2): 108-126, 2013.

[3] Loudritis, S. J., Damage detection in gear systems using empirical mode decomposition, *Engineering Structures*, 26(12): 1833-1841, 2004.

33

Analysis of finite dimensional linear control systems subject to uncertainties via probabilistic densities

J.-C. Cortés[♭], A. Navarro-Quiles[♭], J.-V. Romero[♭1] and M.-D. Roselló[♭]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

Control Theory is a branch of Mathematics that studies the behaviour of a dynamical system with controllers, one or more, applied through actuators. Furthermore, its main objective is to develop control models for controlling such systems using a control action in an optimum manner, that is ensuring the stability. Applications of Control Theory, in irrigation systems, can be found since the ancient Mesopotamia more than 2000 years B.C. But it was not until the 1868 that the first definitive mathematical description of Control Theory was established in the works by J.C. Maxwell, [1]. From this moment Control Theory gained importance, becoming nowadays a fundamental tool to develop new technologies.

A control problem consists in finding controls, say $u(t)$, such that the solution of a model, $x(t;u)$, coincides or gets close to a target value x^1 at a final time instant T, i.e., $x(T;u) = x^1$. Generally, an optimal control problem is defined via a set of differential equations, ordinary or partial, describing the states which depend on the control variables that minimize a particular cost function of the form

$$J(v) = \frac{1}{2}||x(t;u) - x^1||^2 + \frac{\beta}{2}||u||^2,$$

where $\beta \geq 0$ allows us to penalize using too much costly control.

On the other hand, the parameters that appear in this kind of formulations are generally set via experimental data. Therefore, since these values are obtained from certain measurements and samplings, they contain an intrinsic uncertainty. This situation allows us to consider inputs as random variables or stochastic processes rather than constants or deterministic functions, respectively.

[1]e-mail: jvromero@mat.upv.es

2 Probabilistic solution

As it has been pointed previously, a control problem is defined through a set of ordinary or partial differential equations depending on the dimensionality of the system, finite or infinite dimensional, respectively. In this contribution, given its interest, we consider the finite dimensional linear control system

$$\begin{array}{rcl} x'(t) & = & \mathbf{A}x(t) + \mathbf{B}\,u(t), \quad 0 < t < T, \\ x(0) & = & x^0. \end{array} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state of the system, $x^0 \in \mathbb{R}^n$ is the initial data, $\mathbf{A}$ is a $n \times n$ matrix of the free dynamics part, $\mathbf{B}$ is a $n \times m$ matrix, with $m \in \mathbb{N}$ and $m \leq n$, called the control operator and $u(t)$ the m-dimensional control vector.

We study, from a probabilistic point of view, the randomized control problem

$$\begin{array}{rcl} x'(t,\omega) & = & \mathbf{A}(\omega)x(t,\omega) + \mathbf{B}(\omega)\,u(t,\omega), \quad 0 < t < T, \\ x(0,\omega) & = & x^0(\omega). \end{array} \tag{2}$$

where all the input parameters $A_{ij}(\omega)$, $B_{ik}(\omega)$, $0 \leq i,j \leq n$ and $0 \leq k \leq m$, the starting seed $x^0(\omega) = [x_1^0(\omega), \ldots, x_n^0(\omega)]^\top$ and the final target $x^1(\omega) = [x_1^1(\omega), \ldots, x_n^1(\omega)]^\top$ are assumed to be absolutely continuous random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assuming that the system is controllable, we can obtain a solution stochastic process of the initial value problem (2) [2,3],

$$x(t,\omega) = \left(\mathrm{e}^{\mathbf{A}(\omega)t} - H(t;\mathbf{A}(\omega),\mathbf{B}(\omega))\mathrm{e}^{\mathbf{A}(\omega)T}\right) x^0(\omega) + H(t;\mathbf{A}(\omega),\mathbf{B}(\omega))x^1(\omega),$$

where

$$H(t;\mathbf{A}(\omega),\mathbf{B}(\omega)) = \int_0^t \mathrm{e}^{\mathbf{A}(\omega)(t-s)}\mathbf{B}(\omega)\mathbf{B}^*(\omega)\mathrm{e}^{\mathbf{A}^*(\omega)(T-s)}\,\mathrm{d}s\,\Lambda^{-1}(T;\mathbf{A}(\omega),\mathbf{B}(\omega)).$$

and

$$\Lambda(x;\mathbf{A}(\omega),\mathbf{B}(\omega)) = \int_0^x \mathrm{e}^{\mathbf{A}(\omega)(T-t)}\mathbf{B}(\omega)\mathbf{B}^*(\omega)\mathrm{e}^{\mathbf{A}^*(\omega)(T-t)}\mathrm{d}t.$$

Now, we apply the Random Variable Transformation method (see Reference [4]) to obtain the first probability density function of the solution stochastic process

$$\boxed{\begin{array}{rcl} f_1(x,t) & = & \displaystyle\int_{\mathbb{R}^{h_1}} f_{x^0,x^1,\mathbf{A},\mathbf{B}}\left(\left(\mathrm{e}^{\mathbf{A}t} - H(t;\mathbf{A},\mathbf{B})\mathrm{e}^{\mathbf{A}T}\right)^{-1}(x - H(t;\mathbf{A},\mathbf{B})x^1), x^1, \mathbf{A}, \mathbf{B}\right) \\ & & \det\left(\left(\mathrm{e}^{\mathbf{A}t} - H(t;\mathbf{A},\mathbf{B})\mathrm{e}^{\mathbf{A}T}\right)^{-1}\right)\mathrm{d}x^1\,\mathrm{d}\mathbf{A}\,\mathrm{d}\mathbf{B}, \end{array}}$$

3 Numerical example

In this example we consider that A, B and x^1 are deterministic matrices

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad x^1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

In addition, we assume that the random vector $x^0(\omega)$ follows a multivariate Normal distribution with mean $\mu = (1, 1)$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}, \qquad \text{i.e.} \qquad x^0 = (x_1^0, x_2^0) \sim \mathrm{N}(\mu, \Sigma)$$

In Figure (1) the first probability density function is plotted for the time instant $t = 0.1$. Phase portrait is represented in Figure (2). In the phase portrait the expectation and 50% and 90% confidence intervals are shown at the time instants $t \in \{0, 0.1, 0.5, 0.9\}$. We observe that the solution tends to the point $x^1 = (0, 0)$. Notice that as x^1 is deterministic, then, the variability vanishes as time increases.

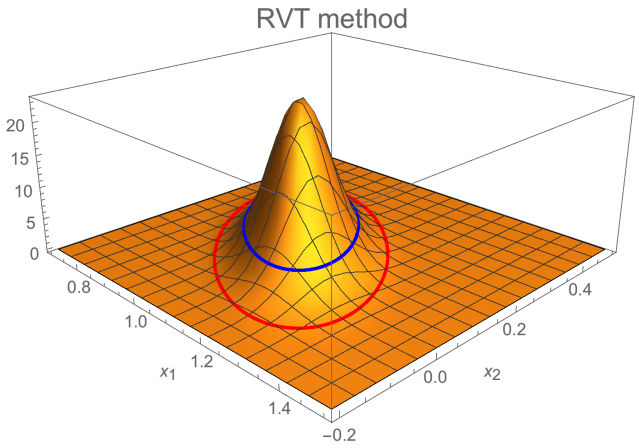


Figure 1: First probability density function of the solution stochastic process at the time instant $t = 0.1$. 50% (blue curve) and 90% (red curve) confidence regions.

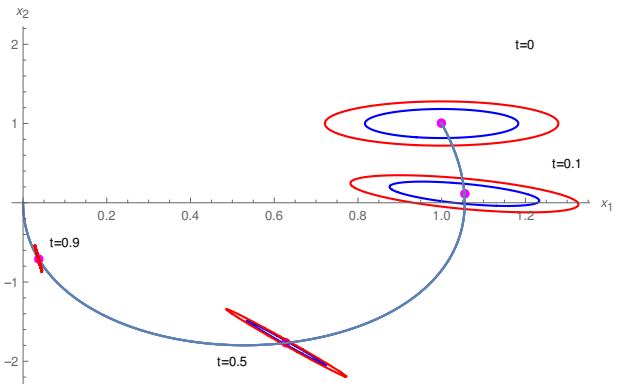


Figure 2: Continuous spiral line represents the expectation of the solution. 50% (blue curve) and 90% (red curve) confidence regions are plotted at the time instants $t \in \{0, 0.1, 0.5, 0.9\}$.

Acknowledgements

This work has been partially supported by the Ministerio de Economía y Competitividad grant MTM2017–89664–P. Ana Navarro Quiles acknowledges the funding received from Generalitat Valenciana through a postdoctoral contract (APOSTD/2019/128).

References

[1] Fernández Cara, E. and Zuazua Iriondo, E. Control Theory, mathematical achievements and perspectives, *Boletín de la Sociedad Española de Matemática Aplicada*, 26: 79–140, 2003.

[2] Zuazua Iriondo, E. Chapter 7 - Controllability and Pbservability of Partial Differential Equations: Some Results and Open Problems, *Handbook of differential equations: Evolutionary equations*, 3: 527–621, 2007.

[3] Lazar, M. and Zuazua Iriondo, E. Greedy controllability of finite dimensional linear systems. *Automatica*, 74: 327–340, 2016.

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

34

Topographic representation of cancer data using Boolean Networks

C. Santamaría[♭1], B. García–Mora[♭], G. Rubio[♭] and A. Falcó[♮]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València,
(♮) Departamento de Matemáticas, Física y Ciencias Tecnológicas,
Universidad CEU Cardenal Herrera.

1 Introduction

Research in biomedicine, and cancer in particular, currently has generated a very large amount of data, mainly due to the different omics. The search for patterns and relationships between them, in short, the search for explanations about the meaning of these data, requires mathematical and computational models. In this sense, mathematical modelling can contribute to the rational design of optimal treatment protocols involving combinations of surgery, chemotherapy and radiotherapy and the development of new therapies. Mathematical models can also be used to determine the mode of action of new compounds and to identify new targets for drug design.

In this regard, there is a lot of research to find a classification in molecular subtypes of bladder cancer that allow stratifying patient populations in a clinically relevant manner. This means to find subtypes with significant differences in survival data among them, that have a different natural history of the disease, and find out if those different subtypes should each be treated distinctly. These classifications are made by looking for clusters from molecular characterizations of tumors (repeated observations of global patterns of gene or protein expression in a tumor type). But the existing classifications are only partly consistent.

Our goal is to explore the problem of the biomolecular classifications of a type of bladder cancer, the Nonmuscle-Invasive Bladder Cancer (NMIBC) with an approach based on Boolean Networks that has been successful in recent years.

2 Modeling framework

Our starting hypothesis is the following: molecular subtypes suggest states in the evolution of a process. Genetic interactions may be considered dynamics in a network, and so a mathematical approach to locate particular states in a network could be appropriate.

[1]e-mail: crisanna@imm.upv.es

Gene Regulatory Networks (GRNs) are abstractions of the structures that underlie the processes of gene expression. Reconstructing these structures from the observation of biological phenomena, and in particular, diseases such as cancer, is a very active field of research (see [1] for an overview of recent developments).

One approach to study GRN is Boolean Networks (BNs). The BN model consists of a set of nodes (representing genes), directed edges (representing interactions among genes) and Boolean functions [2]. Denoting the state of a gene i by x_i, at a given time its expression would be ON if $x_i = 1$, or OFF if $x_i = 0$. For a m-gene BN, the state vector $X^t = (x_1^t, x_2^t, \ldots, x_m^t)$ collects the expression of all the genes in the network at discrete point time t. A Boolean function F_i determines the output value x_i at time $t + 1$. Denoting F the vector function that collects all F_i, the gene expression state of a BN will be given by

$$X^{t+1} = F(X^t).$$

A trajectory in the state space is a sequence of states $X^0, \ldots, X^t, X^{t+1}, \ldots$ The trajectories may converge to a single state, named point attractor. The set of initial states that converge to an attractor is its basin of attraction. The network topology and the Boolean functions at each node determine the attractor structure, which consists of attractors, trajectories and basins of attraction.

Our goal is to explore with bladder cancer data the approach that we summarize below, following the explanation given by their authors in [3]. The state space of a GRN is the space that contains all theoretically possible gene expression patterns of that GRN. One gene expression pattern of the GRN is represented in this framework by a point in a high-dimensional state space. The attractor state is a stable equilibrium state, and in this way it can be seen as a point at the bottom of what could be thought of as a "valley" in the multidimensional space (the basin of attraction). "The GRN of a particular genome maps into one landscape and each geographic position in it represents a unique gene expression profile, i.e., a cell state, in the high-dimensional state space of the genomewide network" [3].

Nowadays it is well established that cell types may be seen as attractors of an underlying genetic network dynamics, and "Boolean networks have become a standard model for studying the statistical properties of attractors and the dynamics of networks" [4].

In parallel the idea that tumors are a consequence of a disrupted regulation of normal cell and tissue development is gaining weight. A tumor cell would be an aberrant cell type. If cell types are attractors, then cancer cells can also be represented by attractors.

A given GRN produces a particular landscape with "valleys" and "hills" that may be so complex that not all attractors are occupied by those cells that represent physiological cell types in the body. The majority of those attractors likely represent abnormal, non-viable gene expression patterns. However, if there is a viable proliferative phenotype associate with some attractor it could represent a cancer attractor.

In this framework a genetic mutation can be seen as a change in the network architecture: remove a node or a connection, reinforce or add a connection, etc. So, mutations introduce

changes in the wiring diagram of the GRN. This implies a change in the landscape topography.

In recent years, various applications have been made based on this or a close approach. Some examples are [5–8]. For a review of modeling approaches in this context see [9].

3 Results and discussion

Our study takes as a reference a recent paper [10] whose overall goal was to determine whether specific DNA mutations and/or copy number variations are enriched in specific molecular subtypes. They used the complete TCGA (The Cancer Genome Atlas) RNA-seq dataset. Taking into account three different published classifiers developed by the research groups participating in the article, they assigned TCGA's bladder cancers to molecular subtypes. Our idea is also to use public databases, with a completely different methodology. But the final objective, like theirs, is to correctly classify the molecular subtypes of the NMIBC. As we said earlier, there is still no satisfactory classification.

This paper is a first step is our purpose of applying this approach to the NMIBC. We have considered different bladder cancer subpathways collected from REACTOME (free online database of biological pathways). From this source and the software *Cytoscape* (software platform for visualizing complex networks) we can select the genes related with a specific disease, in our case NMIBC.

A resulting network is given by the graph of (Figure 1) with 31 genes and their interactions. The software translates these interactions in Boolean Functions. We used the R-package BoolNet [11]. The Boolean network has 2^{31} possible states and we obtain 32 attractors. The sequence of the states from one initial state to its attractor is depicted in the Figure 2.

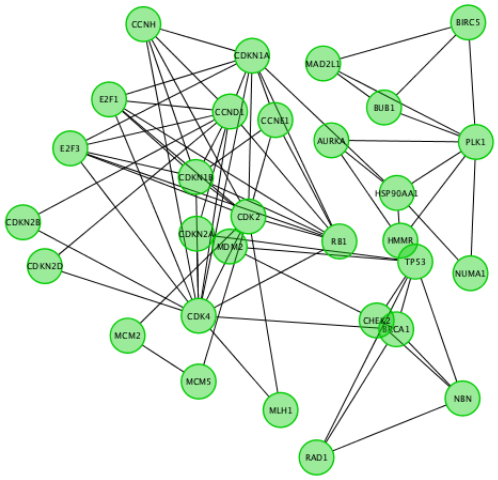


Figure 1: Gene Regulatory Network in Bladder Cancer

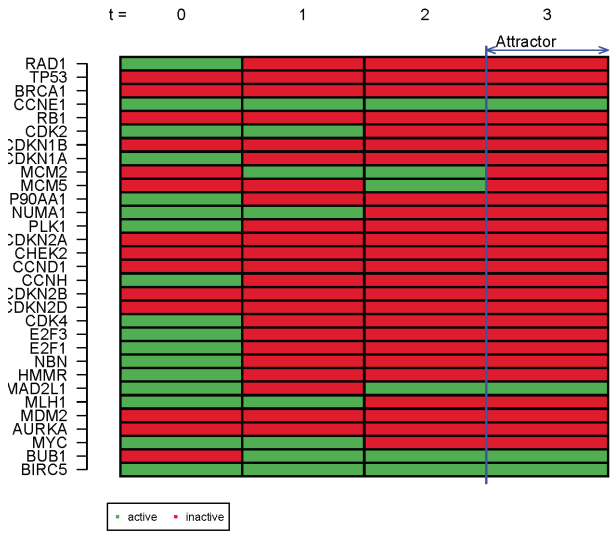


Figure 2: Attractors in the Bladder Cancer

.

4 Conclusions

The objective has been to explore the application of a Boolean network approach in Bladder Cancer establishing the differentiated attractors in this carcinoma. The main objective in the long run is to establish a classification to obtain the risk groups in this tumor, which is very necessary for clinical practice, but it is still a problem awaiting resolution.

References

[1] Sanguinetti, G., Vân Anh Huynh-Thu, Editors. Gene Regulatory Networks, Methods and Protocols. New York, Springer, 2019.

[2] Zhoua, JX., Samal, A., d'Hérouël, AF., Nathan D., Price, ND. and Huang, S., Relative stability of network states in Boolean network models of gene regulation in development. *BioSystems*, 142-143:15-24, 2016.

[3] Huang, S., Ernberg, I. and Kauffman, S., Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Seminars in Cell & Developmental Biology*, 20: 869-876, 2009.

[4] Bornholdt, S. and Kauffman, S., Ensembles, dynamics, and cell types: Revisiting the statistical mechanics perspective on cellular regulation. *Journal of Theoretical Biology*, 467:15-22, 2019.

[5] Font-Clos, F., Zapperi, S. and La Porta, C., Topography of epithelial-mesenchymal plasticity. *PNAS*, vol. 115, no. 23, 2018.

[6] Zhu, X., Yuan, R., Hood, L. and Ao, P., Endogenous molecular-cellular hierarchical modeling of prostate carcinogenesis uncovers robust structure. *Progress in Biophysics and Molecular Biology*, 117: 30-42, 2015.

[7] Poret, A. and Boissel, J-P., An in silico target identification using Boolean network attractors: avoiding pathological phenotypes. *C. R. Biologies*, 337: 661-678, 2014.

[8] Poret, A., Guziolowski, C., Therapeutic target discovery using Boolean network attractors: improvements of kali. *R. Soc. open sci.*, 5: 171852, 2018.

[9] Tyson, J.J., Laomettachit, T and Kraikivski, P. Modeling the dynamic behavior of biochemical regulatory networks. *Journal of Theoretical Biology*, 462: 514-527, 2019.

[10] Choi, W., Ochoa, A., McConkey, D.J., Aine, M., Höglund, M., Kim, W.Y., Real, F.X., Kiltie, A.E., Milsom, I., Dyrskjøt, L. and Lerner, S.P., Genetic Alterations in the Molecular Subtypes of Bladder Cancer: Illustration in the Cancer Genome Atlas Dataset. *European Urology*, 72: 354-365, 2017.

[11] Müssel, C., Hopfensitz, M. and Kestler, H.A.. BoolNet–an R package for generation, reconstruction and analysis of Boolean networks *Bioinformatics*, 26:1378-80, 2010.

35

Trying to stabilize the population and mean temperature of the World

Antonio Caselles ♭ and Maria T. Sanz♮[1]

(♭) IASCYS member (retired), Departament de Matemàtica Aplicada,
Universitat de València,
(♮) Departament de Didàctica de la Matemàtica,
Universitat de València.

1 Introduction

The global average temperature on Earth has increased and there are institutions such as NASA's Goddard Institute for Space Studies (GISS) who observe this fact from the year 1880, claiming that the increase was 0.8 degrees Celsius, being two-thirds of this heating from the year 1975. The goal, according to GISS scientists, is to provide an estimate of the change in temperature that could be related with global climate change. Other scientists [1] claim that climate change is not only related with external variables, such as those mentioned in [2] (the world's population, international trade, population ageing, income and, technological progress), but also with human attitudes, and that such a change could lead to the extinction of the human species. Between these human attitudes, it raises the election of the type of energy to use, being renewable energy which must be enhanced [3, 4].

Our conclusion when trying to assess the present situation in relation to stabilizing temperature and population in the world is that more studies are needed but defining optimal intervention strategies is urgent. For us, the following ones are key words: energy, complex model, uncertainty, dynamics, validation, and optimization. Energy sources and energy consumption have demonstrated to be the root of the problem [5]. A complex model tries to relate a lot of variables in a non-evident, non-stable or non-linear manner. Considering uncertainty means that the model has to be stochastic and results have to be presented either with its respective confidence interval or with its standard deviation (liability is important). Considering dynamics implies that data and results are time series (intervention has to be continuous). Validation means that the model has been proved to be useful for the aims it has been designed (the classical validation method, but not the only one, is comparing calculated results with real ones, the so called ex-post method). The model has to be able to find optimal strategies, that is, a set of values of the input variables along time that demonstrates to reach the proposed goal in the considered period of time. Genetic algorithms, which try to imitate nature's procedures

[1]e-mail: m.teresa.sanz@uv.es

(natural selection, mutation, migration, etc.), have demonstrated to be the most adequate ones for optimizing complex situations like the considered one.

2 Demographic Model

The starting point of the model construction process is to state a demographic dynamic model neither considering ages nor sexes:

$$\frac{dPOPL(t)}{dt} = (BIRR(t) - DEAR(t) + TINH(t) - TEMH(t)) \cdot POPL(t) \tag{1}$$

Where, *BIRR*, is the birth rate; *DEAR*, is the death rate; *TINH*, is the immigration rate; *TEMH*, is the emigration rate; and *POPL*, is the total population.

This model has been applied to solve problems in a given country [6]. This time it will be applied with data of the entire world to try to investigate if it is possible to reach the stability of the average global temperature ($XAGT$) and at the same time that of the world population, through interventions on the amount and type of energy consumption and, in the affirmative case, to determine the best guide line to do it. In order to reach this goal two new variables have been included in the model. On the one hand, *EQUI* [7] pretends to assess, in an aggregate form, the negative impact of human activity on the ecosystem through energy consumption. It considers renewable energies, nuclear energy and fossil energy. Its values vary from 0 to 1, being values near 0 the ideal ones. On the other hand there is the Human Development Index (*HDI*) [8]. Thus, the variables related with both are also introduced in the system based on Eq. (1) as stated in Eq. (2).

$$\frac{dPOPL(t)}{dt} = (BIRR(equi(t), hdi(t)) - DEAR(equi(t), hdi(t)) + NMIR(t)) \cdot POPL(t) \tag{2}$$

HDI is calculated through three subsystems: education, health and economy. The implied variables in it [8] are: Life expectancy at birth (years) (*LEBI*), Literacy Rate (%) (*LIRA*), Gross Enrolment Rate (%) (*GREN*) and Gross Domestic Product per capita (GDP) (PPP $). That is:

$$HDI(lebi, educ, gdp) = (lebi \cdot gdp \cdot educ)^{\frac{1}{3}} \tag{3}$$

$$EDUC(lira, gren) = \frac{2}{3} lira + \frac{1}{3} gren \tag{4}$$

But, GDP is calculated by Eq. 5:

$$GDPR(t) = FCEX(t) + GRCF(t) + EBSE(t) - IBSE(t) \tag{5}$$

This four variables (Final consumption expenditure (*FCEX*), Gross capital formation (*GRCF*), Goods and services exports (*EBSE*), Goods and services imports (*IBSE*)) can be found as historical data but, [7] shows that these variables can be also disaggregated as functions of energy consumption (*CONE*) (see Eq. (6) and (7)).

$$CONE(t) = (ENUS(t) \cdot POPL(t-1) - 1e12)/(1e14 - 1e12) \tag{6}$$

$$GDPR(t) = FCEX(cone(t)) + GRCF(cone(t)) + EBSE(cone(t)) - IBSE(cone(t)) \quad (7)$$

Where *ENUS* is the energy consumed per capita. *EQUI* is improved in this work, in order to gain manageability, as Eq. (8) states.

$$equi = \sqrt{cont \cdot (1 - ncon)} \quad (8)$$

Where *CONT*, measures the degree of use of the energies affecting negatively the environmental quality and *NCON*, measures the degree of use of the energies affecting it non-negatively.

$$cont = \sqrt{yfof \cdot ynuc}; \quad yfof = \frac{fofu}{enus \cdot popl}; \quad ynuc = \frac{nucl}{enus \cdot popl}; \quad yree = \frac{reen}{enus \cdot popl} \quad (9)$$

$$1 = ynuc + yree + yfof \quad (10)$$

Where *YFOF*, *YNUC* and *YREE* are the ratios between the used fossil fuel, nuclear and renewable energies and the total one, respectively. *FOFU* is the used fossil fuel energy and *YFOF* is calculated in the model as a linear combination of the energies emitting CO_2 (oil, carbon and gas). Remark that *NCON* is equal to *YREE* (Eq. (8) and (9)).

The average global temperature (*XAGT*) is calculated as a function of the CO_2 emissions index (*YCO2*), CH_4 emissions index (*YCH4*), and N_2O emissions index (*YN2O*) as Eq. (11) states. Following [5] "the CO_2 concentration is the most important long-lived "forcing" of climate change". With respect to methane, the same Organization comments: "this is much less abundant in the atmosphere". Finally, it qualifies N_2O as "A powerful greenhouse gas". Taking these considerations into account, the variable measuring the mean temperature of the world should be a geometric average which weights corresponding to CO_2 and N_2O are greater than those corresponding to methane.

$$xagt(yco2, ych4, yn2o) = a + b\sqrt[3]{yn2o^{1.2} \cdot ych4^{0.6} \cdot yco2^{1.2}} \quad (11)$$

Finally, birth and death rates are calculated from the *HDI* and *EQUI* indices (see Eq. (2)) as a combination of two logistics fitted to the real data of the calibration period. But, given that the sense of HDI and EQUI is opposed, we use as independent variable in these equations a new variable combining both indices as Eq. (12) states:

$$wellbeing = hdi \cdot (1 - equi) \quad (12)$$

3 Model Validation

Given that the available data, obtained from World Data Bank [9], correspond to the 1990-2015 period, we use the data from years 1990 to 2000 (or from 1991 to 2001 for the case of the referred indices *HDI* and *EQUI*) for calibration, and the data from years 2002 to 2015 for validation.

The validation process is considered successful because the determination coefficients, R_2, are very high, and the maximum relative error does not exceed 5% in any case. Nevertheless, the only non-good fitting is the corresponding to average global temperature vs. time. This is due to the great dispersion of data along time. For instance, in the deterministic validation: a) Population. Maximum relative error: $0.0916177\% < 5\%$, $R_2 = 0.999994$; b) *HDI*, maximum relative error: $0.603967\% < 5\%$, $R_2 = 0.991411$; c) *EQUI*, maximum relative error: $4.00631\% < 5\%$, $R_2 = 0.878743$; d) *XAGT*, maximum relative error: $26.454\% > 5\%$, $R_2 = 0.121093$.

4 Optimization of the use of the different types of energy in the future

According to the proposed objectives, the model is able to optimize by means of a genetic algorithm (GA) the quantity and proportion of the use of the different energy sources in order to keep the global temperature and population stable. **Nevertheless**, the control variables used by the GA to obtain this goal, are: the CO_2 emissions from oil, gas, coal, and other fossil fuel sources, the rate of renewable energy and the CH_4 and N_2O emissions.

These variables enter into the GA through an array of seven rows and three columns. Each row includes: a maximum and a minimum values of the respective variable as well as a percentage of variation over its initial value that determines the yearly search window (Table 1). Data of Table 1 have been obtained by observing the historical data and its trend. Note that maximum values correspond to real data of 2015 (the beginning of the simulated period), that is because they are considered as bounds that should not be exceeded. As for the percentage of maximum yearly relative variation for each variable, it is tentatively stated.

Input variable	**Minimum (Kt)**	**Maximum (Kt)**	**Maximum yearly variation (%)**
PETR	1000000	11807740*	2
GAST	300000	6622602*	2
CARB	1000000	15130042*	2
OTHE	100000	2277207*	2
YREE	0.1769905686* (rate)	1 (rate)	10
XN2O	100000	3153742.479*	2
XCH4	100000	8014066.562*	2

Table 1: Maximum, minimum and maximum yearly variation of control variables. (*): Real data at the beginning of the simulated period (2015).

Furthermore, three objective variables to be minimized are introduced in the model: *OBJ1*, *OBJ2* and *OBJE*. The first one (*OBJ1*) represents temperature, *OBJ2* represents population and, *OBJE* represents a combination of both (previously normalized due to they have different dimensions). See Eq. (13) to (15).

$$OBJ1(t) = \frac{XAGT - \min_{XAGT}}{\max_{XAGT} - \min_{XAGT}} \tag{13}$$

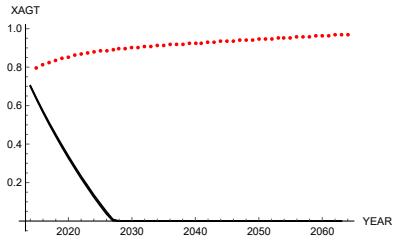
$$OBJ2(t) = \frac{POPL - \min_{POPL}}{\max_{POPL} - \min_{POPL}} \tag{14}$$

$$OBJE(t) = \sqrt{Abs(obj1)^{0.8} \cdot Abs(obj2)^{1.2}} \tag{15}$$

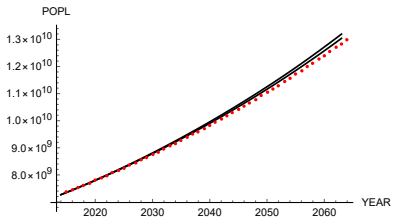
Note that, tentatively, we give a higher weight to stabilizing population. With respect to energy consumption, two tentative strategies are simulated. On the one hand, the energy consumed per capita (*ENUS*) is considered as following its historical trend (*STR1*). On the other hand, its tendency is reduced 10% per year (*STR2*).

Fig. 1 to 4 show the result of the minimization of OBJE by means of a genetic algorithm in the 2016-2045 period. They suggest what should be the actions of the main responsible persons in the world with respect to energy consumption and contamination emissions.

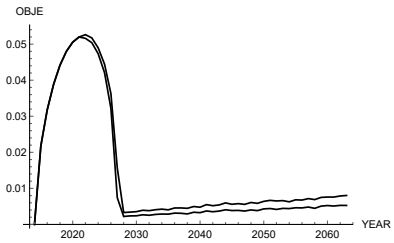
Fig. 1 show the trends of population and global mean temperature as well as the OBJE variable in both strategies. Graphically, significant differences between both strategies cannot be observed neither in OBJE (Figure 1c and 1f), nor in its involved variables (Figure 1a, 1b, 1d and 1e).



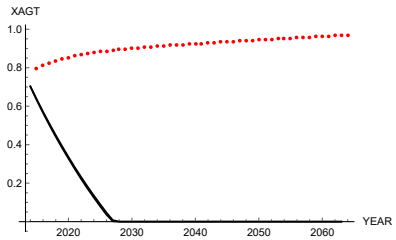
(a) Global average temperature (*STR1*)



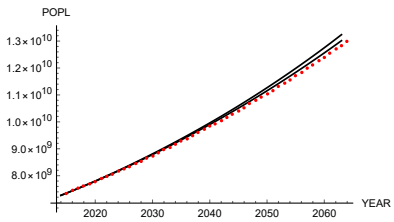
(b) World population (*STR1*)



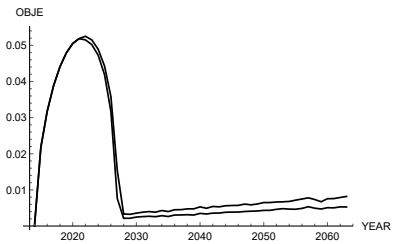
(c) OBJE variable (*STR1*)



(d) Global average temperature (*STR2*)



(e) World population (*STR2*)

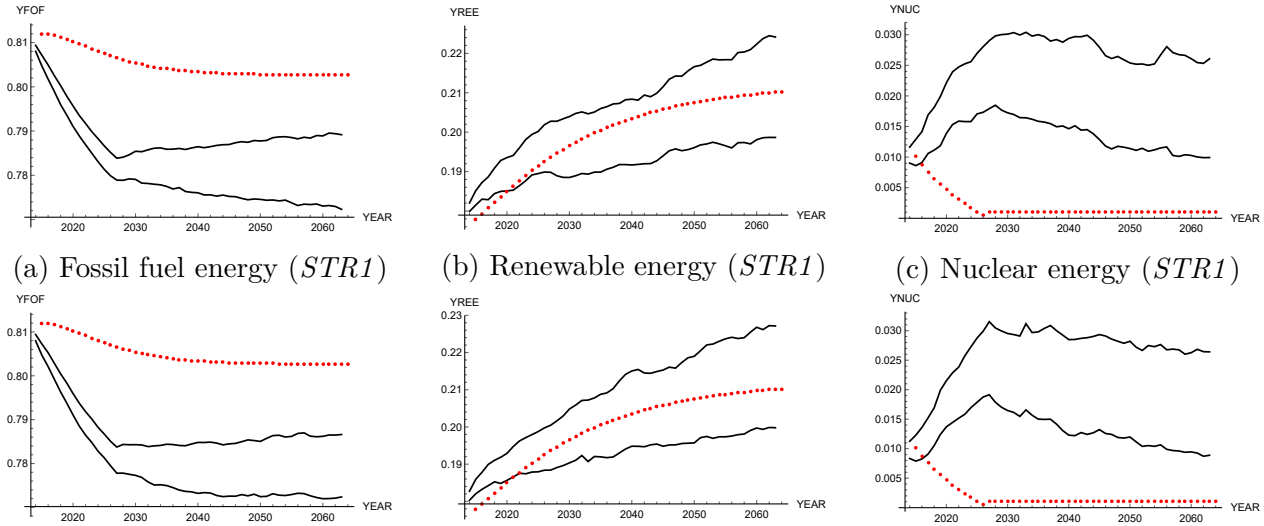


(f) OBJE variable (*STR2*)

Figure 1: Results for the optimal scenario. Simulated trend (dots). Minimum and maximum values for optimal intervention (lines).

Nevertheless, Fig. 2 shows some differences between *STR1* and *STR2* in energy rates. It shows that the proportion of fossil fuel consumption has to be reduced up to around 0.78, that is, 4.5% approximately with respect to its historical trend. Similarly, nuclear energy has to be potentiated up to a proportion of around 0.01 (*STR1*) and around 0.007 (*STR2*). Finally, the optimal evolution of the renewable energy consumption increases over the historical trend, an increase of around 10%, *STR1* and around 15% in the case of *STR2*.

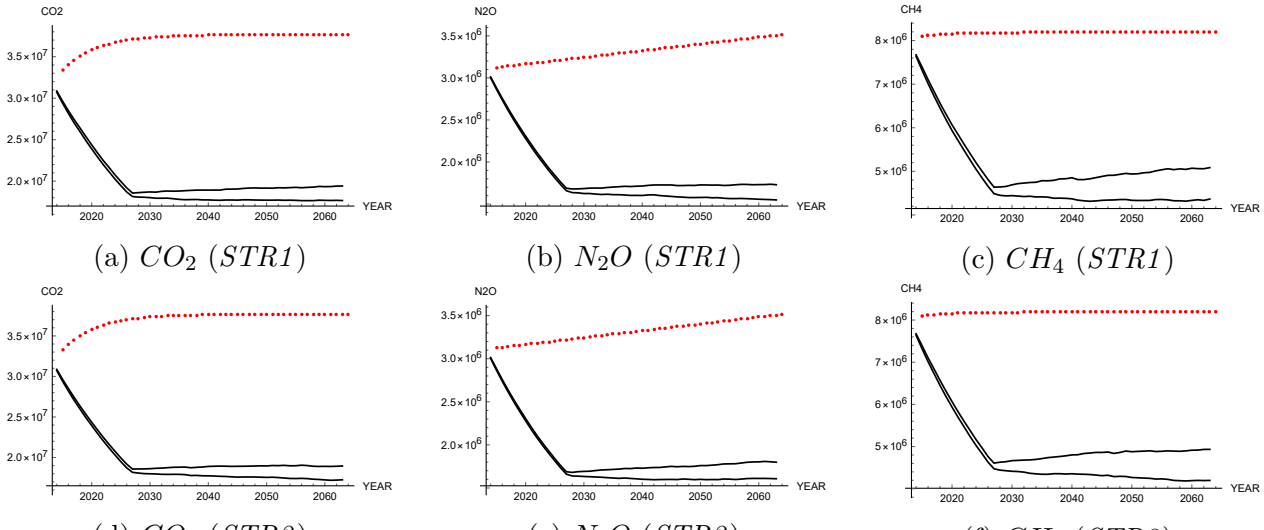
Fig. 3 suggests that all contamination emissions have to be reduced around 44% with respect to its historical trend before 2045 in both strategies.



(a) Fossil fuel energy ($STR1$) (b) Renewable energy ($STR1$) (c) Nuclear energy ($STR1$)

(d) Fossil fuel energy ($STR2$) (e) Renewable energy ($STR2$) (f) Nuclear energy ($STR2$)

Figure 2: Consumption proportion. Simulated trend (dots). Minimum and maximum values for optimal intervention (lines).

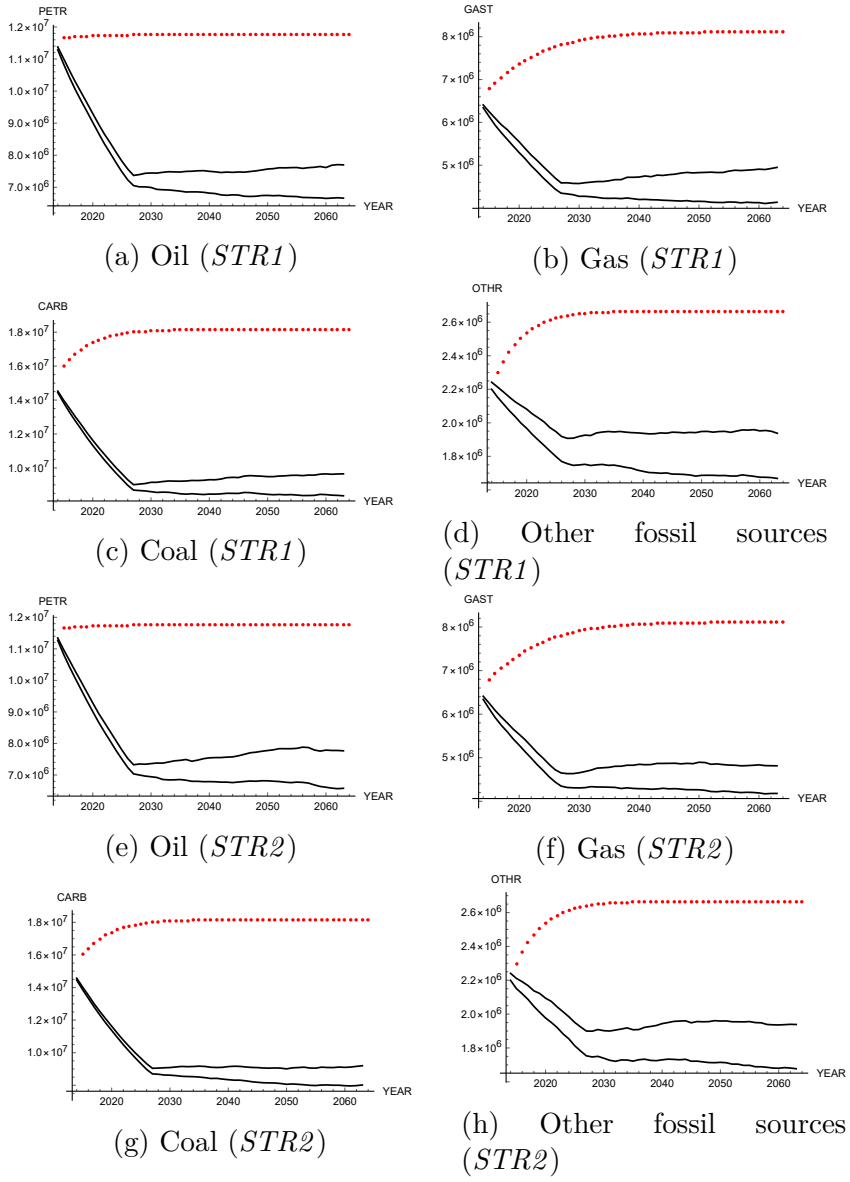


(a) CO_2 ($STR1$) (b) N_2O ($STR1$) (c) CH_4 ($STR1$)

(d) CO_2 ($STR2$) (e) N_2O ($STR2$) (f) CH_4 ($STR2$)

Figure 3: Emissions. Simulated trend (dots). Minimum and maximum values for optimal intervention (lines).

Finally, Fig. 4 suggests that all CO_2 emissions have to be reduced to its half part approximately before 2045 (around 42% those coming from oil, around 44% those coming from gas, around 50% those coming from coal and, around 27% those coming from other fossil sources).



(a) Oil (*STR1*) (b) Gas (*STR1*)

(c) Coal (*STR1*) (d) Other fossil sources (*STR1*)

(e) Oil (*STR2*) (f) Gas (*STR2*)

(g) Coal (*STR2*) (h) Other fossil sources (*STR2*)

Figure 4: CO_2 emissions. Simulated trend (dots). Minimum and maximum values for optimal intervention (lines).

References

[1] Janssen, M. and de Vries, B., The battle of perspectives: a multi-agent model with adaptive responses to climate change. *Ecological Economics*, 26: 43-65, 1998.

[2] Graves, P.E., Implications of global warming: Two eras. *World Development Perspectives*, 7–8: 9–14, 2017.

[3] Arent, D.J., Wise, A. and Gelman, R., The status and prospects of renewable energy for combating global warming. *Energy Economics*, 33: 584–593, 2011.

[4] Trainer, T., A critical analysis of the 2014 IPCC report on capital cost of mitigation and of renewable energy. *Energy Policy*, 104: 214–220, 2017.

[5] https://climate.nasa.gov (accessed 29.04.19).

[6] Micó, J.C., Soler, D. and Caselles, A., Age-Structured Human Population Dynamics. *Journal of Mathematical Sociology*, 30: 1-31, 2006.

[7] Sanz, M.T., Caselles, A., Micó, J. C. and Soler, D., Including an environmental quality index in a demographic model. *Int. J. Global Warming*, 9(3), 2016.

[8] UNDP (2015). United Nations Human Development Report.

[9] https://data.worldbank.org/ (accessed 29.04.19).

36

Optimizing the demographic rates to control the dependency ratio in Spain

Joan C. Micó [♭], David Soler[♭], Maria T. Sanz[♮1], Antonio Caselles[♯] and Salvador Amigó[*]

(♭) Institut Universitari de Matemàtica Multidisciplinar,
Universitat Politècnica de València,
(♮) Departament de Didàctica de la Matemàtica,
Universitat de València,
(♯) IASCYS member (retired), Departament de Matemàtica Aplicada,
Universitat de València,
(∗) Avaluació i Tractaments Psicològics,
Universitat de València.

1 Introduction

The decline in the birth rate and the increase in longevity are a fact in the developed countries and, a growing trend in the developing countries, so the implications of these facts on future well-being are fundamental, in particular the impact on the population pyramid and, even more, on the dependency rate.

According to the studies on EU-28 [3] the proportion of people of working age is decreasing, while the relative number of retired people is increasing. Demographers warn that this is due to a decrease in births [1,2], but not only this demographic phenomenon affects this increase in the dependency rate. For example, migration control is an essential tool. Particularly, an organization such as the Department of Economic and Social Affairs of UN warns that only a fifteen per cent of the Governments control their current immigration to address their population ageing, and only a a thirteen per cent deal with the problem of the long-term population decline [3].

Underlying these facts, a problem arises: what would be the appropriate birth and migration happening for a society such that, within a reasonable period, its dependency ratio changes its trend?

The aim of this work is to adapt the demographic model presented by [4] to solve the described problem. The model modifications here presented include considering the death and migration rates as control variables, which obligates to change some model parts. This new model has been validated for the case of Spain in its deterministic and stochastic formulations. Finally, the model is used to determine the future evolution of the birth, death and migration rates in Spain

[1]e-mail: m.teresa.sanz@uv.es

in order to decrease the dependency ratio. The evolution of these demographic phenomena, which are considered optimal, is calculated by using strategies and scenarios.

2 Demographic Model

The new demographic model (see [4] to appreciate the changes), written in its continuous form, is constituted by the following equations:

$$\frac{\partial w_i(t,x)}{\partial t}+c\frac{\partial w_i(t,x)}{\partial x} = (-d_i(x)\cdot grde_i(t,x)+f_i(x)\cdot gryn_i(t,x)-g_i(x)\cdot grem_i(t,x))\cdot w_i(t,x) \quad (1)$$

$$w_i(t,0) = birt(t)\cdot\frac{b_i(t)}{b_1(t)+b_2(t)}\cdot\int_0^{+\infty}(w_1(t,x)+w_2(t,x))dx\cdot\int_0^{+\infty}\bar{b}_i(x)\cdot w_2(t,x)dx \quad (2)$$

$$w_i(t_0,x) = u_i(x) \quad (3)$$

Where, $i = 1$ represents men and $i = 2$ women.

Eq. (1) is a von Foerster-McKendrick equation that determines the dynamics of population density depending on time and age, $w_i(t,x)$, where $d_i(x)$, $f_i(x)$ and $g_i(x)$ represent respectively the death, immigration and emigration rates, as a function of age. Also, $grde_i(t,x)$, $gryn_i(t,x)$ and $grem_i(t,x)$ are respectively the growth rates for each previous demographic phenomena, as functions of age and time.

Eq. (2) represents the boundary condition, that is, births at $x = 0$. In this equation, $\frac{b_i(t)}{b_1(t)+b_2(t)}$ is the proportion of men or women born (according to $i = 1$ or 2, respectively), that is, births per sex ($b_i(t)$) divided by the total number of births ($b_1(t) + b_2(t)$); $birt(t)$ is the birth rate, i.e., the total numbers of births ($b_1(t) + b_2(t)$) divided by the total population; and $\bar{b}_i(x)$ is the ratio between the fertility rate and births.

Eq. (3) is the initial condition, that is, the initial population density, $u_i(x)$, at $t = t_0$.

Some simplifying hypotheses are made on Eqs. (1) and (2) (similarly to those made in [4]) in order to introduce the death, emigration and immigration rates temporarily defined. Thus, the modifications introduced in the model are the following.

$$d_i(x)\cdot grde_i(t,x) \approx \bar{d}_i(x)\cdot\frac{d_i(t)}{d_1(t)+d_2(t)}\cdot deat(t)\cdot popt(t) \quad (4)$$

$$f_i(x)\cdot gryn_i(t,x) \approx \bar{f}_i(x)\cdot\frac{y_i(t)}{y_1(t)+y_2(t)}\cdot immi(t)\cdot popt(t) \quad (5)$$

$$g_i(x)\cdot grem_i(t,x) \approx \bar{g}_i(x)\cdot\frac{e_i(t)}{e_1(t)+e_2(t)}\cdot emig(t)\cdot popt(t) \quad (6)$$

In these equations, the proportions of deaths, immigration or emigration for men or women, ($\frac{d_i(t)}{d_1(t)+d_2(t)}$, $\frac{y_i(t)}{y_1(t)+y_2(t)}$, $\frac{e_i(t)}{e_1(t)+e_2(t)}$, respectively) (according to $i = 1$ or 2, respectively) are considered, that is, deaths, immigration and emigration per sex ($d_i(t), y_i(t)$ and $e_i(t)$) divided by the

total number of deaths, immigration or emigration. Finally, $\bar{d}_i(x)$, $\bar{f}_i(x)$ and $\bar{g}_i(x)$ are the ratios between the different demographic rates (functions of age) and deaths, immigration and emigration respectively in $t = 0$. Note that, $popt(t)$ can be calculated by the model as:

$$popt(t) = \int_0^{+\infty} (w_1(t,x) + w_2(t,x))dx \tag{7}$$

With these considerations on the initial model, the following equations are obtained:

$$\begin{aligned} \frac{\partial w_i(t,x)}{\partial t} + c\frac{\partial w_i(t,x)}{\partial x} = \Big(&(-\bar{d}_i(x) \cdot \frac{d_i(t)}{d_1(t)+d_2(t)} \cdot deat(t) + \bar{f}_i(x) \cdot \frac{y_i(t)}{y_1(t)+y_2(t)} \cdot inmi(t) \\ &- \bar{g}_i(x) \cdot \frac{e_i(t)}{e_1(t)+e_2(t)} \cdot emig(t)) \cdot \int_0^{+\infty} (w_1(t,x) + w_2(t,x))dx \Big) \cdot w_i(t,x) \end{aligned} \tag{8}$$

$$w_i(t,0) = birt(t) \cdot \frac{b_i(t)}{b_1(t)+b_2(t)} \cdot \int_0^{+\infty} (w_1(t,x) + w_2(t,x))dx \cdot \int_0^{+\infty} (\bar{b}_i(x) + w_2(t,x))dx \tag{9}$$

$$w_i(t_0,x) = u_i(x) \tag{10}$$

3 Model Validation

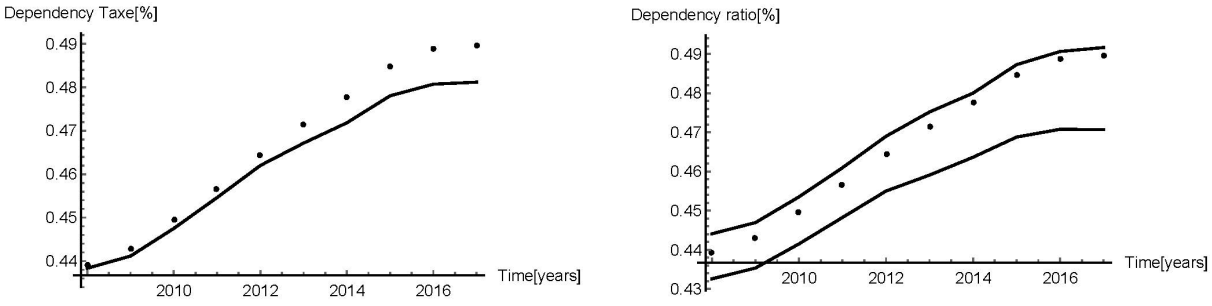
The validation of the model is performed for Spain in the 2007-2017 period, i.e., for those years whose information is available in the World Data Bank [5]. The obtained data are also used to fit input variables to time.

Although the model has been written as a set of differential and functional equations, the solutions have been calculated with the Euler Method, following [6, 7], which explain that the Euler Method is more adequate to solve such equations. In the case of the integral in Eq. (9), it is calculated through the Simpson Composite Rule. This approach results in a set of finite difference equations that has been programmed in Visual Basic 6.0 using Sigem [8, 9].

The corresponding validation has been performed like in [4]. On the one hand, the deterministic formulation of the model is validated through the determination coefficients and the random residuals tests. The real and simulated data are plotted in Figs. 1a and 2a. On the other hand, the stochastic formulation is also validated by checking that the historical data fall between the minimum and maximum simulated values (Figs. 1b and 2b). The validation process is considered successful because the determination coefficients, R^2, are very high, and the maximum relative error does not exceed 4.51% in any case. In the case of the stochastic validation, all the real data are within the 99% generated confidence interval.

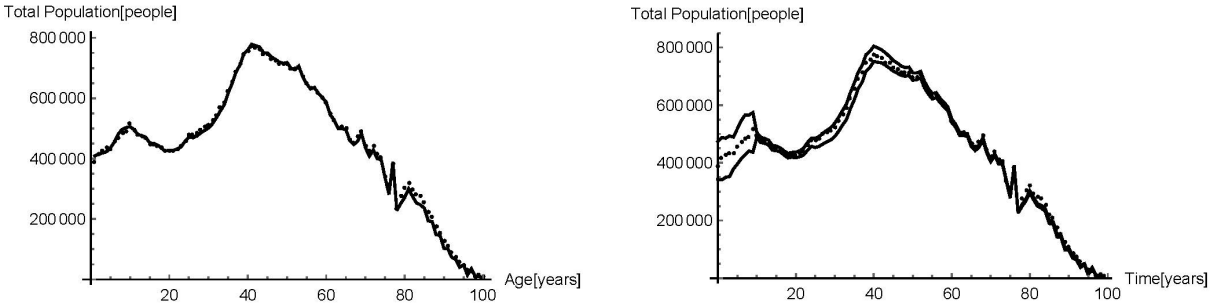
4 Model application

In the application case, the aim is to minimize the dependency ratio. This minimization decreases the pressure on the productive population. Thus, the dependency ratio is defined as



(a) Deterministic validation, $R^2 = 0.997542$ (b) Stochastic validation

Figure 1: Dependency ratio for Spain in the 2008-2017 period.



(a) Deterministic validation, $R^2 = 0.998135$ (b) Stochastic validation

Figure 2: Total population for Spain in 2017.

the dependent population (population with ages from 0 to 15 and with 65 or more) divided by the productive population (population with ages from 16 to 64). That is:

$$obje(t) = \frac{\sum_i (\int_0^{x_m} w_i(t,x)dx + \int_{x_M}^{+\infty} w_i(t,x)dx)}{\sum_i \int_{x_m}^{x_M} w_i(t,x)dx} \tag{11}$$

In Eq. (11), x_m is the minimum working age, generally $x_m = 15$, and x_M the retirement age, generally $x_M = 65$.

The method, that has been used to find the evolution of the input variables (control variables) that minimize the dependency ratio, that is, the objective variable $obje(t)$ is to determine strategies over control variables (see Table 1).

Control variable	**SS1**	**SS2**	**SS3**	**SS4**
Birth Rate	↑	↑	↑	↑
Emigration Rate	↑	↓	↑	↓
Immigration Rate	↑	↓	↓	↑

Table 1: Strategies to minimize the dependency ratio. ↑: to increase 5% the tendency; ↓ to decrease 5% the tendency.

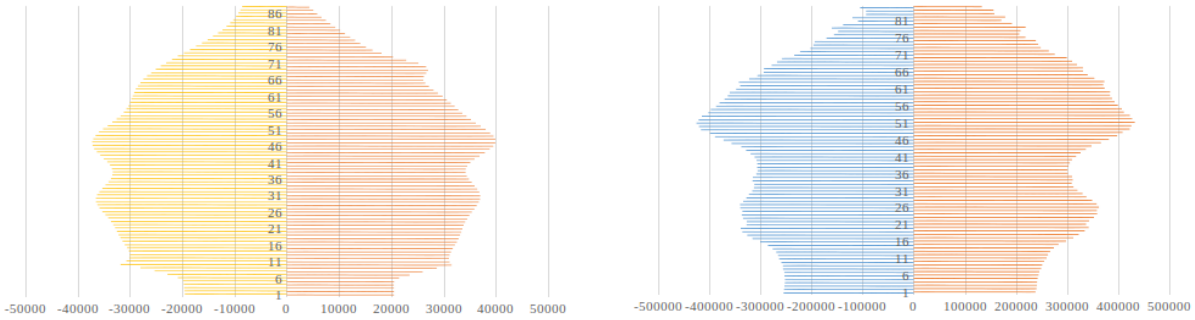
For the application case here presented (the case of Spain), the time t runs in the 2018-2027 period and, the corresponding deterministic model formulation results are shown in Table 2.

year	SS1	SS2	SS3	SS4
2018	0.4891209	0.488799	0.4919379	0.4863304
2019	0.4886366	0.4885366	0.4918719	0.4855817
2020	0.489179	0.4895301	0.4931213	0.4858498
2021	0.4872227	0.4879565	0.4917752	0.4837045
2022	0.4847987	0.4860044	0.4901535	0.4809992
2023	0.4832554	0.4850131	0.4894081	0.4791768
2024	0.4808911	0.4831095	0.4877837	0.476468
2025	0.4779461	0.4807993	0.4857836	0.473359
2026	0.4763169	0.4797138	0.4851585	0.4713179
2027	0.4748512	0.4790213	0.4790213	0.4695968

Table 2: Values of the dependency ratio, $obje(t)$, for the case of Spain in the 2018-2028 period.

To reduce the dependency ratio, i.e. to get more people in working ages with respect to those in non-working ages, it is necessary to apply the SS4 and to modify the trend of the demographic control variables on those terms: increasing the birth and immigration rates and, reducing the deaths (Table 1).

In this situation, the Spain population pyramid has changed as Fig. 3 shows.



(a) 2017 (b) 2018

Figure 3: Pyramid population, female (right) and male (left) population for Spain.

References

[1] KC, S. and Lutz, W., The human core of the shared socioeconomic pathways: Population scenarios by age, sex and level of education for all countries to 2100, *Glob. Environ. Chang.*, 42: 181–192, 2017.

[2] `http://ec.europa.eu/eurostat/statistics-explained/index.php/Population_structure_and_ageing/es` (accessed 27.04.19).

[3] https://esa.un.org/poppolicy/publications.aspx (accessed 27.04.19).

[4] Micó, J. C., Soler, D., Sanz, M. T., Caselles, A. and Amigó, S., Birth rate and population pyramid: A stochastic dynamical model, *Modelling for engineering & Human Behaviour 2018*, 292–297, 2018.

[5] https://data.worldbank.org/ (accessed 27.04.19).

[6] Djidjeli, K., Price, W.G., Temarel, P. and Twizell, E.H., Partially implicit schemes for the numerical solutions of some non-linear differential equations, *Appl. Math. Comput.*, 96: 177–207, 1998.

[7] Letellier, C., Elaydi, S., Aguirre, L.A. and Alaoui, A., Difference equations versus differential equations, a possible equivalence for the Rossler system? *Phy. D: Nonlin. Phen.*, 195: 29–49, 2004.

[8] Caselles, A., A tool for discovery by complex function fitting, in: R. Trappl (Ed.), Cybernetics and Systems Research'98, Austrian Society for Cybernetic Studies, Vienna, 787–792, 1998.

[9] Caselles, A., Modelización y Simulación de Sistemas Complejos (Modeling and Simulation of Complex Systems), Universitat de València, Valencia (Spain), 2008. Available in http://www.uv.es/caselles as well as SIGEM, (accessed 27.04.19).

37

An integer linear programming approach to check the embodied CO_2 emissions of the opaque part of a façade

David Soler [♭1], Andrea Salandin[♮] and Michele Bevivino[♯]

(♭) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València,
(♮) Centro de Tecnologías Físicas,
Universitat Politècnica de València,
(♯) Architect and buildings energy consultant,
Via Fiume 6, 39100 Bolzano, Italy.

1 Introduction

The building sector is responsible of about 40% of the total energy consumption and 45% of the CO_2 emissions in Europe [1]. Therefore, it is a key potential source of both energy saving and avoiding the greenhouse effect, and current European legislation takes into account those important aspects [2]. In the last few years, several papers have used Linear Programming approaches to solve optimization problems relating with energy and buildings [3–6]. In particular, very recently, two papers [7,8] have modelized the problem of finding the best composition of a building's façade, in order to, among other objectives, minimize or to upper bound the thermal transmittance of the façade [9]. Thermal transmittance measures the rate of heat flow through the elements of the building envelope and it is a key magnitude to assess the energy efficiency of the building. All the above cited papers consider the reduction of the energy consumption once the building has been constructed, but they do not take into account the embodied energy of the building, nor its embodied CO_2 emissions, which include the life cycle of any material used in the building construction: excavation, processing, construction, operation, maintenance, demolition and waste or recycling.

As buildings become more energy efficient, the relative proportion of embodied energy and associated carbon emissions arising during the building lifecycle increases [10]. Life cycle assessment (LCA) is a widely recognized and accepted method for the assessment of burdens and impacts throughout the lifecycle of a building [11]. LCA evaluates all resource inputs, including energy, materials and water, in order to calculate the environmental impacts of a building at either the material, product or whole building level. There are two main existing LCA databases available worldwide: University of Bath's Inventory of Carbon and Energy (ICE) [12] and the Swiss Ecoinvent database [13].

[1]e-mail: dsoler@mat.upv.es

It is important to highlight how the primary energy consumed in construction represents between 10% and 30% of the energy required for its functioning during the lifecycle of the building. The embodied energy in a building can be estimated by taking into account the weight of the required materials, the data base information and a pre-dimensioning through a program that can calculate quantities and estimate environmental costs [14]. Furthermore, thermal diffusivity can provide useful information for evaluating the thermal mass behavior, the speed of energy interchange and the accumulation capacity.

The aim of this work is to choose the adequate materials for the different layers of a façade, with their corresponding thicknesses, in order to minimize the embodied CO_2 emissions of the opaque part of a building's façade, at the same time that other restrictions inherent to the construction of the façade are met, such as, current legislation about thermal transmittance, budget limitations and total thickness of the façade. To do this, we formulate the problem as an Integer Linear Programming (ILP) problem with binary variables, following some ideas given in [7,8]. Results of this approach on a case study involving different scenarios for a 5-layer external wall are also presented.

2 ILP formulation

In this section, the problem of minimizing the embodied CO_2 emissions of the opaque part of a façade subject to certain construction restrictions, is modeled as an ILP problem. Concretely, we will minimize the embodied CO_2 emissions of $1m^2$ of the façade, because if the façade has a total of $S\,m^2$ of opaque part, it is only necessary to multiply our result by S. For a better understanding of the formulation, we first present some notations, the used variables and the parameters.

- Let n be the number of layers of the façade, which will be enumerated from inside to outside. Each layer $i \in \{1, \ldots, n\}$ is made of one of the m_i different materials available for this layer, and given a layer $i \in \{1, \ldots, n\}$, the material $j \in \{1, \ldots, m_i\}$ is available in r_{j_i} different thicknesses.
- For each $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m_i\}$ and $k \in \{1, \ldots, r_{j_i}\}$, the following parameters are considered:
  - $k_{i,j,k}$ number of kg that weighs $1m^2$ of material j with type of thickness k available for layer i.
  - $kco2_{i,j}$ number of kg of embodied CO_2 for each kg of material j available for layer i.
  - $t_{i,j,k}$ thickness corresponding to material j with type of thickness k available for layer i (note that k indicates the type of thickness, not the thickness).
  - $c_{i,j,k}$ cost of placing in layer i $1m^2$ of material j with type of thickness k available for layer i.
- The total thickness of the external wall is comprised between bounds $T_{\min}$ and $T_{\max}$.
- Let $U_{\max}$ be the maximum thermal transmittance allowed for the opaque part of the façade.

- Let $B_{\max}$ be the maximum budget allowed to construct $1m^2$ of the opaque part of the façade.

- Given two consecutive layers, there may exist incompatibilities between some materials and thicknesses corresponding to these layers (see examples in [7]).

- The variables of the ILP problem are the binary variables $x_{i,j,k}$ whose value are 1 if layer i is made with material j and type of thickness k, and 0 otherwise, $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,m_i\}$ and $k \in \{1,\ldots,r_{j_i}\}$.

- Given a material j, with $j \in \{1,\ldots,m_i\}$ for some $i \in \{1,\ldots,n\}$, and let λ_j be its thermal conductivity, following the calculations given in [7], the linear constraint to comply with the thermal transmittance upper bound for the opaque part of the façade is:

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{r_{j_i}} \frac{t_{i,j,k}}{\lambda_j} x_{i,j,k} \geq \frac{1}{U_{\max}} - \frac{1}{h_{int}} - \frac{1}{h_{ext}} \tag{1}$$

  Where $1/h_{ext}$ and $1/h_{int}$ (m^2KW^{-1}) represent the standard external and internal conductivity respectively for the air layers connected with the façade.

The problem of minimizing the embodied CO_2 emissions of $1m^2$ of the opaque part of a façade can be formulated mathematically as the following ILP problem, defined through Eqs. (2) to (8):

$$Minimize \quad \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{r_{j_i}} kco2_{i,j} \cdot k_{i,j,k} \cdot x_{i,j,k} \tag{2}$$

$$s.t.: \quad \sum_{j=1}^{m_i}\sum_{k=1}^{r_{j_i}} x_{i,j,k} = 1 \qquad \forall i \in \{1,\ldots,n\} \tag{3}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{r_{j_i}} c_{i,j,k} \cdot x_{i,j,k} \leq B \tag{4}$$

$$T_{\min} \leq \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{r_{j_i}} t_{i,j,k} \cdot x_{i,j,k} \leq T_{\max} \tag{5}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{r_{j_i}} \frac{t_{i,j,k}}{\lambda_j} x_{i,j,k} \geq \frac{1}{U_{\max}} - \frac{1}{h_{int}} - \frac{1}{h_{ext}} \tag{6}$$

$$x_{i,j,k} + x_{(i+1),j',k'} \quad \forall (i,j,k-(i+1),j',k') - incompatible \tag{7}$$

$$x_{i,j,k} \in \{0,1\} \quad \forall i \in \{1,\ldots,n\}, j \in \{1,\ldots,m_i\}, k \in \{1,\ldots,r_{j_i}\} \tag{8}$$

Where:

- Eq. (2) is the objective function, that is, the total embodied CO_2 emissions in kg.

- Eq. (3) ensures that each layer is made exactly of one material with a given thickness.

- Eq. (4) guarantees that the maximum budget is not exceeded.

- Eq. (5) restricts the total thickness of the façade within the established bounds.

- Eq. (6) ensures that the opaque part of the façade does not exceed the maximal allowed thermal transmittance.

- Eq. (7) forbids to place a material j' with thickness k' in the next layer to the one (layer i) containing the material j with thickness k (we denote this fact $(i, j, k-(i+1), j', k')-$ *incompatibility*). At most one of the two materials will appear in the corresponding layer.

- Finally, Eq. (8) defines the variables of the problem as binary.

Note that the above formulation contains the most usual constrains given to construct the opaque part of a façade, but it could include other types of linear constraints to fit as much as possible the real problem.

3 Case study

The case study is based on a section defined by up to five different layers according to different technological and constructive solutions and in order to generate interesting alternative scenarios. At the same time, each layer can have different thicknesses, depending on the material chosen for it. Layer 1, the internal one, may be plaster gypsum or gypsum plaster board. Layer 2 is a gap for building systems, which will not exist in case layer one be made of plaster gypsum and layer 3 be made of brick wall. Layer 3 constitutes the structural element of the wall with five different technologies: X-lam panels, wooden balloon-frame, MHM panels (Massiv-Holz-Mauer), reinforced concrete wall or brick wall. Layer 4 provides the thermal insulation, with natural or synthetic materials (projected polyurethane, extruded polystyrene, expanded polystyrene, mineral wool, expanded cork, coconut fiber, sheep wool or nanoporous gel). Finally, layer 5 is the exterior finishing, made of plaster cement and whitewash.

Taking into account the different thicknesses chosen for the different materials in each layer, a total amount of 19,065 combinations for this external wall are possible. The generic data of CO_2 emissions for this case study have been obtained from [12, 14]. In total, 70 ILP problems have been solved, depending on different allowed maximum U-values (from 0.15 up to 0.94) and thicknesses in intervals of 5 cm (from 15 up to 50 cm). Table 1 shows the amount of embodied CO_2 emissions produced by the optimal solution in each one of the 70 combinations of maximun U and thickness interval. A blank means that the problem is infeasible. We observe how we reach the minumum value for two intervals of thickness (25-30 and 30-35 cm) and for an $U \leq 0.4 Wm^{-2}K^{-1}$. The yellow part of the table is related to the maximum allowed U-values in each one of the 5 winter climate zones in Spain (from less, A, to more sever, E). *Mathematica* 11.3 [15] has been used as ILP solver in these computational experiments and it has been run on a PC Intel® $Core^{TM}$ I7-6700 with 4 processors, 3.46GHz and 8GB RAM. The CPU times to obtain each one of the optimal solutions are shown in Table 2. Note that the time to obtain any optimal solution was little than 0.08s (according to *Mathematica*'s assumption, 0.0 s means that the calculation takes no measurable CPU time), while in the 7 problems without feasible solution, *Mathematica* needed about 12s to conclude that the problem was infeasible.

Note that in this case study we have not taken into account Eq. (4), that is, our aim has been to obtain the minimum embodied CO_2 emissions independently of budged restrictions.

Obviously, if we had considered budged restrictions, the amount of embodied CO_2 emissions would be greater or equal than those obtained in this work in all cases.

	0.15	**0.20**	**0.30**	**0.40**	**0.50**	**0.57(E)**	**0.66(D)**	**0.73(C)**	**0.82(B)**	**0.94(A)**
[0.15, 0.20[			45.495	31.599	28.035	27.315	27.315	27.315	27.315	27.315
[0.20, 0.25[			36.603	28.035	27.315	27.315	27.315	27.315	27.315	27.315
[0.25, 0.30[		43.047	21.001	28.035	14.841	14.841	14.841	14.841	14.841	14.841
[0.30, 0.35[		29.601	16.281	14.841	14.841	14.841	14.841	14.841	14.841	14.841
[0.35, 0.40[		25.38	16.281	16.281	16.281	16.281	16.281	16.281	16.281	16.281
[0.40, 0.45[	34.761	20.571	18.999	18.999	18.999	18.999	18.999	18.999	18.999	18.999
[0.45, 0.50[	37.065	23.796	20.439	20.439	20.439	20.439	20.439	20.439	20.439	20.439

Table 1: Minimum embodied kg of CO_2 in $1m^2$ of the external wall given an interval of thickness (row) and a maximal allowed thermal transmittance (column).

	0.15	**0.20**	**0.30**	**0.40**	**0.50**	**0.57**	**0.66**	**0.73**	**0.82**	**0.94**
[0.15, 0.20[	11.28125	12.1875	0.01562	0.0	0.01562	0.0	0.01562	0.01562	0.01562	0.0
[0.20, 0.25[	11.35937	12.17187	0.03125	0.03125	0.01562	0.01562	0.0	0.0	0.01562	0.0
[0.25, 0.30[	11.89062	0.03125	0.03125	0.01562	0.0	0.0	0.01562	0.01562	0.0	0.01562
[0.30, 0.35[	12.04687	0.0625	0.01562	0.01562	0.01562	0.01562	0.0	0.01562	0.0	0.01562
[0.35, 0.40[	11.46875	0.078125	0.01562	0.0	0.0	0.0	0.01562	0.01562	0.01562	0.0
[0.40, 0.45[	0.01562	0.0	0.03125	0.03125	0.03125	0.01562	0.01562	0.01562	0.01562	0.01562
[0.45, 0.50[	0.0	0.03125	0.03125	0.01562	0.01562	0.01562	0.01562	0.01562	0.01562	0.0

Table 2: Time in seconds to find the optimal solutions shown in Table 1, given an interval of thickness (row) and a maximal allowed thermal transmittance (column).

Table 3 shows all data, including material and thickness of each layer, corresponding to the optimal solution with the lowest embodied CO_2 optimal solution for 3 different scenarios. The cells containing these solutions are emphasized with blue color in Table 1. Note that in case of a tie the optimal solution with the lowest U-value has been chosen. The three scenarios are:

- The lowest embodied CO_2 among all the optimal solutions.
- The lowest embodied CO_2 among the optimal solutions in the minimum thickness interval.
- The lowest embodied CO_2 among the optimal solutions with the minimum maximal allowed U-value.

We observe how the ballon frame solution for layer 3 is most represented as well as the cork as insulating material for layer 4. The minumum enbodied CO_2 emission (14.841 $kg\ CO_2/m^2$) can be reached for U-values allowed in all Spanish climate zones (A to E). The lowest U-value of 0.15 $Wm^{-2}K^{-1}$ is reached with balloon frame and projected polyurethane but a large thickness of 44.25 cm. Finally, in order to reduce thickness (18.25 cm) we have more embodied CO_2 and a higher U-value.

Scenario	Lowest CO_2	Lowest CO_2 for the lowest thickness	Lowest CO_2 for the lowest U-value
Thickness (cm)	33.25	18.25	44.25
U-value ($Wm^{-2}K^{-1}$)	0.330	0.551	0.147
Embodied $kgCO_2/m^2$	14.841	27.315	34.761
Layer 1	Gypsum plaster board (1.25 cm)	Gypsum plaster board (1.25 cm)	Gypsum plaster board (1.25 cm)
Layer 2	LV air gap (10 cm)	LV air gap (5 cm)	LV air gap (10 cm)
Layer 3	Balloon Frame (20 cm)	X-lam (10 cm)	Balloon Frame (20 cm)
Layer 4	Cork (1cm)	Cork (1 cm)	Projected polyurethane (12 cm)
Layer 5	Plaster (1 cm)	Plaster (1 cm)	Plaster (1 cm)

Table 3: Data of the best solutions for three diferent scenarios.

4 Conclusions

The ILP shows it applicability also for the problem of embodied energy in building with the possibility to compare different constructive solutions and to give more elements for the final choise taking into account environmental aspects.

From a constructive point of view, the balloon frame (layer 3 of our case study) is for its lightness the most represented constructive solution, beeing all woden solutions (balloon frame, MHM and X-lam) the less CO_2 consuming.

Othe trends are related with the importance of the air gap as CO_2 neutral layer but with a great influence in the thermal behaviour of the façade.

References

[1] Energy in Figures – Statistical Pocketbook 2015 edition, Brussels: European Comision.

[2] European Parliament, Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings, in Directive 2010/31/EU. 2010: Brussels.

[3] Privitera, G., Dayb, A.R., Dhesic, G. and Long, D., Optimising the installation costs of renewable energy technologies in buildings: A Linear Programming approach, *Energy Build.* 43: 838-843, 2011.

[4] Ashouri, A., Fux, S.S., Benz, M.J. and Guzzella, L., Optimal design and operation of building services using mixed-integer linear programming techniques, *Energy*, 59: 365-376, 2013.

[5] Lindberg, K.B., Doorman, G., Fischerc, D., Korpås, M., Ånestad, A. and Sartori, I., Methodology for optimal energy system design of Zero Energy Buildings using mixed-integer linear programming, *Energy Build.* 127: 194-205, 2016.

[6] Ogunjuyigbe, A.S.O., Ayodele, T.R. and Oladimeji, O.E., Management of loads in residential buildings installed with PV system under intermittent solar irradiation using mixed integer linear programming, *Energy Build.* 130: 253-271, 2016.

[7] Soler, D., Salandin, A. and Micó, J.C., Lowest thermal transmittance of an external wall under budget, material and thickness restrictions: An Integer Linear Programming approach, *Energy Build*, 158: 222–233, 2018.

[8] Salandin, A. and Soler, D., Computing the minimum construction cost of a building's external wall taking into account its energy efficiency,*J. Comput. Appl. Math.*, 338: 199–211, 2018.

[9] McMullan, R., *Environmental Science in Building*, Palgrave Macmillan, Basingtoke, 2012.

[10] Lolli, N., Fufa, S.M. and Inman, M., A parametric tool for the assessment of operational energy use, embodied energy and embodied material emissions in building, *Energy Procedia*, 111: 21-30, 2017.

[11] Ahmad Faiz Abd, R. and Sumiani, Y., A review of life cycle assessment method for building industry, *Renew. Sust. Energ. Rev.*, 45: 244–248, 2015.

[12] Bath Inventory of Carbon and Energy (ICE), U.o.B.S.E.R. Team, Editor. 2010.

[13] Ecoinvent, Ecoinvent database version 3.1., E. Centre, Editor. 2013: www.ecoinvent.org, Zurich, Switzerland.

[14] able of embodied energy or primary energy of materials. http://www.tectonica-online.com/topics/energy/embodied-energy-materials-enrique-azpilicueta/table/31/

[15] Wolfram, Mathematica, http://www.wolfram.com/mathematica

38

Acoustics on the Poincaré Disk

Michael M. Tung [♭1]

(♭) Instituto de Matemática Multidisciplinar,
Universitat Politècnica de València.

1 Introduction

The Poincaré disk model is a straightforward model of hyperbolic geometry [1] on the 2-dimensional disk taking over certain properties from the Poincaré half-plane [2, 3]. The corresponding metric of this manifold weighs radial distances from the center of the disk to its circumference in a characteristic manner. Not only from the mathematical viewpoint has this model fundamental relevance, but we will argue that its prominent feature can be interpreted as somewhat that of a black hole turned upside down, a quality absent from the related half-plane analogue [4]. This particular geometry and topology make it an attractive candidate for acoustic wave simulation with metamaterial devices—devices composed of extraordinary materials which allow to implement curved background spacetimes in acoustics, see [5–9] and references therein.

In this work we study the feasibility to implement acoustics on the Poincaré disk and investigate the wave propagation in such a medium. We explore the main differential-geometric features of this spacetime with its asymptotic behaviour and causal boundaries very much alike to those of acoustic black holes [9]. By employing the framework developed in [6, 7] we find the acoustic laboratory parameters (mass-density tensor and bulk modulus) corresponding to the underlying spacetime structure. We also derive the equations of motion which govern acoustic wave propagation on the Poincaré disk. This work concludes with numerical simulations for one illustrative example.

2 Spacetime geometry

The Poincaré disk, henceforth denoted by $\mathbb{D}_{\mathrm{P}}$, is the resulting image of the stereographic projection $(X, Y, Z) \mapsto (x, y)$ of the upper part of a circular hyperboloid of two sheets, represented by the equation $X^2 + Y^2 - Z^2 = -a^2$, onto the xy-plane.

Fig. 1 provides a schematic view of the stereographic mapping, and shows the similar triangles to yield the following relations between the coordinates of the hyperboloid and its projection

[1] e-mail: mtung@imm.upv.es

to the plane, *i.e.*

$$\frac{x}{a} = \frac{X}{Z+a}, \quad \frac{y}{a} = \frac{Y}{Z+a}. \tag{1}$$

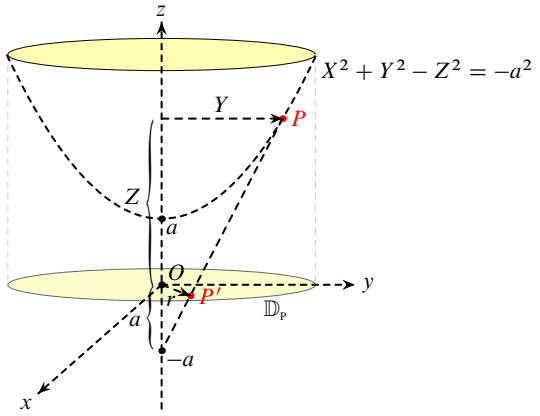


Figure 1: Schematic view of the stereographic projection of point P on the upper sheet of a circular hyperboloid to point P' located on the plane $z = 0$. The projection gives the Poincaré disk, $\mathbb{D}_{\mathrm{P}}$, endowed with a characteristic metric.

Using the conventional radial polar coordinate r on the xy-plane, it is not difficult to find

$$\left.\begin{aligned} X &= \frac{2x}{1 - r^2/a^2} \\ Y &= \frac{2y}{1 - r^2/a^2} \\ Z &= \frac{1 + r^2/a^2}{1 - r^2/a^2} a \end{aligned}\right\} \quad \text{for} \quad 0 \le r < a, \tag{2}$$

This induces the following spatial line element for distances ℓ on Poincaré disk $\mathbb{D}_{\mathrm{P}}$:

$$d\ell^2 = dX^2 + dY^2 = 4\frac{dx^2 + dy^2}{\left(1 - r^2/a^2\right)^2}. \tag{3}$$

Now, it is customary to introduce the geodesic radius $\varrho = \operatorname{artanh}(r/a)$, which converts (3) to the much simpler form

$$d\ell^2 = a^2 d\varrho^2 + a^2 \sinh^2\varrho \, d\varphi^2. \tag{4}$$

Note that (4) represents the metric of hyperbolic geometry underlying much of the famous artwork by the Dutch artist M.C. Escher [10]. Then, adding the time component, the full spacetime metric for the Lorentzian manifold $M = \mathbb{D}_{\mathrm{P}} \times \mathbb{R}$ is given by

$$\boldsymbol{g} = -\big(c\,dt\big) \otimes \underbrace{\big(c\,dt\big)}_{\theta^0} + \big(a\,d\varrho\big) \otimes \underbrace{\big(a\,d\varrho\big)}_{\theta^1} + \big(a \sinh\varrho\, d\varphi\big) \otimes \underbrace{\big(a \sinh\varrho\, d\varphi\big)}_{\theta^2}, \tag{5}$$

where θ^μ ($\mu = 0, 1, 2$) indicates the dual-base forms of the local coframe. Recall that $a > 0$ is a physical length scale, further ϱ and φ are geodesic polar coordinates. Moreover, $c > 0$ is a constant speed. This completes the spacetime setup necessary for the succeeding wave simulation.

3 Physical construction

In combination with Cartan's structure equations, the dual base $\{\theta^0, \theta^1, \theta^2\}$ introduced by (5) allows to straightforwardly compute the Riemann curvature tensor $\hat{R}^{\alpha}{}_{\beta\gamma\delta}$ in the coframe of manifold $M = \mathbb{D}_{\mathrm{P}} \times \mathbb{R}$. The only non-zero and independent component turns out to be

$$\hat{R}^{1}{}_{212} = -\frac{1}{a^2} \quad \Rightarrow \quad G_{00} = \hat{G}_{00} = -\frac{1}{a^2}. \tag{6}$$

In the final step of (6), we have given the corresponding Einstein tensor $\hat{G}_{\alpha\beta}$, whose 00-components in the local coframe and coordinate frame are identical. As an immediate consequence of (6) the underlying energy-matter density ρ_0 is exotic:

$$\underbrace{G_{00}}_{-1/a^2} = \frac{8\pi G}{c^4} \underbrace{T_{00}}_{\rho_0 c^2} \quad \Rightarrow \quad \rho_0 = -\frac{c^2}{8\pi G a^2} < 0, \tag{7}$$

where G is the usual gravitational constant. As expected, if the disk extends to infinity, *viz.* $a \to \infty$, the energy-matter content will vanish and M becomes asymptotically flat.

Obviously, current—and likely any future technology—does not permit to implement such physical configuration with negative energy-matter density. However, fine-tuning the acoustic parameters of a suitable metamaterial will presumably allow to do so in the near future. In Ref. [7], we have shown that there exists a general 1-to-1 correspondence between spacetime metric $\mathbf{g}$ and the parameters κ (bulk modulus) and $\boldsymbol{\rho}$ (density tensor) of an acoustic metamaterial device. In this case, using (5), we obtain for $0 \leq r < a$:

$$\kappa = \frac{4}{\left(1 - r^2/a^2\right)^2} \kappa_0, \quad \rho_0 \rho^{ij} = \tfrac{1}{4} \left(1 - r^2/a^2\right)^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{8}$$

Note that the constants κ_0 and ρ_0 are fixed by the physical properties of the corresponding flat space. This completes the implementation of the Poincaré disk $M = \mathbb{D}_{\mathrm{P}} \times \mathbb{R}$ for metamaterial acoustics.

4 Wave simulation

Once the acoustic metamaterial is configured, wave propagation can be simulated in such media. Acoustic phenomena will be governed by an elementary variational principle, namely that for a spacetime M, endowed with metric $\boldsymbol{g}$, the action will be stationary with respect to variations of the acoustic potential $\phi : M \to \mathbb{R}$, such that integration over bounded spacetime domain $\Omega \subseteq M$ with volume element $d\mathrm{vol}_g$ satisfies [7]:

$$\frac{\delta}{\delta\phi} \int_{\Omega} d\mathrm{vol}_g \, \boldsymbol{g}(\nabla\phi, \nabla\phi) = 0. \tag{9}$$

Next, the variational principle, (9), amounts to solving

$$*d*d\phi = 0, \tag{10}$$

where $*$ is the Hodge dual. In local coordinates, (10) takes the form of the wave equation with the Laplace-Beltrami Δ_M operator for the curved spacetime background M.

To take advantage of the underlying rotational symmetry, we choose concentric waves centered around the origin for acoustic probing. Then, all non-trivial behaviour of the acoustic potential ϕ is contained in a radial factor, which we denote by $\phi_1(\varrho)$. Furthermore, considering rotational symmetry, we can show that the exact solutions for $\phi_1(\varrho)$, satisfying (10), are combinations of Legendre polynomials with complex arguments.

Apart from the exact result, we have derived a relatively simple and accurate approximate solution for the radial dependence of the potential:

$$\phi_1(\varrho) = e^{-\varrho/2}\left[Ae^{\sqrt{1-4a^2\omega^2/c^2}\,\varrho} + Be^{-\sqrt{1-4a^2\omega^2/c^2}\,\varrho}\right]. \tag{11}$$

Here, ω specifies the frequency of the monochromatic sound waves, and A and B are constants determined by the boundary conditions. Extreme damping occurs in the asymptotic limit $\varrho \to \infty$ ($r \to a$), and consequently it will never be possible to escape $\mathbb{D}_{\mathrm{P}}$. Moreover, oscillatory behaviour emerges when $\omega > \frac{c}{2a}$. For a numerical simulation, we assume $a = 1$, $c = 1$, and $\omega = 1 > 1/2$, so that naturally harmonic wave features will materialize. Fig. 2 captures exact and approximate results for the boundary conditions $\phi_1(1) = 1$ and $\phi_1'(1) = 0$. The absolute error is exceptionally good and ranges between $2.65 \cdot 10^{-8}$ and 0.018 (only close to $\varrho = 3.4$).

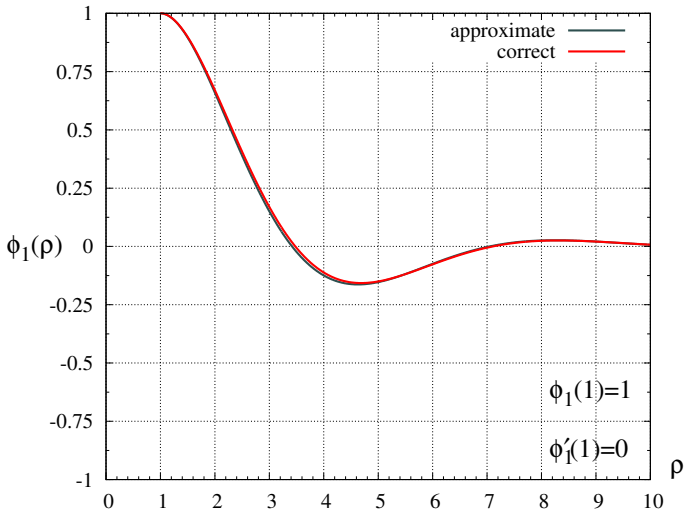


Figure 2: Non-trivial radial dependence of the acoustic potential for scale $a = 1$, speed $c = 1$, frequency $\omega = 1$, with conditions $\phi_1(1) = 1$, $\phi_1'(1) = 0$.

References

[1] Anderson, J.W., *Hyperbolic Geometry.* Berlin, Springer-Verlag, 2005.

[2] Kelly, P.J. and Matthews, G., *The Non-Euclidean Hyperbolic Plane: Its Structure and Consistency.* Berlin, Springer-Verlag, 1981.

[3] Stahl, S., *A Gateway to Modern Geometry: The Poincaré Half-Plane*. Burlington, MD, Jones & Bartlett, 2008.

[4] Tung, M.M., Modelling acoustics on the Poincaré half-plane, *J. Comput. Appl. Math.*, 337: 336–372, 2018.

[5] Cummer, S., A sound future for acoustic metamaterials, *J. Acoust. Soc. Am.*, 141(5): 3451, 2017.

[6] Tung, M.M. and Peinado, J., A covariant spacetime approach to transformation acoustics. *Progress in Industrial Mathematics at ECMI 2012*, M. Fontes, M. Günther, N. Marheineke, (eds.), Mathematics in Industry, 19: 335–340, 2014.

[7] Tung, M.M., A fundamental Lagrangian approach to transformation acoustics and spherical spacetime cloaking, *Europhys. Lett.*, 98: 34002–34006, 2012.

[8] Tung, M.M. and Weinmüller, E.B., Gravitational frequency shifts in transformation acoustics, *Europhys. Lett.*, 101: 54006–54011, 2013.

[9] Tung, M.M. and Weinmüller, E.B., Acoustic metamaterial models on the (2+1)D Schwarzschild plane, *J. Comput. Appl. Math.*, 346: 162–170, 2019.

[10] Schattschneider, D., The mathematical side of M.C. Escher, *Notices Am. Math. Soc.*, 57(6): 706–718, 2010.

39

Network computational model to estimate the effectiveness of the influenza vaccine *a posteriori*

D. Martínez-Rodríguez[♭], R. San Julián-Garcés[♭], E. López-Navarro[♭] and R.J. Villanueva[♭1]

(♭) Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València.

1 Introduction

Influenza or seasonal influenza, known as the flu, is an infectious disease caused by two different virus, which usually belong to the *Orthomyxoviridae* family. It affects mainly the respiratory system, although it can also affect the circulatory and muscular system. The symptoms that produce this virus can be diverse, from high fever, headache and throat irritation to muscle aches and feeling tired [1].

The first records of this disease date from ancient Egypt, and from then until now this disease has been one of the most complex to study and treat due to the high mutations of different nature that the virus suffers every year, making it highly unpredictable. This virus appears every year in the temperate zones of the whole globe, where the difference in the average temperature between the summer and winter is high [1]. These climatic factors, together with the contact between individuals in a population, favor the spread of the disease rapidly. Annually, about three to five million people worldwide are infected by the influenza virus and about 250 to 500 thousand die.

Every year the World Health Organization (WHO) predicts which strains of the influenza virus are most likely to circulate among the population. Because of this, a new vaccine should be developed every season that is capable of immunizing as many individuals as possible. It takes about six months to formulate and produce the required doses, but the vaccine must be ready before the flu outbreak. It is not possible to analyze its effectiveness before its production although it is possible to do it *a posteriori*.

The time between the start of the vaccine development process and its release to the market, as well as the lack of real data (such as vaccine coverage) makes necessary to calculate the effectiveness of the vaccine once the flu season has ended, which usually happens between October and April in the northern hemisphere. In addition, the partial immunization of some people in

[1]e-mail: rjvillan@imm.upv.es

the population who have previously had similar flu periods and the transmission of antibodies generated by the parents to the next generation are added to these problems.

The World Health Organization (WHO) strongly recommends to the population to be vaccinated every year against the flu, in order to prevent the possibility of infection and to reduce the number of infected people [2].

2 Computational network model

Until now, several techniques have been proposed to determine the effectiveness of the influenza vaccine *a posteriori*, although their reliability is currently under discussion. Many of the proposed techniques are based on statistical analysis, however we propose a new and efficient technique based on a computer network model capable of representing the spread of flu in a population.

The network model consists of a graph formed by vertices and edges. Each of the vertices corresponds to an individual of the population and each edge represents the contact (effective or not) of the transmission of the disease between the vertices that join the edge. It is possible to simulate in a network the evolution of the transmission dynamics of an infectious disease such as influenza over time using computer programs.

The generated network has been designed in a flexible way, which allows to introduce specific vaccine strategies and change them if necessary in a quick and simple way. This model builds a population made up of one million individuals, where the total number of relationships changes depending on an average degree. These relationships are generated randomly, in the same way as individuals, through an algorithm for obtaining random numbers. The age of the individuals of the population follows the demographic data of the Community of Valencia.

The data used represent the weekly reported cases over a period of 26 weeks between October 2016 and January 2017. The data includes 95% of the confidence intervals.

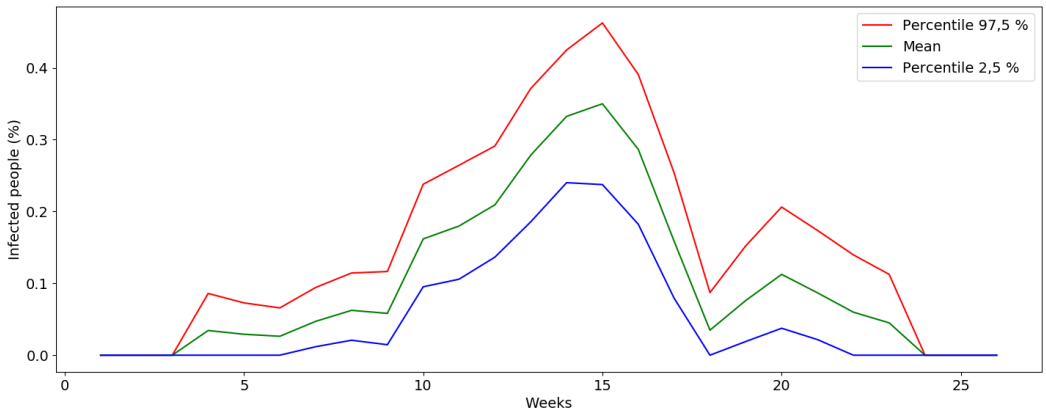


Figure 1: Weekly flu reported cases 2016-2017 [3].

To simulate the flu transmission dynamics over time, the model behaves in the same way as in reality. Each individual can go through four different states:

- Susceptible: The individual is healthy and can be infected by infected neighbors.
- Vaccinated: The individual has been effectively vaccinated, that is, the individual is protected against infection.
- Infected: The individual is infected. After a week the individual recovers.
- Recovered: The individual has passed the flu and is no longer infected.

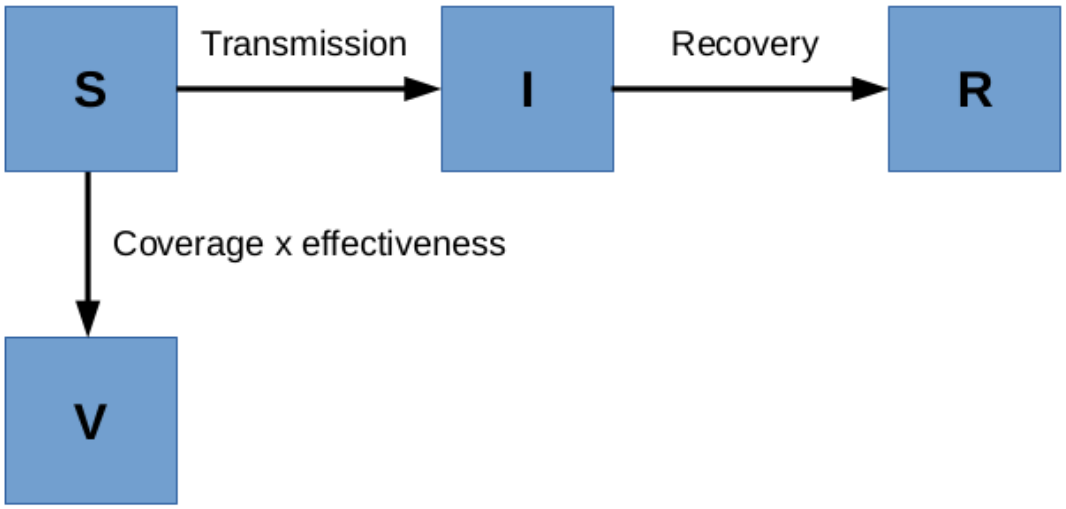


Figure 2: Flow diagram of the influenza transmission dynamics.

A part of the population that is vaccinated is not protected from the flu. This is because the vaccine is not 100% effective and there are cases where it can take effect and others where it has no effect. In cases where it does have effect, people are protected against the flu of the current season. Otherwise, they remain susceptible.

In the network model there are two types of parameters to work with. The unknown parameters, which are the average degree of the computer network, the weekly transmission rate (different for each week) and the effectiveness of the vaccine. The known parameters are the coverage of the vaccine, that is, how many people in the total population should be vaccinated and the recovery time.

3 Model calibration

To calibrate the unknown parameters of the model, we repeatedly applied an optimization algorithm (PSO) with an error function that measures the difference of the model and the confidence intervals of the data. This task is evaluated 50 thousand times.

Following, we select one hundred sets of input parameters of the model in such a way that the 95% confidence interval of the outputs should be as close as possible to the 95% confidence interval of the real data.

4 Results

After selecting the one hundred outputs that best capture the uncertainty of the data, the obtained results can be seen in Fig. 3 and 4:

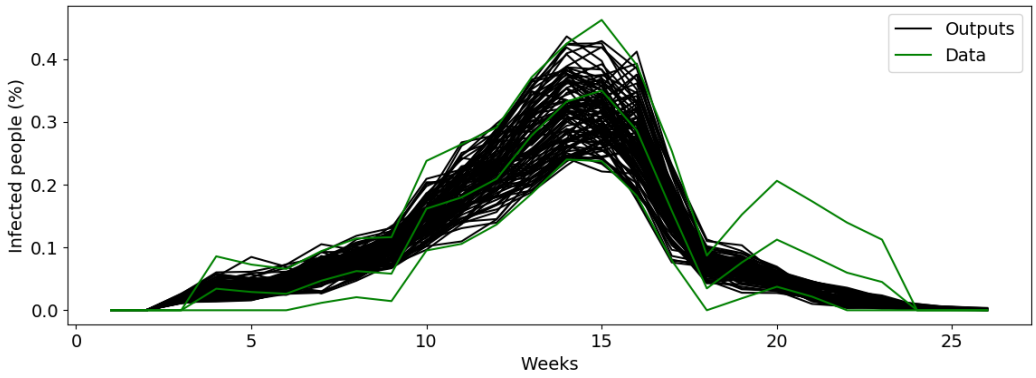


Figure 3: Hundred best model outputs (in black). In green, the data and the 95% CI.

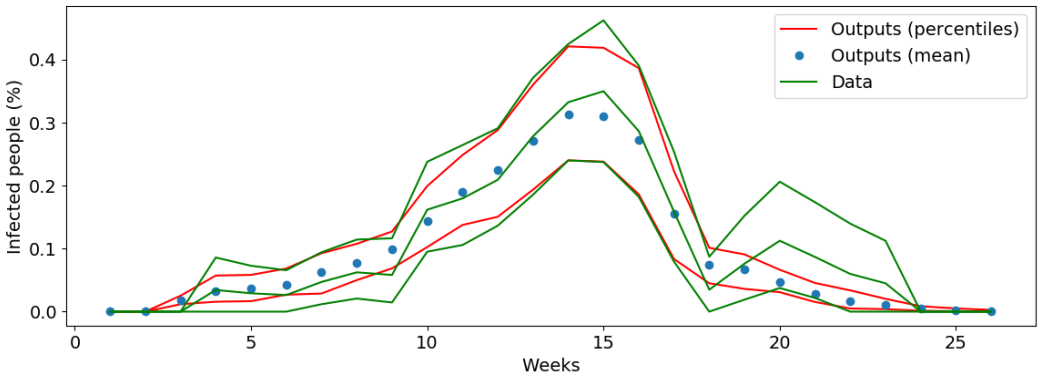


Figure 4: The same figure as Fig. 3, but instead of the 100 model outputs, its 95% CI of the hundred best model outputs.

As we can see in Fig. 3 the one hundred outputs are the black lines and the real data are the green lines. The outputs capture the uncertainty of the data in most of the band. In Fig. 4 the difference with the Fig. 3 is the black lines. We represent the percentiles and the mean of the outputs and the real data (the mean and the percentiles).

The calibration gives us values for the unknown parameters:

	Percentile 2,5%	**Mean**	**Percentile 97,5%**
Average degree	43,47	59,18	67,52
Vaccine effectiveness	14%	29%	57%

Table 1

In Table 1 we can see the values calibrated for the average degree and the vaccine effectiveness. Also, in Fig. 5 we can see how the transmission rate of influenza evolves over the disease season.

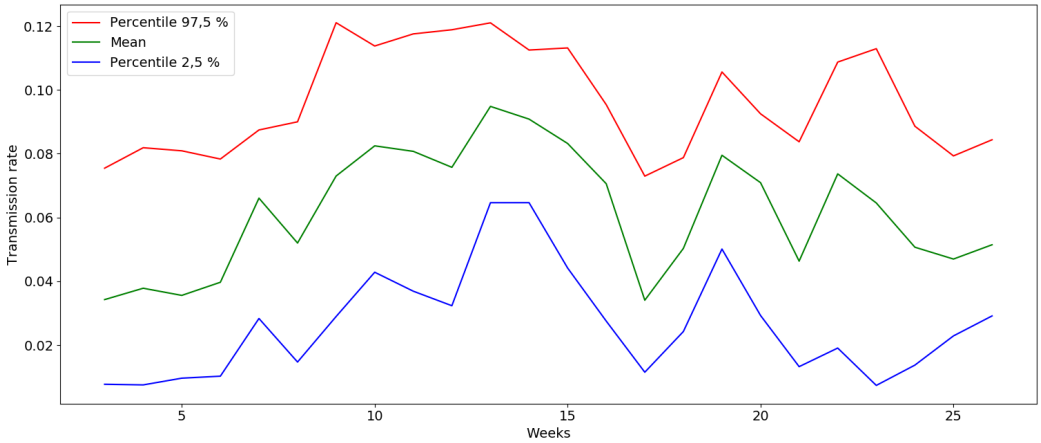


Figure 5: 95% CI of weekly transmission rate over the disease season.

5 Conclusions

We have used a novel technique based on a computational network model to determine the effectiveness of the flu vaccine in a population of individuals.

The techniques used to perform the calibration have taken into account the data and the model building uncertainty.

The estimated effectiveness of the influenza vaccine is low, about 30%.

We will consider to divide the population in to age groups because the coverage and the effectiveness of the vaccine change with the age of the individual.

Acknowledgments

This paper has been supported by the European Union through the Operational Program of the [European Regional Development Fund (ERDF) / European Social Fund (ESF)] of the Valencian Community 2014-2020.

Esta actuación está cofinanciada por la Unión Europea a través del Programa Operativo del [Fondo Europeo de Desarrollo Regional (FEDER) / Fondo Social Europeo (FSE)] de la Comunidad Valenciana 2014-2020.

Aquesta actuació està cofinançada per la Unió Europea a través del Programa Operatiu del [Fons Europeu de Desenvolupament Regional (FEDER) / Fons Social Europeu (FSE)] de la

Comunitat Valenciana 2014-2020.

Records: GJIDI/2018/A/009 and GJIDI/2018/A/010.

References

[1] Moghadami, M., A Narrative Review of Influenza: A Seasonal and Pandemic Disease, Iranian Journal of Medical Sciences, 42 (1): 2–13, 2017.

[2] Recommended composition of influenza virus vaccines for use in the 2006-2007 influenza season, World Health Organization, 2006. https://web.archive.org/web/20060528111237/http://www.who.int/csr/disease/influenza/2007northreport.pdf.

[3] Portero, A., Alguacil A., Sanchis, A., Pastor, E., López, A., Miralles, M., Martin, M., Alberich, C., Roig, F., Lluch, J., Paredes, J. and Vanaclocha, H., Prevención y vigilancia de la gripe en la Comunitat Valenciana. Temporada 2016-2017, 150: 28, Valencia, 2017.