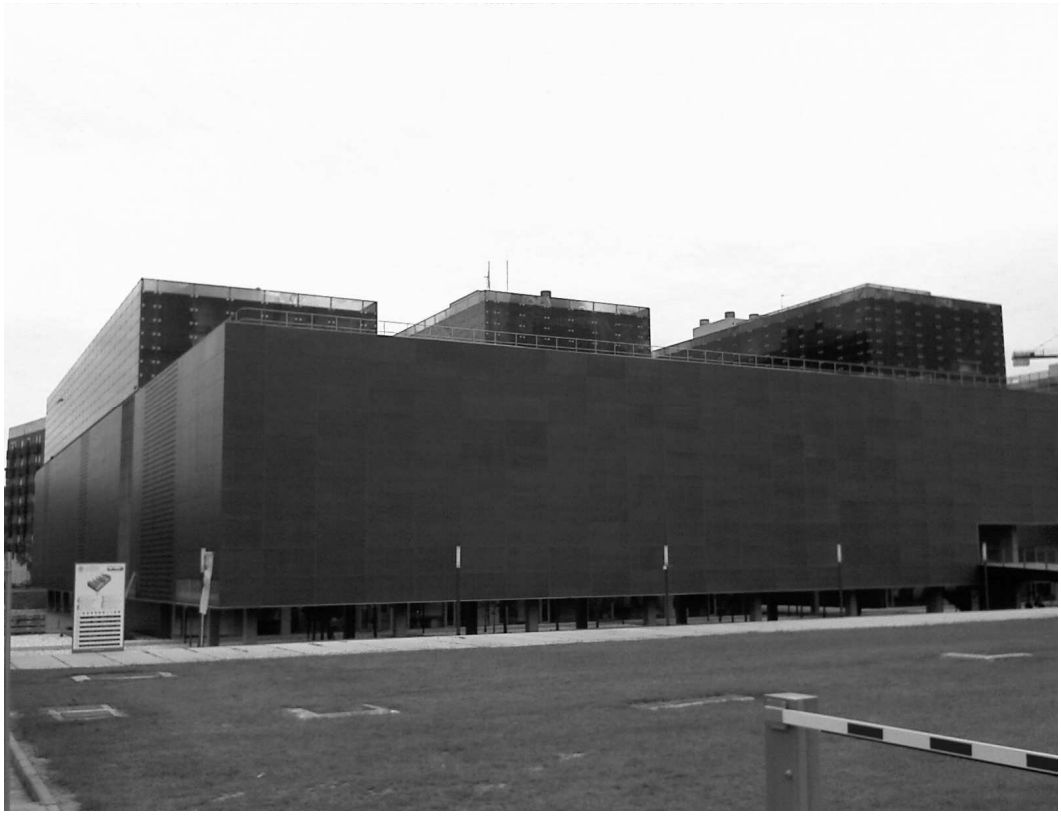
# MODELLING FOR ENGINEERING AND MEDICINE 2008

*Instituto de Matemática Multidisciplinar*



**Edited by: L. Jódar**

im² Instituto de Matemática Multidisciplinar

UNIVERSIDAD POLITECNICA DE VALENCIA

GENERALITAT VALENCIANA

# MODELLING FOR ENGINEERING AND MEDICINE 2008

Instituto de Matemática Multidisciplinar
Universidad Politécnica de Valencia
Valencia 46022, SPAIN

Edited by
Lucas Jódar,
Instituto de Matemática Multidisciplinar, Director
I.S.B.N.: 978-84-691-8345-8

# CONTENTS

# Neural Networks as Cellular Automata

L. Acedo*

Instituto de Matemática Multidisciplinar

Universidad Politécnica de Valencia, Edificio 8G, 2°

Camino de Vera, 46022 Valencia, España

December 11, 2008

## 1 Introduction

Mathematical models for the dynamical evolution of a set of simple Boolean (or multiple states) automata according to some defined deterministic or stochastic rules have been an area of steady interest since the seminal works of Wolfram [1]. Particularly fecund are the applications to mathematical biology: epidemic spread [2], pattern formation and developmental biology [3], or forest-fire models [4]. The last group of models are also generally known as excitable media and they could be traced back to the proposal of Greenberg and Hastings [5]. These models have been also successfully applied to the propagation and breaking of spiral waves in the heart tissue [6]. In sharp contrast with the unremitting effort in so many areas of mathematical biology, the applications of cellular automata to neural networks and the brain are relatively scarce and mostly concentrated on some aspects of brain function, such as memory [7], instead of the global modeling of neural dynamics [8, 9]. In this paper we revisit a stochastic cellular automaton in a complete graph that have been recently proposed as a rough model of the brain cortex [10]. In the original version of the model we only considered the counterpart of excitatory neurons whose states could be firing or quiescent. The transitions from the firing state to the resting state ,or viceversa, are controlled by two

*e-mail: luiacrod@imm.upv.es

probabilities: the probability that a firing neuron excites one of its quiescent neighbours into the firing state, $\alpha$, and the rate of spontaneous deactivation of a firing neuron, $\beta$. Time is discrete in the model so both probabilities are defined per unit time.

The original version of our model only considered Boolean automata at every site of a complete graph corresponding to excitatory neurons. It is a well-known fact that about a 20 per cent of cortical neurons are inhibitory [11], their primary role being to avoid the overexcitation of the cortex. Consequently, in this paper we generalize our stochastic Boolean cellular automaton by including a fraction of inhibitory sites in the complete graph. If the automata at these sites are in the firing state, they can inhibit any of its firing neighbours with a probability $\gamma$ per time-step. We also observe a second-order phase transition for this generalized model but the variance exhibits a richer dependence with $\alpha$. Moreover, the average number of firing neurons is reduced in the excitatory subnetwork as expected from the interaction with the inhibitory sites but, remarkably, the critical point, $\alpha_c$, keeps the same and does not depend on $\gamma$. So, a nice first prediction of the excitatory-inhibitory model is that inhibitory neurons could reduce the excitation of the cortex, as neurophysiology tell us, but they can never prevent the cortex from working, i.e., they never reduce the excitation to zero. As inhibition is taking into account, we find that fluctuations are larger than in purely excitatory networks. The distribution of these fluctuations are gaussian as expected and they are approached subdiffusively. Subdiffusive behaviour is also found in disordered systems and fractals [12] and arise from the random walk of particles in the convoluted structure of these substrates [13]. In our case, the mean-square displacement reaches a plateau and we could parallel this behaviour with the subdiffusion of a particle in a closed domain. The comparison of the mean-square displacement obtained from large sequences of EEG will confirm that these signals also manifest this kind of confinement. This will strenghten our thesis that EEG is, to some extent, a statistical epiphenomenon derived from the collective interactions of large networks of neurons.

## 2 Mean-Field Equations and Fluctuations

Our model is defined as a set of $N + M$ Boolean automata in a complete graph. These automata are placed upon the vertices of the graph with the edges mimicking neural synapses among them. The complete graph topology

have already been used both in comparative neuroanatomy [14] and neural networks [7] and could be a reasonable description of compartments containing the square root of the total number of neurons in the brain [14]. As suggested by the partition we have made, the number $N$ represents the excitatory automata neurons and $M < N$ are the inhibitory ones. Both kind of neurons are found in any of two states: firing or resting. Time is assumed to be discrete and dynamical evolution proceeds according to the following stochastic rules: (i) The transition resting $\rightarrow$ firing takes place with a probability $\alpha$ per unit time if the resting neuron is connected with a single firing excitatory neuron. Taking into account that all sites are connected, and assuming statistical independence of the influences exerted by the excitatory firing neurons on the resting one, we have that the probability for a given quiescent neuron to become firing is $1-(1-\alpha)^E$, where $E$ is the number of firing excitatory neurons in the previous time-step. (ii) The transition firing $\rightarrow$ resting happens in two ways: spontaneously (with a probability $\beta$) or by the influence of the firing inhibitory neurons with a probability $\gamma$ per unit time. So, the global probability for this transition is $1-(1-\gamma)^I(1-\beta)$, where $I$ is the total number of inhibitory automata in the firing state. These rules are, of course, somewhat arbitrary but they have been chosen as the simpler way to incorporate inhibition into the model.

We denote as $E(t)$, $I(t)$ the number of excitatory and inhibitory automata in the firing state at time step $t$, respectively. The rules lead easily to the following Markovian evolution equations:

$$\begin{aligned} E(t+1) &= E(t) + \left(1-(1-\alpha)^{E(t)}\right)(N-E(t)) \\ &- \left(1-(1-\gamma)^{I(t)}(1-\beta)\right)E(t) \\ I(t+1) &= I(t) + \left(1-(1-\alpha)^{E(t)}\right)(M-I(t)) \\ &- \left(1-(1-\gamma)^{I(t)}(1-\beta)\right)I(t)\,. \end{aligned} \tag{1}$$

We find a critical value of the excitatory probability, $\alpha_c$, such that for $\alpha > \alpha_c$ a nontrivial stable fixed point for the discrete system in Eq. (1) is found:

$$I_0 = E_0 \simeq \frac{\alpha-\alpha_c}{\alpha N + M\gamma}\,, \quad \alpha > \alpha_c = \beta/N\,, \tag{2}$$

where we have approximated $\alpha_c$ for small $\beta$. We must notice that in order to avoid probability overcounting (neurons excited several times at the same time-step) one must derive Eq. (1) from a master equation of the system as a whole [15]. This approach implies that $\alpha_c = \beta/N$ but it is not too a drastic change if transition probabilities are small.

We have also studied the fluctuations in the number of firing neurons in this model with the objective of finding a parallelism with EEG. Electroencephalography or EEG is a non-invasive technique with a long tradition in clinical neurology in which brain electrical activity is recorded by means of a set of electrodes attached to the scalp with adhesive materials [16]. In this technique the tiny voltage differences between pair of electrodes are amplified and registered with a temporal resolution $\sim$ 4ms. Striking differences are found among the EEG of an awaken normal subject (we have alpha rhythm in this case which is characterized by its low amplitude and frequencies in the range 8-12 Hz) and that of deep sleep, also known as sleep stage 4 (delta waves, approximately two times the amplitude of alpha rhythm and with average frequencies in the range of 2 Hz). For the reason of its small frequency, the deepest phase of sleep also receives the name of slow wave sleep. It is generally accepted that some subcortical structures in the brain, such as the thalamus of the brainstem, could play the role of pacemakers for some rhythms. However, EEG is far from being periodic. Moreover, EEG is the most complex physiological signal and it does not seem strange odd to hypothesize a statistical origin for the EEG background. This would imply that general statistical properties of the fluctuations in network models could be mapped to some aspects of the real EEG. To test this idea, we will establish a correspondence between the time-dependent variance of the fluctuations in the excitatory subnetwork of the model and the mean-square displacement of cortical EEG signals as follows:

$$\begin{aligned} \left\langle \left(V(t)-V_0\right)^2 \right\rangle &= N^2\rho^2\sigma^2(t) = \frac{N\beta}{\beta+M\gamma}\rho^2\left(1-e^{-t/\tau}\right) , \\ \tau &= \frac{1}{2N\alpha-\beta} , \end{aligned} \tag{3}$$

where $\rho$ denotes the contribution of a single neuron to the global electric potential integrated at the electrodes in the EEG protocol. In order to test the validity of this model we have studied two EEG records from the same subject [18]: One of them corresponding to the waking state and consisting of six minutes of alpha rhythm sampled every 4 ms and the other consisting of a four minutes record of delta rhythm. These records were divided in portions of 1 second each and the mean-square displacement from the initial value of the voltage was calculated for every portion and, finally, averaged over all portions. This statistical analysis show a very good agreement with the predictions of Box-Lucas model [17] in Eq. (3) and we conclude that Neural

networks cellular automata could be relevant in realistic neural dynamics of assemblies of neurons in the brain.

# 3 Concluding remarks

In this paper we have studied a cellular automata with a complete graph topology and stochastic rules as a simple mathematical model for neural dynamics. We have generalized a previous version of the model, in which only excitatory neurons were considered, to take into account a population of inhibitory neural automata. Our proposal is perhaps the most simple, but nontrivial, cellular automata incorporating the minimum requirements to display interesting behaviours and mimic, to some extent, the collective effects of the dynamics of the real cortex: Boolean automata occupy the vertices of the graph, the edges of this graph are analogous to the synapses, each neuron receives inputs from their excitatory and inhibitory neighbours and changes its state from resting to firing with a probability $\alpha$ for every connection with an excitatory firing neuron. Firing neurons become quiescent with probability $\gamma$ for every firing inhibitory neuron. Refractoriness is simulated by a probability rate $\beta$ for the spontaneous deactivation of firing neurons. The complete graph topology has been suggested as an appropiate model for the distribution of compartments in the brain containing roughly the square root of the total number of neurons ($\sim 10^5$) [14]. This seems more realistic than two-dimensional geometries used in previous cellular automata models [8]. The main advantage of discrete cellular automata models is that we can explicitly calculate the fluctuations in the number of firing neurons. Our interest in these fluctuations arise from our hypothesis that electroencephalographic (EEG) records and other global monitoring methods for brain activity (Magnetoencephalograms, Positron emission tomography) should exhibit these fluctuations as a noisy background. The prediction of a gaussian subdiffusive noise is, at least, consistent with the statistical analysis of EEG, if we consider it as a random walk. Moreover, the ratio of the variances between $\delta$ and $\alpha$ rhythms is also similar to the ratio of the maximum and the minimum variance in our cellular automata model when the population of inhibitory neurons is, approximately, to a 20 per cent of that of excitatory ones.

Statistical analysis of EEG has traditionally been restricted to the frequency spectrum and the so-called amplitude [16], which is an unfortunate

term for a clearly nonperiodic signal. Our approach could contribute to a better understanding of EEG and the relation between the states of the brain and global activity signals. In this philosophy, recent clinical studies of comatose patients [19] have stressed the importance of EEG as a tool to evaluate the possible recovery from coma: alpha coma is followed by diffuse delta waves and, finally, theta activity. Our model could accommodate this process in a scenario in which the excitation, inhibition and spontaneous deactivation probabilities vary as the patient recovers.

Present work could develop along the following experimental and theoretical research lines: Firstly, it would be interesting to analyze both pathological and normal EEG with the techniques proposed in this paper using longer records from more subjects. Of particular interest is the study of alpha coma records in order to test the idea that this is the noisy output of a brain with low activity as suggested by our model. Statistical correlations and the mean-shape of a fluctuation [20] are other independent tools from the theory of random walks whose application to EEG could reveal some hidden features. Nevertheless, it is clear that EEG cannot be merely noise and as we move to smaller and smaller scales organized activity should be observed. In order to fill the gap between collective macroscopic models and more realistic microscopic descriptions of neurons, a second step in the modeling of collective fluctuations would be to substitute our simple automata with Hodgkin-Huxley neurons [21, 22], that is, to formulate a detailed neural population model. Population models have been studied since the pioneer work of Wilson and Cowan [23] but litte attention have been paid to global spontaneous fluctuations, albeit diffusive noise for single neurons is well-known [21]. Further work should disclose if population models also exhibits the main features of cellular automata: second-order phase transitions, and subdiffusive neural dynamics. Work along these lines is in progress and should be published elsewhere.

# References

[1] S. Wolfram, *Cellular Automata and Complexity*, Collected Papers, Westview Press, Boulder, CO, U.S.A., (1994).

[2] E. Ahmed and H. N. Agiza, On modeling epidemics including latency, incubation and variable susceptibility, *Physica A* **253**, 347-352, (1998).

[3] A. Deutsch and S. Dormann, *Cellular Automaton Modeling of Biological Pattern Formation*, Birkhäuser Verlag, Basel, Germany, (2004).

[4] P. Bak, K. Chen, and C. Tang, A forest-fire model and some thoughts on turbulence, *Phys. Lett. A* **147**, 297-300, (1990).

[5] J. M. Greenberg and S. P. Hastings, Spatial Patterns for Discrete Models of Diffusion in Excitable Media, *SIAM J. Appl. Math.* **34**, 515, (1978).

[6] G. Bub, A. Shrier, and L. Glass, Spiral Wave Generation in Heterogeneous Excitable Media, *Phys. Rev. Lett.* **88**, 058101, (2002).

[7] Y. Bar-Yam, *Dynamics of Complex Systems*, Addison-Wesley, Reading, MA, (1997).

[8] M. I. Hoffmann, A Cellular Automaton Model Based on Cortical Physiology, *Complex Systems* **1**, 187, (1987).

[9] M. Tatsuno, Y. Nagai and W. Aizawa, Rule-dynamical Approach to Hippocampal Network, *Neurocomputing* **38**, 965-971, (2001).

[10] L. Acedo, A second-order phase transition in the complete graph stochastic epidemic model, *Physica A* **370**, 613-624, (2006).

[11] M. Abeles, *Corticonics: Neural Circuits of the Cerebral Cortex*, Cambridge University Press, Cambridge, (1991).

[12] S. Havlin and D. Ben-Avraham, *Diffusion and Reactions in Fractals and Disordered Systems*, Cambridge University Press, Cambridge, (2000).

[13] H. C. Berg, *Random Walks in Biology*, Princeton University Press, Princeton, NJ, (1983).

[14] V. Braitenberg, Brain Size and Number of Neurons: An Exercise in Synthetic Neuroanatomy, *J. Comput. Neurosci.* **10**, 71-77, (2001).

[15] M. A. M. de Aguiar, E. M. Rauch, and Y. Bar-Yam, Invasion and Extinction in the Mean Field Approximation for a Spatial Host-Pathogen Model, *J. Stat. Phys.* **114**, 1417-1451, (2004).

[16] A. H. Ropper and R. H. Brown, *Adams and Victor's Principles of Neurology*, McGraw-Hill, New York, (2005).

[17] D. A. Ratkowsky, *Handbook of non-linear regression models*, M. Dekker, New York, (1990).

[18] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, PhysioBank, PhysioToolkit, and Physionet: Components of a New Research Resource for Complex Physiologic Signals, *Circulation* **101** (23), e215-e220, (2000).

[19] S. Fossi, A. Amantini, A. Grippo, C. Cossu, N. Boni, and F. Pinto, Anoxic-Ischemic alpha coma: prognostic significance of the incomplete variant, *Neurol. Sci.* **24**, 397-400, (2003).

[20] A. Baldassarri, F. Colaiori, and C. Castellano, Average Shape of a Fluctuation: Universality in Excursions of Stochastic Processes, *Phys. Rev. Lett.* **90**, 060601, (2003).

[21] W. N. Kistler and W. Gerstner, *Spiking Neuron Models: Single Neurons, Populations, Plasticity*, Cambridge University Press, Cambridge, (2002).

[22] A. L. Hodgkin and A. F. Huxley, A quantitative description of ion currents and its applications to conduction and excitation in nerve membranes, *J. Physio. Lond.* **117**, 500-544, (1952).

[23] H. R. Wilson and J. D. Cowan, A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue, *Kybernetik* **13**, 55-80, (1973).

# Training of an Attractor Neural Network in a Stochastic Environment

L. Acedo*

Instituto de Matemática Multidisciplinar

Universidad Politécnica de Valencia, Edificio 8G, 2°

Camino de Vera, 46022 Valencia, España

December 11, 2008

## 1 Introduction

In 1894 Ramón y Cajal suggested that the strenghtening of connections among neurons was the basis of memory storing into the brain [1]. Half a century later, this idea was developed by Donald Hebb [2] in his theory of synaptic plasticity. Since then, Hebbian Theory has consolidated as the paradigm for learning and memory formation in the brain. The essence of Hebb's theory has been paraphrased into a very simple adage: "cells that fire together, wire together". This means that: (i) Memories are stored into the synapses among neurons which, in turn, is a more efficient mechanism than using the neurons themselves as memory storage devices (taking into account that every neuron projects between $10^3$ to $10^4$ synapses to other neurons and there are about $10^{10}$ neurons in the human brain). (ii) The chemical synapse between a couple of neurons becomes more effective when both neurons fire simultaneously. This is called Long Term Potentiation (LTP) by Neurophysiologists [3]. Moreover, experiments on insects such as the honey bee [4] suggest that learning takes place even on the most simple nervous systems ($\sim 950000$ neurons [5]).

*e-mail: luiacrod@imm.upv.es

In 1973, Lomo and Bliss performed a experiment in the rabbit hippocampus [6] and found the first evidence of the enhancement of a postsynaptic cell response after their excitation by a volley of high-frequency stimulus. Despite the conjectural status of Hebb's proposal, the work on LTP made it more believable. Just two years after the work of Lomo and Bliss, a multilayered neural network including a training algorithm, the so-called Cognitron, was proposed by Fukushima [7]. The most popular mathematical model for memory was proposed in 1982 by Hopfield and it is known as Attractor Neural Network or Hopfield Networks. This model is inspired in the Ising model for magnetic systems [10] and models neurons as binary automata with two states: 0 for resting or quiescent and 1 for firing. One of the nice achievements of this model is that it behaves as an associative memory, i. e., imprinted memories can be recalled from a pattern that includes only a fraction of the original memory [9]. Very recently, experiments performed in cultured networks, formed by real neurons grown on artificial substrate, have shown that collective modes of firing can be imprinted by chemical stimulation and recalled for days [11]. The recording of the activity of an ensemble of neurons in the hippocampus of a living rat have also become possible thanks to the technique developed in the group of Tsien [12]. Moreover, this group have also find the ensemble of neurons that fires collectively in a mouse hippocampus in connection with the concept of nest [13]. This means that when the mouse finds a cavity that can be used as a nest these neurons start to fire and they do not cease their firing until the cavity is filled.

This evidence is an important backup to the Hebb-Hopfield paradigm for memory but there are still many questions to be addressed. Mathematical models, albeit oversimplified, could play an important role by answering some of these questions but, equally relevant, by posing new meaningful questions that neurobiologists, psychologists or ethologists could transform into experiments. If this synergy is succesful we will perform another step toward the consolidation of this paradigm or perhaps we will discover it must be changed. In this paper we study the learning process of patterns that are never fully presented as perceptual inputs to the brain. On the contrary, we consider that only a random fraction $p$ of the pattern is observed. Animals rarely finds ideal patterns in their natural environment. Nevertheless, they are able to integrate the partial information they receive into complete neural maps and categories of objects, sounds, etc. Moreover, In this paper we show that Hopfield networks are very efficient in storing complete memories of patterns that have only excited a partial set of the neurons in the whole

neural map. In order to perform so well, we show that at least a random fifty per cent of the pattern must be perceived at different moments.

## 2 Hebbian learning model and the learning phase transition

We consider the standard neural network model with two possible states for the neurons: $s_i = +1$ for the firing of the $i-$th neuron and $s_i = -1$ for quiescent neurons. This boolean automata neurons are placed upon the nodes of a complete graph in which the edges represent synapses. The strength of these synapses is described by means of a matrix $J_{ij}$, $i = 1, \ldots, N$, $j = 1, \ldots, N$. The self-interaction of neurons upon themselves is not considered so $J_{ii} = 0$ and the diagonal elements of this matrix are null. Moreover, we suppose that the strength of these synapses does not depend on the direction of propagation of the electrochemical impulse and, consequently, the corresponding matrix is symmetric: $J_{ij} = J_{ji}$ for any $i$ and $j$.

A pattern of activation of a neural assembly is reinforced by the Long Term Potentiation mechanism. Mathematically, this is usually implemented as follows:

$$J_{ij}(t+1) = J_{ij}(t) + cs_i(t)s_j(t) \ , \tag{1}$$

where $c$ is an small constant and time is discrete with a unit step of the order of the natural time scale for neural processes ($\sim$ 1 ms). In order to study the learning of a pattern in noisy conditions we have considered a lattice arrangement of the neurons in a square with $N/2$ neurons at every side. We take simple circle and square patterns in this geometry and imprint then according to rule in Eq. (1) but only a fraction $p$, randomly chosen, of the set of neurons describing the pattern is firing at a given time step.

After this imprinting stage we prepare the system in an arbitrary random state (neurons are firing or quiescent with probability 0.5) and evolve it according to Glauber dynamics [9]:

$$s_i(t+1) = \text{sign}\left(\sum_{j=1}^{N} J_{ij}s_j(t)\right) \ . \tag{2}$$

Once the fixed point for this evolution is attained we calculate Hamming's

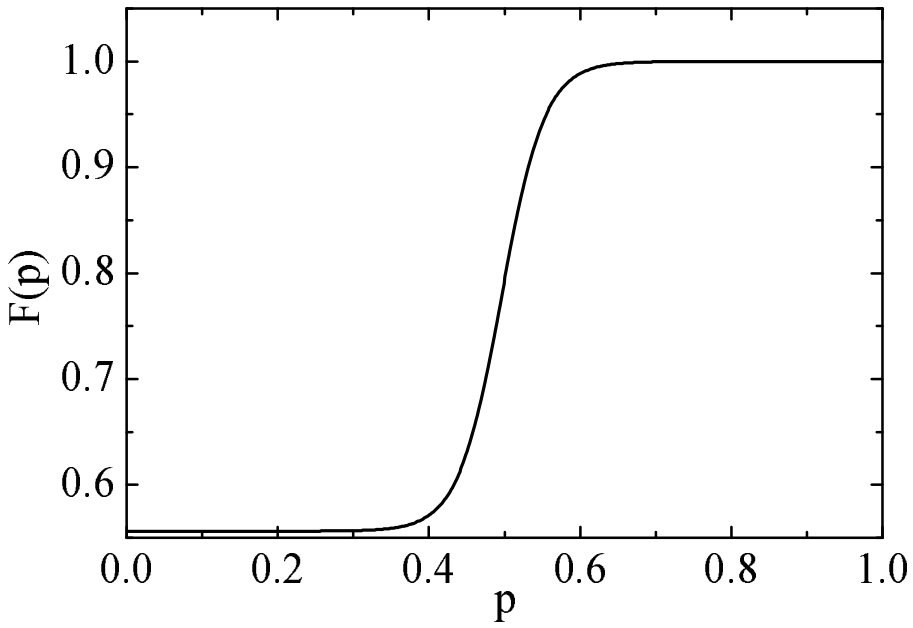



Figure 1: Average retrieval fraction of a pattern, $F(p)$ which stimulates a random fraction $p$ of neurons at a time. We considered a circular pattern and $c = 0.001$ for the imprinting stage (which last for 100 time steps).

distance between the ideal pattern and the recovered pattern:

$$d(\mathbf{s}, \mathbf{s}') = \frac{N}{2} - \sum_{i=1}^{N} s_i s_i' \, . \tag{3}$$

If we repeat the last steps a sufficiently large number of times to achieve a good statistics we can define $F(p)$ as the average of the Hamming distance between the fixed point of Glaubers's dynamics and the real pattern. This is a measure of the efficiency of the learning process in a random or noisy environment. As shown in Fig.1, we conclude that only for the simultaneous activation of a fraction $p > 0.5$ neurons during the learning stage we retrieve the original pattern. Otherwise, almost nothing is learned albeit all sites of the pattern are "seen" by the other at one or other time. This is a presumably testable consequence of Hopfield model in an adequate psychological experiment.

# 3 Conclusions and Remarks

Attractor Neural Networks are a classical model for pattern learning and retrieval in the brain. Neurophysiological basis for this mechanism is the long term potentiation of synapses among neurons which fire together, the so-called Hebbian learning paradigm. Despite this hypothesis was proposed more than fifty years ago, it has received substantial experimental backup very recently. As it is well-known, memories are represented by collective firing patterns in this network. These patterns can be excited by setting the attractor neural network in a state in which only a fraction of neurons in the pattern are firing.

In this communication, we study the Hebbian learning process in a stochastic environment, i. e., only a fraction $p$ of the real pattern is observed at every time step. This is achieved by a random excitation of neurons with probability $p$ at the learning stage. After this learning stage, the neural network is evolved towards its attractor using Glauber dynamics. The learning function $F(p)$, the fraction of the real pattern learned for a random observation of a fraction $p$, fits well to a sigmoidal with a critical value $p = 0.5$. For $p > 0.5$ most of the pattern is learned but for $p < 0.5$ almost nothing is learned. This behaviour corresponds to a phase transition in the learning process.

This phase transition could be relevant for psychology and ethology because animals and humans face with an everchanging environment and, consequently, their nervous systems are evolved to achieve efficient learning of patterns and regularities, even in a world that is always shifting and never presents the same information at different times. Taking into account that learning is not exclusive of humans but quite a general skill in the animal kingdom, the possibility of testing this transition in a controlled experiment are real. The learning of simple visual patterns in insects is perhaps the better scenario for this experiment.

Another interesting possibility not pondered in this work is the learning of several patterns from noisy presentations to the subject. This patterns could be presented alternatively, changing between one pattern and another at different time steps, or in different stages, presenting one pattern for certain amount of time and changing to the other for the same amount of time. Interference between the two patterns can be expected. Work along this line is in progress and will be published elsewhere.

# References

[1] S. Ramón y Cajal, The Croonian Lecture: La Fine Structure des Centres Nerveux, *Proceedings of the Royal Society of London* **55**, 444-468, (1894).

[2] D. O. Hebb, *The Organization of Behaviour*, Wiley, New York, (1949).

[3] O.Paulsen and T. J. Sejnowski, Natural patterns of activity and long-term synaptic plasticity, *Current Opinion in Neurobiology* **10** (2), 172-179, (2000).

[4] K. von Frisch, *The Dance Language and Orientation of Bees*, Harvard Univ. Press, Cambridge, MA, (1967).

[5] R. Menzel and M. Giurfa, Cognitive architecture of a mini-brain: the honeybee, *Trd. Cog. Sci.* **5**, 62-71, (2001).

[6] T. Lomo, The discovery of long-term potentiation, *Phil. Trans. R. Soc. Lond. B* **358**, 617-620, (2003).

[7] K. Fukushima, Cognitron: A Self-Organizing Multilayered Neural Network, *Biological Cybernetics* **20**, 121-136, (1975).

[8] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences of the USA* **79** (8), 2554-2558, (1982).

[9] Y. Bar-Yam, *Dynamics of Complex Systems*, Addison-Wesley, Reading, MA, (1997).

[10] K. Huang, *Statistical Mechanics*, 2nd ed., John Wiley & Sons, New York, (1987).

[11] I. Baruchi and E. Ben-Jacob, Towards neuro-memory-chip: Imprinting multiple memories in cultured neural networks, *Phys. Rev. E* **75**, 050901(R), (2007).

[12] L. Lin, R. Osan, S. Shoham, W. Jin, W. Zuo and J. Z. Tsien, Identification of network-level coding units for real-time representation of episodic experience in the hippocampus, *Proceedings of the National Academy of Sciences USA* **102** (17), 6125-6130, (2005).

[13] L. Lin, G. Chen, H. Kuang, D. Wang and J. Z. Tsien, Neural encoding of the concept of nest in the mouse brain, *Proceedings of the National Academy of Sciences USA* **104** (14), 6066-6071, (2007).

# Numerical and analytical solutions of forcing seasonal diseases using the differential transformation method

Abraham J. Arenas[*], Gilberto González-Parra[†], Benito M. Chen-Charpentier[‡]

(*) Departamento de Matemáticas y Estadística, Universidad de Córdoba, Montería, Colombia

(†) Departamento de Cálculo, Universidad de los Andes, Mérida, Venezuela

(*,†) Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, España

(‡) Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019, USA

December 11, 2008

## 1 Introduction

The numerical solution of seasonal epidemic models has been obtained in several papers in order to investigate numerically the reliability and efficiency of the different methods. For instance in [1], a nonstandard numerical method was tested numerically using a seasonally forced epidemic model. Additionally, in [2] a Fourier transform method was studied and applied to analyze the population dynamics of nematode infections of ruminants with the effect of seasonality in the free-living stages. Also, in [3] a nonstandard numerical method for the solution of a mathematical model for the *RSV*

[*]e-mail: aarenas@sinu.unicordoba.edu.co

[†]gcarlos@ula.ve

[‡]bchen@uwyo.edu

epidemiological transmission is used to investigate the numerical efficiency of the method.

In this work seasonal epidemiological models are solved using the $DTM$ for approximating the solutions in a sequence of time intervals. It is showed that the $DTM$ is easy to apply and their numerical solutions preserve the properties of the continuous models, such as periodic behavior, positivity and boundedness. Furthermore, the proposed numerical method is used in some cases with arbitrarily large time step sizes, saving computational cost when integrating over long time periods. It is important to remark that this method is applied directly to system of nonlinear ordinary differential equations without requiring linearization, discretization or perturbation.

The $DTM$ is a semi-analytical numerical technique depending on Taylor series that promises to be useful in various fields of mathematics. The $DTM$ derives from the differential equation system with initial conditions a system of recurrence equations that finally leads to a system of algebraic equations whose solutions are the coefficients of a power series solution. However, the classical $DTM$ has some drawbacks: the obtained truncated series solution does not exhibit the periodic behavior which is characteristic of seasonal disease models and gives a good approximation to the true solution, only in a small region. Therefore, in order to accelerate the rate of convergence and improve the accuracy of the calculations, it is necessary to divide the entire domain $H$ into $n$ subdomains. The main advantage of domain split process is that only a few series terms are required to get the solution in a small time interval $H_i$. Therefore, the system of differential equations can then be solved in each subdomain [4]. After the system of recurrence equations has been solved, each solution $x^j(t)$ can be obtained by a finite-term Taylor series. Thus this proposed $DTM$ does not have the above drawbacks.

The differential transformation technique is applied here to solve seasonal epidemiological models. The first two models are related to the $RSV$ epidemiological transmission [5, 6] and the third model to obesity dynamics at the population level [9].

## 2 Basic definitions of DTM

For the sake of clarity in the presentation of the $DTM$ and in order to help to the reader we summarize the main issues of the method that may be found in [7].

**Definition 2.1** *Let $x(t)$ be analytic in the time domain $D$, then it has derivatives of all orders with respect to time $t$. Let*

$$\varphi(t,k) = \frac{d^k x(t)}{dt^k}, \qquad \forall t \in D. \tag{1}$$

*For $t = t_i$, then $\varphi(t,k) = \varphi(t_i,k)$, where $k$ belongs to a set of non-negative integers, denoted as the $K$ domain. Therefore, (1) can be rewritten as*

$$X(k) = \varphi(t_i,k) = \left[\frac{d^k x(t)}{dt^k}\right]_{t=t_i} \tag{2}$$

*where $X(k)$ is called the spectrum of $x(t)$ at $t = t_i$.*

**Definition 2.2** *Suppose that $x(t)$ is analytic in the time domain $D$, then it can be represented as*

$$x(t) = \sum_{k=0}^{\infty} \frac{(t-t_i)^k}{k!} X(k). \tag{3}$$

Thus, the equation (3) represents the inverse transformation of $X(k)$.

**Definition 2.3** *If $X(k)$ is defined as*

$$X(k) = M(k)\left[\frac{d^k x(t)}{dt^k}\right]_{t=t_i} \tag{4}$$

*where $k \in \mathbb{Z}^+ \cup \{0\}$, then the function $x(t)$ can be described as*

$$x(t) = \frac{1}{q(t)} \sum_{k=0}^{\infty} \frac{(t-t_i)^k}{k!} \frac{X(k)}{M(k)}, \tag{5}$$

*where $M(k) \neq 0$ and $q(t) \neq 0$. $M(k)$is the weighting factor and $q(t)$ is regarded as a kernel corresponding to $x(t)$.*

Note, that if $M(k) = 1$ and $q(t) = 1$, then Eqs. (2) and (3) and (4) and (5) are equivalent.

**Definition 2.4** *Let $[0, H]$ the interval of simulation with $H$ the time horizon of interest. We take a partition of the bounded interval $[0, H]$ as $\{0 = t_0, t_1, \ldots, t_n = H\}$ such that $t_i < t_{i+1}$ and $H_i = t_{i+1} - t_i$ for $i = 0, \ldots, n$. Let $M(k) = \frac{H_i^k}{k!}$, $q(t) = 1$ and $x(t)$ be a analytic function in $[0, H]$. It then defines the differential transformation as*

$$X(k) = \frac{H_i^k}{k!}\left[\frac{d^k x(t)}{dt^k}\right]_{t=t_i} \quad \text{where } k \in \mathbb{Z}^+ \cup \{0\}, \tag{6}$$

*and its differential inverse transformation of $X(k)$ is defined as follow*

$$x(t) = \sum_{k=0}^{\infty}\left(\frac{t}{H_i}\right)^k X(k), \text{ for } t \in [t_i, t_{i+1}]. \tag{7}$$

From the definitions above, we can see that the concept of differential transformation is based upon the Taylor series expansion. Thus, applying the $DTM$ a system of differential equations in the domain of interest can be transformed to a algebraic equation system in the $K$ domain and each $x^j(t)$ can be obtained by the finite-term Taylor series plus a remainder, i.e.,

$$x^j(t) = \frac{1}{q(t)}\sum_{k=0}^{n}\frac{(t-t_i)^k}{k!}\frac{X^j(k)}{M(k)} + R_{n+1} = \sum_{k=0}^{n}\left(\frac{t}{H}\right)^k X^j(k) + R_{n+1}, \tag{8}$$

where

$$R_{n+1} = \sum_{k=n+1}^{\infty}\left(\frac{t}{H}\right)^k X^j(k), \text{ and } R_{n+1} \to 0 \text{ as } n \to \infty.$$

For practical problems of simulation, the computation interval $[0, H]$ is not always small, and to accelerate the rate of convergence and improve the accuracy of the calculations, it is necessary to divide the entire domain $H$ into $n$ subdomains. The main advantage of domain split process is that only a few Taylor series terms are required to construct the solution in a small time interval $H_i$, where $H = \sum_{i=1}^{n} H_i$. It is important to remark that, $H_i$ can be chosen arbitrarily small if necessary. Thus, the system of differential equations can then be solved in each subdomain [4]. The approach described above is known as the $D$ spectra method. Considering the function $x^j(t)$ in the first sub-domain ($0 \le t \le t_1$, $t_0 = 0$), the one dimensional differential transformation is given by

$$x^j(t) = \sum_{k=0}^{n}\left(\frac{t}{H_0}\right)^k X_0^j(k)\,, \text{ where } X_0^j(0) = x_0^j(0). \tag{9}$$

Therefore, the differential transformation and system dynamic equations can be solved for the first subdomain and $X_0^j$ can be solved entirely in the first subdomain. The end point of function $x^j(t)$ in the first subdomain is $x_1^j$, and the value of $t$ is $H_0$. Thus, $x_1^j(t)$ is obtained by the differential transformation method as

$$x_1^j(H_0) = x^j(H_0) = \sum_{k=0}^{n} X_0^j(k). \tag{10}$$

Since that $x_1^j(H_0)$ represents the initial condition in the second subdomain, then $X_1^j(0) = x_1^j(H_0)$. And so the function $x^j(t)$ can be expressed in the second sub-domain as

$$x_2^j(H_1) = x^j(H_1) = \sum_{k=0}^{n} X_1^j(k). \tag{11}$$

In general, the function $x^j(t)$ can be expressed in the $i-1$ subdomain as

$$x_i^j(H_i) = x_{i-1}^j(H_{i-1}) + \sum_{k=1}^{n} X_{i-1}^j(k) = X_{i-1}^j(0) + \sum_{k=1}^{n} X_{i-1}^j(k),\ i = 1, 2, ..., n. \tag{12}$$

Using the $D$ spectra method described above, the functions $x^j(t)$ can be obtained throughout the entire domain, for all $j$.

# 3 The operation properties of the differential transformation

We consider $q(t) = 1$, $M(k) = \frac{H_i^k}{k!}$ and $x^1(t)$, $x^2(t)$, $x^3(t)$ three uncorrelated functions of time $t$ and $X^1(k)$, $X^2(k)$, $X^2(k)$ the transformed functions corresponding to $x^1(t)$, $x^2(t)$, $x^2(t)$. With $\mathcal{D}$ we denote the Differential Transformation Operator. Thus, the following basic properties hold:

1. **Linearity**. If $X^1(k) = \mathcal{D}[x^1(t)]$, $X^2(k) = \mathcal{D}[x^2(t)]$ and $c_1$ and $c_2$ are independent of $t$ and $k$ then

$$\mathcal{D}[c_1 x^1(t) \pm c_2 x^2(t)] = c_1 X^1(k) \pm c_2 X^2(k). \tag{13}$$

   Thus, if $c$ is a constant, then $\mathcal{D}[c] = c\delta(k)$, where $\delta(k)$ is the Dirac delta function.

2. **Convolution**. If $X^1(k) = \mathcal{D}[x^1(t)]$, $X^2(k) = \mathcal{D}[x^2(t)]$, then

$$\mathcal{D}[x^1(t)x^2(t)] = X^1(k) * X^2(k) = \sum_{l=0}^{k} X^1(l)X^2(k-l). \text{ Therefore,}$$

$$\begin{aligned}\mathcal{D}[x^1(t)x^2(t)x^3(t)] &= X^1(k) * \Big(X^2(k) * X^3(k)\Big) \\ &= \sum_{k_2=0}^{k}\sum_{k_1=0}^{k_2} X^3(k_1)X^2(k_2-k_1)X^3(k-k_2). \end{aligned} \tag{14}$$

3. **Derivative**. If $x^1(t) \in C^n[0, H]$, then

$$\mathcal{D}\left[\frac{d^n x^1(t)}{dt^n}\right] = \frac{(k+1)(k+2)\cdots(k+n)}{H_i^n} X^1(k+n). \tag{15}$$

4. If $x^1(t) = \cos(\omega t + \alpha)$, then

$$\mathcal{D}[x^1(t)] = \frac{(H_i\omega)^k}{k!}\cos\left(\frac{\pi k}{2} + \alpha + 2\pi i H_i\right), \tag{16}$$

where $i$ denotes the $i$-th split domain.

# 4 Applications to seasonal epidemic models

In this section, the differential transformation technique is applied to solve several nonlinear differential equations system that arise from seasonal epidemiological models. The seasonality of the models is given by the transmission rate $\beta(t)$ and biological considerations mean that must be a continuous function, positive, nonconstant and periodic of period $T$. Thus, in this work

$$\beta(t) = b_0\left(1 + b_1\, cos\left(\frac{2\pi}{T}(t+\phi)\right)\right), \tag{17}$$

where $b_0 \geq 0$ is the baseline transmission parameter, $0 \leq b_1 \leq 1$ measures the amplitude of the seasonal variation in the transmission and $0 \leq \varphi \leq 1$ is the phase angle normalized.

## 4.1 Models for the transmission of Respiratory Syncytial Virus ($RSV$)

The model is presented as follows

$$\begin{aligned}
\dot{S}(t) &= \mu - \mu S(t) - \beta(t)S(t)I(t) + \gamma R(t), && S(0) = S_0 > 0 \\
\dot{I}(t) &= \beta(t)S(t)I(t) - \nu I(t) - \mu I(t), && I(0) = I_0 > 0 \\
\dot{R}(t) &= \nu I(t) - \gamma R(t) - \mu R(t), && R(0) = R_0 > 0.
\end{aligned} \tag{18}$$

From section (3) it deduces its spectrum

$$\mathbf{S}(k+1) = \frac{H_i}{k+1}\left\{\mu(\delta(k) - \mathbf{S}(k)) + \gamma\mathbf{R}(k) - \sum_{k_2=0}^{k}\sum_{k_1=0}^{k_2}\mathbf{B}(k_1)\mathbf{S}(k_2-k_1)\mathbf{I}(k-k_2)\right\}$$

$$\mathbf{I}(k+1) = \frac{H_i}{k+1}\left\{\sum_{k_2=0}^{k}\sum_{k_1=0}^{k_2}\mathbf{B}(k_1)\mathbf{S}(k_2-k_1)\mathbf{I}(k-k_2) - (\mu+\nu)\mathbf{I}(k)\right\}, \tag{19}$$

$$\mathbf{R}(k+1) = \frac{H_i}{k+1}\left\{\nu\mathbf{I}(k) - (\mu+\gamma)\mathbf{R}(k)\right\},$$

with $\mathbf{S}(0) = S_0$, $\mathbf{I}(0) = I_0$, $\mathbf{R}(0) = R_0$ and

$$\mathbf{B}(k_1) = b_0\delta(k) + b_0 b_1 \frac{(H_i\omega)^k}{k!}\cos\left(\frac{\pi k}{2} + \phi + 2\pi i H_i\right).$$

Thus, from a process of inverse differential transformation, it can be obtained the solutions of each sub-domain taking $n+1$ terms for the power series like Eq. (9), i.e.,

$$\begin{aligned}
S_i(t) &= \sum_{k=0}^{n}\left(\frac{t}{H_i}\right)^k \mathbf{S}_i(k),\ 0 \le t \le H_i, \\
I_i(t) &= \sum_{k=0}^{n}\left(\frac{t}{H_i}\right)^k \mathbf{I}_i(k),\ 0 \le t \le H_i, \\
R_i(t) &= \sum_{k=0}^{n}\left(\frac{t}{H_i}\right)^k \mathbf{R}_i(k),\ 0 \le t \le H_i,
\end{aligned} \tag{20}$$

provided that the solutions holds with:

$$S(t) = \sum_{i=0}^{n} S_i(t),\ I(t) = \sum_{i=0}^{n} I_i(t),\ R(t) = \sum_{i=0}^{n} R_i(t). \tag{21}$$

As a second case, in [6] a nested model was presented for study the dynamical transmission of $RSV$ at the population level. This model is structured with a set of four ordinary differential equations that include homotopy parameters which provide paths for different types of $RSV$ transmission models, and where the transmission rate is a continuous positive periodic function. The model is as follows

$$\begin{aligned}
\dot{S}(t) &= \mu P - \frac{\Lambda(t)S(t)}{P} + \frac{\alpha\tau}{\rho}(I_S(t) + I_R(t) + R(t)) - \mu S(t),\\
\dot{I}_S(t) &= \frac{\Lambda(t)S(t)}{P} - (\tau + \mu)I_S(t),\\
\dot{I}_R(t) &= \frac{\sigma\Lambda(t)R(t)}{P} - \left(\frac{\tau}{\rho} + \mu\right) I_R(t),\\
\dot{R}(t) &= \left(1 - \frac{\alpha}{\rho}\right)\tau I_S(t) + \frac{(1-\alpha)\tau}{\rho} I_R(t) - R(t)\left(\frac{\sigma\Lambda(t)}{P} + \frac{\alpha\tau}{\rho} + \mu\right),
\end{aligned} \tag{22}$$

## 4.2 Mathematical model for obesity population

As a third case, a mathematical model for obesity population with seasonality is considered. The population is characterized by the following classes: $N\,(t)$ is the proportion of normal weight individuals, $L\,(t)$ the proportion of latent individuals, $S\,(t)$ the proportion of overweight individuals,$O\,(t)$ the proportion of obese individuals, $D_S\,(t)$ the proportion of overweight on diet individuals, and $D_O\,(t)$ the proportion of obese on diet individuals. Thus, the mathematical model, considering constant population size, is given by the following system of nonlinear ordinary differential equations

$$\begin{aligned}
\dot{N}\,(t) &= \mu + \varepsilon D_S(t) - \mu N(t) - \beta(t)N(t)[L(t) + S(t) + O(t)],\\
\dot{L}\,(t) &= \beta(t)N(t)[L(t) + S(t) + O(t)] - [\mu + \gamma_L]L(t),\\
\dot{S}\,(t) &= \gamma_L L(t) + \varphi D_S(t) - [\mu + \gamma_S + \alpha]S(t),\\
\dot{O}\,(t) &= \gamma_S S(t) + \delta D_O(t) - [\mu + \sigma]O(t),\\
\dot{D_S}\,(t) &= \gamma_D D_O(t) + \alpha S(t) - [\mu + \varepsilon + \varphi]D_S(t),\\
\dot{D_O}\,(t) &= \sigma O(t) - [\mu + \gamma_D + \delta]D_O(t),
\end{aligned} \tag{23}$$

where the whole population is normalized to unity, i.e., $N(t) + L(t) + S(t) + O(t) + D_S(t) + D_O(t) = 1$, and $\beta(t)$ is as in (17). And again, the spectrum of above models are obtained as in (19).

# 5 Conclusions

In this work, seasonal epidemiological models are solved numerically using the $DTM$ for approximating the solutions in a sequence of time intervals. In order to obtain very accurate solutions, the domain region has been splitted into subintervals and the approximating solutions are obtained in a sequence of time intervals. The $DTM$ produces from the system of differential equations with initial conditions a system of recurrence equations that finally leads to a system of algebraic equations whose solutions are the coefficients of a power series solution, and applying a process of inverse transformations it obtain the solutions. Moreover, the $DTM$ does not evaluate the derivatives symbolically and this give advantages over other methods such Taylor, power series or Adomian method.

Here, it is showed that the $DTM$ is easy to apply and their numerical solutions preserves the properties of the continuous models, such as periodic behavior, positivity and boundedness, which when using Runge-Kutta and other numerical methods, we cannot guarantee these properties especially with step size $h$ relatively large. Furthermore, the calculated results demonstrate the reliability and efficiency of the method when is applied to seasonal epidemiological models.

Based on the numerical results it can be concluded that the $DTM$ is a mathematical tool which enables to find approximate accurate analytical solutions for seasonal epidemiological models represented by systems of nonautonomous nonlinear ordinary differential equations. In general, by splitting the time domain, the numerical solutions can be approximated quite well using a small number of terms and small time interval $H_i$. Furthermore, high accuracy can be obtained without using large computer power and the $DTM$ has the advantage of giving an analytical form of the solution within each time interval which is not possible in purely numerical techniques like $RK4$.

# References

[1] W. Piyawong, E.H. Twizell, A.B. Gumel, An unconditionally convergent finite-difference scheme for the SIR model, *Appl. Math. Comput.* 146(23) 611-625 (2003).

[2] M.G. Roberts and B.T. Grenfell, The population dynamics of nematode infections of ruminants: The effect of seasonally in the free-living stages, *Math Med Biol* 9(1) 29-41 (1992).

[3] A.J. Arenas, J.A. Moraño, J.C. Cortés, Non-standard numerical method for a mathematical model of RSV epidemiological transmission, *Comp. Math. Appl.* 56(3)670-678 (2008).

[4] C.L. Chen, S.H. Lin, C.K. Chen, Application of Taylor transformation to nonlinear predictive control problem, *Appl. Math. Model.* 20(9) 699-710 (1996).

[5] A. Weber, M. Weber, P. Milligan, Modeling epidemics caused by respiratory syncytial virus (RSV), *Math. Biosci.* 172(2) 95-113 (2001).

[6] L.J. White, J.N. Nandl, M.G. Gomes, A.T. Bodley-Tickell, P.A. Cane, P. Perez-Brena, J.C. Aguilar, M.M. Siqueira, S.A. Portes, S.M. Straliotto, M. Waris, D.J. Nokes, G.F. Medley. Understanding the transmission dynamics of respiratory syncytial virus using multiple time series and nested models. *Math. Biosc.* 209(1) 222-239 (2007).

[7] J.K. Zhou, Differential Transformation and its Applications for Electrical Circuits, Huazhong University Press, Wuhan,(in Chinese), 1986.

[8] I. Hwang, J. Li, D. Du, A numerical algorithm for optimal control of a class of hybrid systems: differential transformation based approach, *International Journal of Control* 81(2) 277-293 (2008).

[9] L. Jódar, F. Santonja, G. González-Parra, Modeling dynamics of infant obesity in the region of Valencia, Spain, *Comp. Math. Appl.* 56(3) 679-689 (2008).

# T-Wave Alternans diagnostic using Wavelet Transform

Macarena Boix[*], Begoña Cantó[†], David Cuesta[‡], and Pau Micó[‡]

(*) Dpto. de Matemática Aplicada, UPV, Valencia, Spain

(†) Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia, Valencia, España, 46071

(‡) Departamento de Informática de Sistemas y Computadores, Universidad Politécnica de Valencia, Valencia, España, 46071

December 11, 2008

## 1 Introduction

Human beings have a heart with four cavities, two atria and two ventricles. The QRS complex corresponds to a power source that causes the ventricular depolarization. The most important part of any analysis system of the ECG signal is the detection of QRS complex.

T-wave represents the ventricular repolarization. T-wave alternans (TWA) are periodic beat-to-beat variations in the amplitude of the T-wave in a surface ECG. Since the early twentieth century, a large number of experimental and clinical studies checked the usefulness of T-wave morphological changes, as a parameter for evaluating risk of sudden death and malignant arrhythmia. There are evidences of TWA appearances: Long QT Syndrome, Myocardial Ischemia and Infarction, Cardiomyopathies, Heart Failure, Sudden Infant

[*]e-mail: mboix@mat.upv.es

[†]e-mail: bcanto@imm.upv.es

Death Syndrome, and so on. For that, the aim of this paper is to detect TWA and we use the Wavelet Transform (WT).

In our study, we have been forced to generate synthetic ECG signals with TWA, due to the lack of databases containing ECG real signals with changes in the T-wave. Firstly, the ECG signal from the database MIT-BIH Arrhythmia DB is taken and the TWA beat-to-beat with the considered electrical level is added, in this case it is called augmented T-wave. It has been used to generate signals to software ECGLab given in [1], and TWA is generated with or without the addition of the sinusoidal wave with the required electrical level.

The difference between the amplitude of the augmented T-waves and the normal ones is called TWA-level. The TWA-level presented at a ECG signal has been verified by measuring the difference between the average amplitudes of the augmented T-waves (augmented level) and the normal ones (normal level). Several electrical levels: 1mV, 0.5mV, 0.2mV, 0.1mV, 0.05mV, 0.02mV and 0.01mV are used for the experiments presented below.

Next, we give some definitions and mathematical tools used along this paper. In [2] it is showed the following definitions.

**Definition 1**. A function $\Psi(t) \in L^2(\mathbb{R})$ is said to be a wavelet if only if its Fourier transform $\widehat{\Psi}(\omega)$ satisfies

$$\int_0^{+\infty} \frac{\left|\widehat{\Psi}(\omega)\right|^2}{\omega}\, d\omega = \int_{-\infty}^{0} \frac{\left|\widehat{\Psi}(\omega)\right|^2}{|\omega|}\, d\omega = C_\Psi < +\infty.$$

**Definition 2**. Given $\Psi(t) \in L^2(\mathbb{R})$ we denote $\Psi_a(t)$ the dilation of $\Psi(t)$ by a factor $a$ (where $a$ is a positive number): $\Psi_a(t) = \frac{1}{a}\Psi\left(\frac{t}{a}\right)$.

**Definition 3**. A smoothing function is a function $\theta(t)$ whose integral is equal to 1 which converges to 0 at infinity and we assume also $\theta(t)$ to be differentiable.

**Definition 4**. The wavelet transform of a function f(t) at scale $a$ is given by the convolution product $W_a f(t) = f * \Psi_a(t)$.

**Remark 1**. The dilation of a smoothing function is still smoothing function and the derivative of a smoothing function is a wavelet.

For practical applications the scale parameter (and the traslation parameter) must be discretized: the scale parameter can be sampled along the dyadic sequence $\{2^j\}_{j\in\mathbb{Z}}$ without modifying the overall properties of the transforma-

tion. Generally, the parameter $j$ is known as level of decomposition.

**Theorem 1**. [2] Let $\theta(t)$ be a smoothing function. Let $\Psi(t)$ be the first-order derivative of $\theta(t)$ (consecuently a wavelet). Then, the wavelet transform of a function f(t) at scale $a$ is given by

$$W_a f(t) = f * \Psi_a(t) = f * \left( a \frac{d\theta_a}{dt} \right)(t) = a \frac{d}{dt} \left( f * \theta_a \right)(t).$$

This relation tells that the wavelet transform of a function is proportional to the derivative of the smoothed function. By changing the scale $a$, we can obtain the derivatives of the smoothed function at different scales. We refer to [2] for a complete treatment of the wavelet theory.

# 2 Algorithm to detect R-peaks and T-peaks

The ECG signal is a not stationary signal, that is the cause why WT is convenient for analysis. At small scales, the WT reflects the high frequency components of the signal and, at large scales it reflects the low frequency components of the signal. For example, T-waves have different frequency content and could be detected at higher scales of decomposition.

The WT modulus' local maxima at different scales can be used to locate the sharp variation points of ECG signals [3]. In [2] is showed that if the wavelet is the first derivative of a smoothed function, the maximum local of dyadic WT indicates the abrupt signal changes, while the minimum local indicates slow variations. Besides, the WT submits zeros at different scales in the positions where signal is maximum local or minimum local. On the time–frequency plane of the wavelet transform, the wave rising edge of the QRS complex corresponds to a negative minimum, and the dropping edge corresponds to a positive maximum at different scales. The zero-crossing points of these pairs are found to give the location of R peaks [3].

Usually, the onset of the QRS complex contains the high-frequency components, which are detected at finer scales, here we use the level of decomposition $j = 1$.

A similar procedure has been used in the detection of T-waves, using $j = 9$ as the level of decomposition.

We describe below the process followed in the experiments.

1). Firstly, QRS complexes are detected and removed from the signal $x$. The R peaks has been detected using a algorithm which is inspired in

work [4]. The mother wavelet used is 'rbio3.1' (reverse biorthogonal spline wavelet), a real wavelet.

To detect QRS complex, the signal is taken and it is passed through filter passband FIR (order 100), between 1 and 20 Hz and the 100 first samples (400 ms) are removed. Next, the first 5 seconds of this signal are subdivided in intervals of 200 ms since this period is considered as a refractory period. In each interval, the local maxima of the $|WT|$ are obtained. And between these maxima values that exceed a threshold, are selected and they are labelled $x_i$. The considered threshold is $u1 = 0.8$ times the mean($|WT|$) of the total signal. We consider the Wavelet Transform ($WT(x)$) and we take windows from 0.16 s, beginning from a point $i/|WT(i)|$ exceed a second threshold: $u2 = 0.8$ times the mean($|WT(x_i)|$). If in that window exists the zero-crossing of the WT, we will consider the intervals $[a - i, a + i]$, where $a = 0.12$ s. The maximum of signal $x$ in this interval is founded and this one will be a possible R peak. The maxima founded such that they exceed a threshold $u3$ times the maximum amplitude of the signal x, are the R peaks. We take $u3 = 0.3$. Next, we compare with a cardiologists' detections and we accept an error of 5 samples (20 ms). Once detected R peaks of the ECG signal, they are eliminated and we call to the new signal $g$.

2). Secondly, T-waves are detected (both augmented as normal).

We return to apply to signal $g$ the same filter and the 100 first samples are eliminated. We apply the detector at level of decomposition $j = 9$. The procedure is similar to that described above.

We consider the wavelet transform $WT(g)$ and we take windows from 0.16 s, beginning from a point $i/|WT(i)| > \text{mean}|WT(g)|$. If in that window exists the zero-crossing of the WT, we will consider the intervals $[a-i, a+i]$, where $a = 0.12$ s. The maximum of function $g$ in this interval is founded and this one will be a possible T peak. The maxima founded such that they exceed a threshold $u$ times the maximum amplitude of the signal g, are the T peaks. We take $u = 0.11$ for the TWA-level = $\{1, 0.5\}$ mV; $u = 0.16$ for the TWA-level = $\{0.2\}$ mV and $u = 0.2$ for TWA-level = $\{0.1, 0.05, 0.02, 0.01\}$ mV. The range of variation allowed in error is 5 samples.

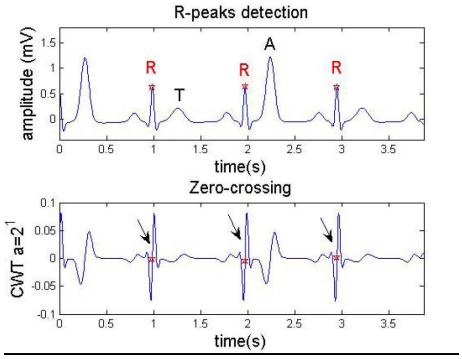



Figure 1. R-peaks detection using WT.

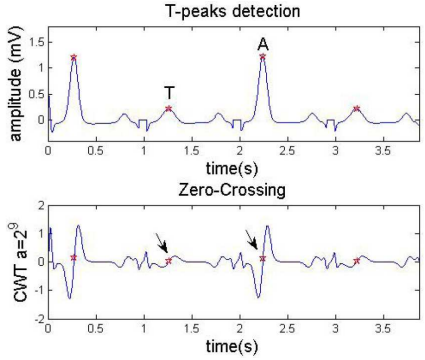



Figure 2. T-peaks detection using WT.

# 3 Conclusions

We use the following notations: TP (true positive): peaks detected correctly; FP (false positive): peaks that algorithm marks incorrectly; FN (false negative): peaks not detected to obtain the error (E) $E = \dfrac{FP + FN}{TP + FN} 100$.

Firstly, we have detected the R peaks, being the average error from 0.06%. Next, we have detected T peaks obtaining an average error of 0.36%.

We think that the results are quite satisfactory. However, we must bear in mind that they have been obtained in synthetic ECG signals, and to verify the proper detector's operation, it would have to be tested in real ECG signals.

# References

[1] Mc P.E. Sharry, G.D. Clifford, L. Tarassenko and L. Smith, A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*. **50** (3) 289-294 (2003).

[2] S. Mallat, and S. Zhong, Charaterization of Signals from Multiscale Edges. *IEEE Trans. on pattern analysis and machine intelligence*. **14**(7) 710-732 (1992).

[3] C. Li, Ch. Zheng, and Ch. Tai, Detection of ECG Characteristic Points Using Wavelet Transforms. *IEEE Trans. on biomedical engineering*. **42**(1) 21-28 (1995).

[4] G. Camps, M. Martínez, E. Soria, R. Magdalena, J. Calpe and J. Guerrero, Foetal ECG recovery using dynamic neural networks. *Artificial Intelligence in Medicine*. **31**(3) 197-209 (2004).

# On the stability of a biomedical mathematical model *

Begoña Cantó†, Carmen Coll‡, and Elena Sánchez§

Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia,

Camino de Vera s/n, 46022, Valencia, España

December 11, 2008

## 1 Introduction

Mathematical models are used in biomedicine to simulate diferent processes which study the movements of the human body under diferents situations. Mathematical models are an important tool to model the hemodialysis process. In this case we want to analyze a model that allows us to determine the amount of waste, potassium and/or urea are eliminated in the blood. This method consists in subject the patient to a process in order to remove wastes and extra fluids through a filter called dialyzer. In these models several compartments are considered, normally one or two, where the liquid runs freely.

Hemodialysis is an artificial way to clean the blood when kidneys can no longer do that on their own. This treatment is characterized by a low quality of life and a high cost. Mathematical modelling helps physicians to predict the effect of dialysis procedures and to decide the correct dialysis dose to the patient.

*Supported in part by Grant MMT2007-64477.

†bcanto@imm.upv.es

‡mccoll@imm.upv.es

§esanchezj@imm.upv.es

In most of the common dialyzers (called artificial kidneys) the renoval of toxical substances from the blood is achieved by extracting it from the body and introducing it into the interior of a kidney machine where it floats along one side of a membrane as being schematically shown in Figure 1.

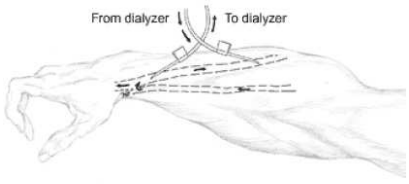



Figure 1. Hemodialysis treatment.

Usually, when the evolution of a process in the nature is analyzed it is use dynamic systems. In these systems the definition of a stable equilibrium point is given in Lyapunov sense and it is important the concept of asymptotic stability.

A biomedical system may have many equilibrium points to which it moves under small perturbations and mathematically this means that the system is unstable in Lyapunov sense. However, in some cases, from a biological point of view, this fact is acceptable to study the stable behaviour of the system since trajectories fall into a cyclical equilibrium. The points that verify this behaviour are called attractive points.

In recent years several works have been developed in order to analyze the dialysis process. For instance, a model of water circulation that it is used to keep certain variables, such as pressure, above the critical values in a patient is showed in [1], several mathematical models that combine physiological knowledge to information held by the patients on certain substances such as urea or sodium are given in [2] and a model of a three compartment model or central nervous system in order to study the time-optimal control of hemodialysis is presented in [3]. In the hemodialysis model some coefficients are not known, and these coefficients play an important role in the behaviour of the system.

The hemodialysis model can be described as a mathematical equation that shows the exchange process between the patient and the dialyzer

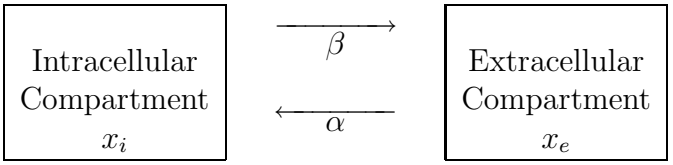
$$x(k+1) = f(x(k), u(k), p).$$

In this equation, $x(k)$ is the percentage of initial wastes at time $k$, $u(k)$ represents the dialyzer action at time $k$ and $p$ represents the unknown parameters of the model.

In this paper, a dynamic linear system with two compartments will be used to describe the hemodynamical phenomena occurring during dialysis. Some conditions will be obtained to ensure the existence of attractive equilibrium points to this model and conditions that the system can lead to an equilibrium state will be analyzed.

## 2 Two compartmental model

The mathematical model to describe dialysis treatment is based on a two compartmental model where the compartments are separated by a semipermeable membrane. Urea works particularly well for this model because of its high permeability in aqueous environments. Its permeability allows it to move freely between the compartments, forward and reverse directions, all the time. Consequences of this reversible behaviour can cause an equilibrium. The model can be described as follows




where $\alpha$ and $\beta$ are the intercompartmental mass transfer coefficients and they take positive and bounded values from 0 to 1. Moreover, $x_i + x_e = 1$. This model is represented by the following state equation

$$x(k+1) = A(p)x(k), \tag{1}$$

being $x(k) = \begin{pmatrix} x_e(k) \\ x_i(k) \end{pmatrix}$ and $A(p) = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{pmatrix}$.

The coefficient matrix depends on the parametric vector $p = (\alpha \ \ \beta)$. This vector belongs to a set which depends on the characteristics of the system. Thus, we want to obtain the domain of variation of the parameters in order to obtain a well defined structure to improve the behavior of the model. If $x_0$ is the initial amount of waste in the blood, the trajectory of the system is given by

$$x(k) = A^k(p)x_0.$$

The relation between two compartments allows achieving equilibrium points. An *equilibrium point* of the system (1), denoted by $x^*$, verifies $x^* = A(p)x^*$. This equilibrium point $x^*$ is said to be asymptotically stable if every trajectory starting in a neighborhood of it is around the $x^*$ and converges on $x^*$.

From a mathematical point of view, the system (1) is asymptotically stable if the eigenvalues of the matrix $A(p)$ are inside the unit circle, that is, if $\rho(A(p)) < 1$.

In our case, the eigenvalues of the matrix $A(p)$ depends on the considered parameters, because they are given by $\lambda = 1$ and $\lambda = 1 - \alpha - \beta$. That is, the system (1) is not asymptotically stable in mathematical sense, which confirms the assertion that a hemodialysis process can have many equilibrium points always present in the environment. However we are interested to study the domain of attraction of the equilibrium points.

An equilibrium point $x^*$ is an attractive equilibrium if there exists $\mu$ such that $\|x_0 - x^*\| < \mu$ then $\lim_{k\to\infty} x(k) = x^*$ (for more details see [4]).

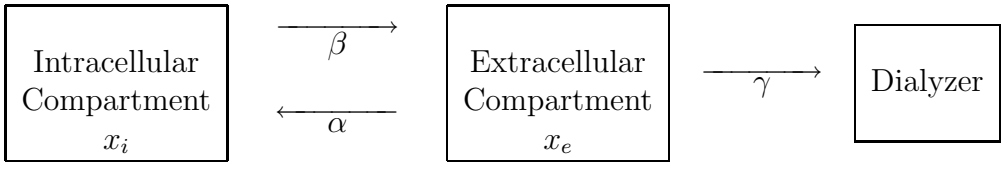
If the parameter vector $p = (\alpha \quad \beta)$, satisfies that $\alpha + \beta \leq 1$, then the attractive equilibrium points of the system (1) are obtained using the expression given by the following equation

$$\lim_{k\to\infty} x(k) = \lim_{k\to\infty} A(p)^k x_0 = \frac{1}{\alpha+\beta}\begin{pmatrix} \beta \\ \alpha \end{pmatrix}.$$

Hence, the point $\left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$ is an attractive equilibrium point because the trajectory of the system is moving towards this point after a sufficiently long time.

# 3 Using a dialyzer

When kidneys fall dialysis treatment cleans blood by a dialysis machine. In this case the model is modified by adding a new compartment connected to the extracellular compartment, as shown in the following diagram



where $\gamma$ is the dialyzer urea clearance coefficient. This coefficient takes positive and bounded values from 0 to 1.

This circuit is represented by the equation (1) with the addition of a feedback that represents the flow of blood to the dialyzer

$$x(k+1) = A(p)x(k) + B(p)u(k), \tag{2}$$

being $B(p) = \begin{pmatrix} -\gamma \\ 0 \end{pmatrix}$, and in this case the coefficient matrices depend on the parameter vector $p = (\alpha \ \ \beta \ \ \gamma)$. The process will be completed when the blood cleaned in the dialyzer returns to the patient. Mathematically this fact is represented by a state feedback that it will provide a dynamic closed-loop system.

If we consider a state feedback $u(k) = Fx(k)$ with $F = [1 \quad 0]$, then the dynamic closed-loop system associated to the system (2) is given by the following equation

$$x(k+1) = (A(p) + B(p)F)\, x(k) = \begin{pmatrix} 1-\alpha-\gamma & \beta \\ \alpha & 1-\beta \end{pmatrix} x(k).$$

As we can observe the system is an autonomous one and this fact allows us to analyze its development over time from the eigenvalues of the coefficient matrix of the system. Thus, the stability of the system (2) is given by the eigenvalues of the matrix $A(p) + B(p)F$

$$\lambda = \frac{1}{2}\left(2 - \alpha - \beta - \gamma \pm \sqrt{\alpha^2 + (\beta-\gamma)^2 + 2\alpha(\beta+\gamma)}\right).$$

We can check that the parameter vector $p = (\alpha \ \ \beta \ \ \gamma)$ has all its entries smaller than one, that is, its eigenvalues are inside the unit circle because they satisfy $0 < \lambda < 1$. Hence, the closed-loop system is asymptotically stable, that is, the trajectory of this system tends to a point and converges to it. To achieve the equilibrium point obtained previously, and the stability behavior of the system, it will be necessary to subject the patient to treatment dialysis for a long period of time. This affects the patient's health because it must undergo lengthy and continuous dialysis sessions, with the consequent deterioration in his health. For that, it will be interesting to get some mathematical models to establish the optimal dose and duration of the process in a dialysis patient, knowing what is the concentration of waste and excess water would be desirable to remove. Thus, the quality of life of patients can be increased.

But this is an open question that needs clinical investigations to contrast the results of the model.

In short, in this work a dynamic control system associated to a process of kidney malfunction represented by two compartments has been considered. Initially a free control system has been studied and secondly a system with a state feedback linked to a parametric vector has been used. For these two systems, conditions to detect and obtained attractive equilibrium points have been obtained. Note that the state feedback is connected to the dialyzer and it will be determine by its business model.

# References

[1] U. Moissl, P. Wabel and R. Isrmann, Model-based control of hemodialysis. *Proceedings of American Control Conference.* **1** 25-27, Arlington (2001).

[2] J. Waniewski, Mathematical modeling of fluid and solute transport in hemodialysis and peritoneal dialysis. *Journal of Membrane Science.* **274** 24-37 (2006).

[3] M. Zilko, J. Pietrzyk and P. Dyras, Time-optimal control of hemodialysis. *Proceedings of the Third IEEE Conference on Control Applications.* **3** 1685-1688 (1994).

[4] E. Kaszkurewicz and A. Bhaya, *Matrix diagonal stability in systems and computation*, Birkhauser Verlag, USA (2000).

# Computing matrix functions solving coupled differential equations for engineering models *

Emilio Defez [†], Jorge Sastre [‡], Javier Ibáñez[§] and Pedro A. Ruiz[¶]

(†) Instituto de Matemática Multidisciplinar.

(‡) Instituto de Telecomunicaciones y Aplicaciones Multimedia.

(§, ¶) Instituto de Aplicaciones de las

Tecnologías de la Información y de las Comunicaciones Avanzadas,

Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, España.

December 11, 2008

It is well known that the wave equation

$$v^2 \frac{\partial^2 \psi}{\partial x^2} = \frac{\partial^2 \psi}{\partial t^2} \ , \tag{1}$$

plays an important role in many areas of engineering and applied sciences. The matrix differential problem

$$Y''(t) + AY(t) = 0 \ , \ Y(0) = Y_0 \ , \ Y'(0) = Y_1 \ , \tag{2}$$

where $A$ is a matrix and $Y_0$ and $Y_1$ are vectors, arises from spatially semi-discretization of the wave equation (1), see [1]. Matrix problem (2) has the exact solution

$$Y(t) = \cos\left(\sqrt{A}t\right)Y_0 + \left(\sqrt{A}\right)^{-1} \sin\left(\sqrt{A}t\right)Y_1, \tag{3}$$

*Acknowledgments. This work has been supported by the *Generalitat Valenciana* project GVPRE/2008/340.

[†]e-mail: edefez@imm.upv.es

[‡]jorsasma@iteam.upv.es

[§]jjibanez@dsic.upv.es

[¶]pruiz@dsic.upv.es

where $\sqrt{A}$ denotes any square root of a non-singular matrix $A$ (see *e.g.* equation 1.2 of [2]). More general problems of type (2), with a forcing term $F(t)$ on the right-hand side arise from mechanical systems without damping, and their solutions can be expressed in terms of integrals involving the matrix sine and cosine [3]. Thus, trigonometric matrix functions play an important role in second order differential systems, similar to matrix exponentials in first order differential problems.

A general algorithm for computing the matrix cosine which uses rational approximations and the double-angle formula $\cos(2A) = 2\cos^2(A) - I$ was proposed by Serbin and Blalock [1]. Higham in [2,4,5] developed a particular version of this algorithm based on the Padé approximation.

In this work, that may be regarded as a continuation of [6], we use Hermite matrix polynomial expansions of the matrix cosine and sine in order to perform a very accurate and competitive method for computing them compared to the results given by the function *funm* of MATLAB.

Throughout this work, $[x]$ denotes the integer part of $x$. The matrices $I_r$ and $\theta_{r\times r}$ in $\mathbb{C}^{r\times r}$ denote the matrix identity and the null matrix of order $r$, respectively. Following [7], for a matrix $A$ in $\mathbb{C}^{r\times r}$, its infinite-norm will be denoted by $\|A\|_\infty$ and its 2-norm will be denoted by $\|A\|_2$. Finally, if $A(k,n)$ are matrices in $\mathbb{C}^{r\times r}$ for $n \geq 0, k \geq 0$, from [6] it follows that

$$\sum_{n\geq 0}\sum_{k\geq 0} A(k,n) \;=\; \sum_{n\geq 0}\sum_{k=0}^{n} A(k,n-k)\;. \tag{4}$$

## Hermite matrix polynomials series expansions of matrix sine and matrix cosine.

For the sake of clarity in the presentation of the following results we recall some properties of Hermite matrix polynomials which have been established in [6] and [8]. From (3.4) of [8, p. 25] the $n$th Hermite matrix polynomial satisfies

$$H_n\left(x, \frac{1}{2}A^2\right) = n!\sum_{k=0}^{[\frac{n}{2}]} \frac{(-1)^k\,(xA)^{n-2k}}{k!(n-2k)!}\;, \tag{5}$$

for an arbitrary matrix $A$ in $\mathbb{C}^{r\times r}$. Taking into account the three-term recurrence relationship (3.12) of [8, p. 26], it follows that

$$\left.\begin{array}{c} H_n\left(x,\frac{1}{2}A^2\right)=xAH_{n-1}\left(x,\frac{1}{2}A^2\right)-2(n-1)H_{n-2}\left(x,\frac{1}{2}A^2\right) \quad , \quad n\geq 1 \\ H_{-1}(x,\frac{1}{2}A^2)=\theta_{r\times r} \quad , \quad H_0(x,\frac{1}{2}A^2)=I_r \end{array}\right] , \tag{6}$$

and from its generating function in (3.1) and (3.2) [8, p. 24] one gets

$$e^{xtA-t^2I}=\sum_{n\geq 0}H_n\left(x,\frac{1}{2}A^2\right)t^n/n!, \quad |t|<\infty, \tag{7}$$

where $x,t\in\mathbb{C}$. The $n$th scalar Hermite polynomial is given by [9, p. 60]

$$H_n(x)=n!\sum_{k=0}^{[\frac{n}{2}]}\frac{(-1)^k\,(2x)^{n-2k}}{k!(n-2k)!} \ , \ n\geq 0 \ , \tag{8}$$

which coincide with the $n-$th matrix Hermite polynomial (5) when $r=1$ and $A=2$.

Taking $y=tx$ and $\mu=1/t$ in (7) it follows that

$$e^{Ay}=\mathrm{e}^{\frac{1}{\mu^2}}\sum_{n\geq 0}\frac{1}{\mu^n n!}H_n\left(\mu y,\frac{1}{2}A^2\right) , \ \mu\in\mathbb{C}, \ \ y\in\mathbb{C}, \ \ A\in\mathbb{C}^{r\times r} \ . \tag{9}$$

Now, we look for the Hermite matrix polynomials series expansion of the matrix cosine $\cos(Ax)$. Given an arbitrary matrix $A\in\mathbb{C}^{r\times r}$, with

$$\cos(Ay)=\frac{e^{iAy}+e^{-iAy}}{2} \ ,$$

and using (9) in combination with [8, p. 25], it follows that

$$H_n(-x,A)=(-1)^nH_n(x,A) \ .$$

Thus, one gets

$$\cos(Ay) \ =e^{\frac{1}{\mu^2}}\sum_{n\geq 0}\frac{1}{\mu^{2n}(2n)!}H_{2n}\left(iy\mu,\frac{1}{2}A^2\right) \ . \tag{10}$$

Taking $\lambda = i\mu$ in (10), we obtain the looked for expression:

$$\cos(Ay) \ = e^{-\frac{1}{\lambda^2}} \sum_{n\geq 0} \frac{(-1)^n}{\lambda^{2n}(2n)!} H_{2n}\left(y\lambda, \frac{1}{2}A^2\right) \ . \tag{11}$$

In a similar form, taking into account that

$$\sin(Ay) = \frac{e^{iAy} - e^{-iAy}}{2i} \ ,$$

it follows that

$$\sin(Ay) \ = e^{-\frac{1}{\lambda^2}} \sum_{n\geq 0} \frac{(-1)^n}{\lambda^{2n+1}(2n+1)!} H_{2n+1}\left(y\lambda, \frac{1}{2}A^2\right). \tag{12}$$

Denoting by $C_N(A,\lambda)$ the $N$th partial sum of series (11) for $y = 1$, one gets

$$C_N(\lambda, A) = e^{-\frac{1}{\lambda^2}} \sum_{n=0}^{N} \frac{(-1)^n}{\lambda^{2n}(2n)!} H_{2n}\left(\lambda, \frac{1}{2}A^2\right) \approx \cos(A), \ \lambda \in \mathbb{C}, \quad A \in \mathbb{C}^{r\times r}. \tag{13}$$

Observe that the case $\lambda = 1$ corresponds with the matrix cosine approximation $C(A;1;N)$ given in [6]. Denoting by $S_N(A,\lambda)$ the $N$th partial sum of series (12) for $y = 1$, one gets

$$S_N(\lambda, A) = e^{-\frac{1}{\lambda^2}} \sum_{n=0}^{N} \frac{(-1)^n}{\lambda^{2n+1}(2n+1)!} H_{2n+1}\left(\lambda, \frac{1}{2}A^2\right) \approx \sin(A), \ \lambda \in \mathbb{C}, \ \ A \in \mathbb{C}^{r\times r}. \tag{14}$$

The introduction of the additional parameter $\lambda$ will improve the results given in [6].

By (5) and (8), it follows that

$$\left\| H_n\left(x, \frac{1}{2}A^2\right)\right\|_2 \leq \sum_{k=0}^{[n/2]} \frac{n!\left(|x|\,\|A\|_2\right)^{n-2k}}{k!(n-2k)!} \ , \tag{15}$$

and thus

$$\left\| H_{2n}\left(\lambda, \frac{1}{2}A^2\right)\right\|_2 \leq \sum_{k=0}^{n} \frac{(2n)!\left(\lambda\,\|A\|_2\right)^{2(n-k)}}{k!(2(n-k))!} \ . \tag{16}$$

Using (4), the following expression holds

$$\sum_{n\geq 0}\sum_{k=0}^{n}\frac{\|A\|_2^{2(n-k)}}{\lambda^{2k}k!\,(2(n-k))!} \;=\; \cosh\left(\|A\|_2\right)e^{\frac{1}{\lambda^2}}\;. \tag{17}$$

Taking the approximate value $C_N(\lambda, A)$ given by (13) and taking into account (16), it follows that

$$\|\cos(A) - C_N(\lambda, A)\|_2 \;\leq\; \mathrm{e}^{-\frac{1}{\lambda^2}}\left[\sum_{n\geq 0}\sum_{k=0}^{n}\frac{\|A\|_2^{2(n-k)}}{\lambda^{2k}k!(2(n-k))!} - \sum_{n=0}^{N}\sum_{k=0}^{n}\frac{\|A\|_2^{2(n-k)}}{\lambda^{2k}k!(2(n-k))!}\right] \tag{18}$$

Considering the previous expression, one gets an error bound for approximation (13):

$$\|\cos(A) - C_N(\lambda, A)\|_2 \leq \mathrm{e}^{-\frac{1}{\lambda^2}}\left[\cosh\left(\|A\|_2\right)e^{\frac{1}{\lambda^2}} - \sum_{n=0}^{N}\sum_{k=0}^{n}\frac{\|A\|_2^{2(n-k)}}{\lambda^{2k}k!(2(n-k))!}\right]. \tag{19}$$

Now, let $\varepsilon > 0$ be an *a priori* error bound. Using (19), if $N$ is the first positive integer so that

$$\sum_{n=0}^{N}\sum_{k=0}^{n}\frac{\|A\|_2^{2(n-k)}}{\lambda^{2k}k!(2(n-k))!} \geq \cosh\left(\|A\|_2\right)e^{\frac{1}{\lambda^2}} - \varepsilon\,\mathrm{e}^{\frac{1}{\lambda^2}}\;, \tag{20}$$

from (19) and (20) one gets

$$\|\cos(A) - C_N(\lambda, A)\|_2 \;\leq\; \varepsilon\;.$$

Summarizing, the next result, similar to theorem 3.1 of [6], has been proved:

**Theorem 1** *Let $A$ be a matrix in $\mathbb{C}^{r\times r}$ and let $\lambda > 0$. Let $\varepsilon > 0$. If $N$ is the first positive integer so that inequality (20) holds. Then*

$$\|\cos(A) - C_N(\lambda, A)\|_2 \;\leq\; \varepsilon\;. \tag{21}$$

Furthermore, using that relation $\sin(A) = \cos\left(A - \frac{\pi}{2}I\right)$, it is possible avoid the computation of the matrix sine. On the other hand, we can obtain a similar result to theorem 1 for the case of the matrix sine:

**Theorem 2** *Let $A$ be a matrix in $\mathbb{C}^{r\times r}$ and let $\lambda > 0$. Let $\varepsilon > 0$. If $N$ is the first positive integer so that inequality*

$$\sum_{n=0}^{N}\sum_{k=0}^{n}\frac{\|A\|_2^{2(n-k)+1}}{\lambda^{2k}k!(2(n-k)+1)!} \geq \sinh\left(\|A\|_2\right)e^{\frac{1}{\lambda^2}} - \varepsilon\, \mathrm{e}^{\frac{1}{\lambda^2}} \ ,$$

*holds. Then, approximation $S_N(\lambda, A)$ given by (14) satisfies*

$$\left\|\sin\left(A\right) - S_N(\lambda, A)\right\|_2 \ \leq \ \varepsilon \ . \tag{22}$$

Starting with expressions (13) and (14), it is possible to simultaneously compute the matrix cosine and sine.

## Numerical examples.

In this section we provide results for numerical experimentation of the computational method based on expansion (11) compared with the results given by the function *funm* of MATLAB. This function allows to compute general matrix functions by the Schur-Parlett algorithm and it is the only function that MATLAB has to compute matrix sine and cosine. The implementations have been tested on an Intel Core 2 Duo T5600 with 2 GB main memory, using 7.5 (R2007b) MATLAB version.

In the first example, we apply the computation of the matrix cosine of a matrix $A$ treated in [6] using the expansion (11). Note that there are different possible choices for the parameter $\lambda$.

**Example 1** *Let $A$ be a matrix defined by*

$$A \ = \ \begin{pmatrix} 3 & -1 & 1 \\ 2 & 0 & 1 \\ 1 & -1 & 2 \end{pmatrix} \ , \tag{23}$$

*with $\sigma(A) = \{1, 2\}$. Matrix $A$ is non-diagonalizable. Using the minimal theorem , see also [6], the exact value of $\cos\left(A\right)$ is*

$$\begin{aligned}\cos\left(A\right) \ &= \ \begin{pmatrix} \cos\left(2\right) - \sin\left(2\right) & \sin\left(2\right) & -\sin\left(2\right) \\ -\cos\left(1\right) + \cos\left(2\right) - \sin\left(2\right) & \cos\left(1\right) + \sin\left(2\right) & -\sin\left(2\right) \\ -\cos\left(1\right) + \cos\left(2\right) & \cos\left(1\right) - \cos\left(2\right) & \cos\left(2\right) \end{pmatrix} \\ &= \ \begin{pmatrix} -1.325444263372824 & 0.909297426825682 & -0.909297426825682 \\ -1.865746569240964 & 1.449599732693821 & -0.909297426825682 \\ -0.956449142415282 & 0.956449142415282 & -0.4161468365471424 \end{pmatrix} \ .\end{aligned}$$

**Algorithm 1** computes sine and cosine of a matrix by means of Hermite approximants.

**Function** $[C, S] = sincosher(A, N, \lambda)$
**Inputs**: Matrix $A \in \mathbb{R}^{r\times r}$; $2N+1$ is the order of the Hermite approximation ($N \in \mathbb{N}$) of sine/cosine function; parameter $\lambda \in \mathbb{R}$
**Output**: Matrices $C = \cos(A) \in \mathbb{R}^{r\times r}$ and $S = \sin(A) \in \mathbb{R}^{r\times r}$

1: $H_0 = I_r$
2: $H_1 = \lambda A$
3: $C = H_0$
4: $S = H_1/\lambda$
5: $aux = 1/\lambda$
6: **for** $n = 2 : 2N+1$ **do**
7: $H = \lambda A H_1 - 2(n-1)H_0$
8: $H_0 = H_1$;
9: $H_1 = H$
10: $aux = aux/(\lambda n)$
11: **if** $\bmod(n, 4) < 2$ **then**
12: **if** $\bmod(n, 2) == 0$ **then**
13: $C = C + auxH$;
14: **else**
15: $C = C - auxH$;
16: **end if**
17: **else**
18: **if** $\bmod(n, 2) == 0$ **then**
19: $S = S + auxH$;
20: **else**
21: $S = S - auxH$;
22: **end if**
23: **end if**
24: **end for**
25: $C = e^{-1/l^2}C$
26: $S = e^{-1/l^2}S$

*In [6], for an admissible error $\varepsilon = 10^{-5}$, we need $N = 15$ to provide the required accuracy. In practice, the number of terms required to obtain a prefixed accuracy uses to be smaller than the one provided by Theorem 3.1 of [6]. So for instance, taking $N = 9$ one gets:*

$$C_9(1, A) = \begin{pmatrix} -1.3254444650245485 & 0.9092974459509594 & -0.9092974459509594 \\ -1.8657468968644513 & 1.4495998777908623 & -0.9092974459509594 \\ -0.9564494509134919 & 0.9564494509134919 & -0.4161470190735891 \end{pmatrix},$$

*and*

$$\|\cos(A) - C_9(1, A)\|_2 = 7.995228661905607 \times 10^{-7}.$$

*We will compare these results obtained letting $\lambda = 1$ in Theorem 3.1 of [6] with the new Theorem 1. Taking $\lambda = 2000$, using Theorem 1 we need $N = 10$ to obtain the same prefixed accuracy. Again, the number of terms required to obtain a prefixed accuracy uses to be smaller than the one provided by (21). For instance, taking $N = 7$ one gets*

$$C_7(2000, A) = \begin{pmatrix} -1.3254442633775207 & 0.9092974268299509 & -0.9092974268299509 \\ -1.8657465692456603 & 1.4495997326980907 & -0.9092974268299509 \\ -0.9564491424157093 & 0.9564491424157093 & -0.41614683654756973 \end{pmatrix},$$

*and*

$$\|\cos(A) - C_7(2000, A)\|_2 = 7.717270333884585 \times 10^{-8}.$$

*The choice of parameter $\lambda$ can still be refined. For example, taking $\lambda = 4.1$ one gets*

$$\|\cos(A) - C_7(4.1, A)\|_2 = 7.098351906265066 \times 10^{-10}.$$

*Figure 1 presents the error 2-norm of approximation (11) for $N = 8$ fixed and $\lambda \in ]0, 25]$. This figure illustrates how the error norm depends on the varying parameter $\lambda$ and it becomes evident that an adequate choice of $\lambda$ may provide results with higher accuracy.*

Figure 2 shows the $2-$norm error bound of $C_N(\lambda, A)$ for the fixed value of $\lambda = 4.1$ varying $N$. For $N = 10$, we obtain

$$\|\cos(A) - C_{10}(4.1, A)\|_2 = 1.7763568394002505 \times 10^{-15}.$$

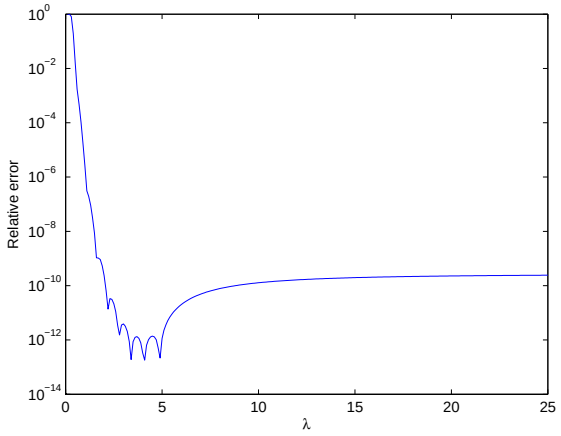



Figure 1: For $N = 8$ fixed and varying $\lambda$.

**Example 2** *In this experiment we consider* 100 *random matrices of the form*

$$A = PDP^{-1}, \tag{24}$$

*where $D$ is a diagonal matrix with uniform random values in the interval* $[-5, 5]$ *and $P$ is a matrix with uniform random values in the same interval. The dimensions of all matrices are* $100 \times 100$. *We have computated the approximation of the matrix cosine $C_N(\lambda, A)$ and sine $S_N(\lambda, A)$ with $N = 20$ and the experimental value of $\lambda$ was $\lambda = 0.7936$.*

*It is well known that the exact solutions are*

$$\cos(A) = P\cos(D)P^{-1} \quad , \quad \sin(A) = P\sin(D)P^{-1} \ .$$

*In the experiment each exact solution has been obtained at* $256-$*digit precision using MATLAB's Symbolic Math Toolbox.*

*Figure 3 shows the comparison between the relative errors of function* `funm` *of MATLAB and series (11) with* $\lambda = 0.7936$ *using the infinite norm:*

$$E_r(x^\star) = \frac{\|x - x^\star\|_\infty}{\|x\|_\infty} \ . \tag{25}$$

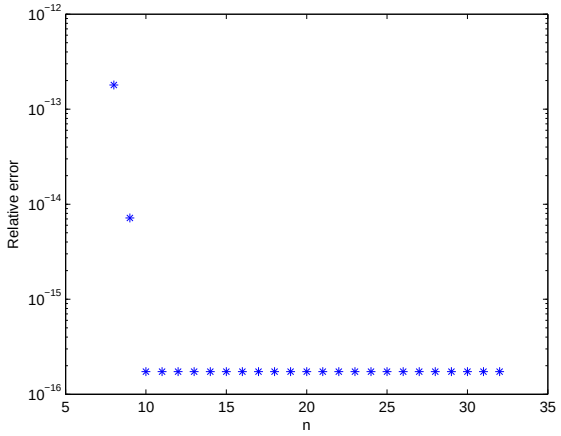



Figure 2: Relative error of Hermite series (11) for example 1 for $\lambda = 4.1$.

*The mean processing time for* **funm** *was* 0.114550 *seconds and the mean processing time for the Hermite approximation was* 0.023535 *seconds. The first average time corresponds only to the computation of* $\cos(A)$ *using the function* **funm**. *The second value corresponds to the computation of* $\cos(A)$ *and* $\sin(A)$ *using Hermite expansion. Our proposed implementation was* 4.8672 *times faster. In the computation of* $\cos(A)$, *the Hermite method gave a smaller error than* **funm** *in* 70% *of the test cases. In the computation of* $\sin(A)$, *the Hermite method gave a smaller error than* **funm** *in* 67% *of the test cases.*

**Example 3** *We consider 100 randomly matrices in the same conditions as in experiment 2. We have computed the approximation of the matrix cosine* $C_N(\lambda, A)$ *and sine* $S_N(\lambda, A)$ *with* $N = 25$. *We choose in this new experiment* $\lambda = 0.6175$.

*Figure 4 shows the comparison between the relative errors of function* **funm** *of MATLAB and series (11) with* $\lambda = 0.6175$ *using infinite norm (25).*

*Now, the mean processing time for* **funm** *was* 0.113997 *seconds and the mean processing time for the Hermite approximation was* 0.027875 *seconds. The first average time corresponds only to the computation of* $\cos(A)$ *us-*

*ing the function* **funm**. *The second value corresponds to the computation of* $\cos(A)$ *and* $\sin(A)$ *using Hermite expansion. Our proposed implementation was* 4.0896 *times faster. In the computation of* $\cos(A)$, *the Hermite method gave a smaller error than* **funm** *in* 74% *of the test cases. In the computation of* $\sin(A)$, *the Hermite method gave a smaller error than* **funm** *in* 74% *of the test cases.*

## Conclusions.

Matrix functions $f(A)$ are often motivated by PDE's, but computation of these functions is not always easy. The Cauchy integral theorem is a useful tool for complex analysis and for computing $f(a)$, but it can also be effective for computing $f(A)$. These functions are generalizations of the scalar case, but they are not always easy to compute. The matrix square root is not the square root of each entry of $A$, but rather is the matrix $B$ such that $B^2 = A$. Similarly the matrix exponential $e^A$ is not $e$ to the power of the $j$, $k-$entry of $A$, but is the matrix $B = I + A + A2/2! + A3/3! + \ldots$. These functions require significant effort to produce highly accurate solutions, and they can be very prone to rounding error in some cases.

Good examples are the matrix cosine and sine, $\cos(A)$ and $\sin(A)$. These functions are defined in the complex plane as

$$\cos(Ay) = \frac{e^{iAy} + e^{-iAy}}{2}, \sin(Ay) = \frac{e^{iAy} - e^{-iAy}}{2i}.$$

This naturally leads to the conclusion that perhaps these functions could be computed by first computing $e^{iA}$ and $e^{-iA}$ and adding them together, but Higham notes that this suffers from cancellation errors in floating point arithmetic. Computational tests revealed that this is indeed true. The Talbot contours will essentially be useless as well, since neither $\cos(z)$ nor $\sin(z)$ decays as $z \to \infty$. The standard circular contour is used to compute these matrix functions, which means that its speed and accuracy will depend highly upon the location of the eigenvalues of $A$. However, because of the periodicity of these functions, an elliptic contour appears more optimal.

In this paper a modification of the method proposed in [6] for computing matrix cosine and sine based on Hermite matrix polynomial expansion is presented. Numerical tests and an algorithm are given. The described method

allows the simultaneous evaluation of the matrix sine and cosine and it has been compared with the function `funm` of MATLAB. The method depends on the parameter $\lambda$, whose impact on the numerical efficiency is currently studied. Furthermore, pending work focuses on the optimal scaling of the matrix and the study of the evaluation of the approximations (13) and (14). To do parallel implementation of the algorithms presented in this work in a distributed memory platform, using the message passing paradigm, MPI and BLACS for communications, and PBLAS and ScaLAPACK for computations.

# References

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

[2] G. I. Hargreaves, N. J. Higham, Efficient algorithms for the matrix cosine and sine, Numerical Algorithms 40 (2005) 383–400.

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

[4] N. Higham, M. Smith, Computing the matrix cosine, Numerical Algorithms 34 (1) (2003) 13–26.

[5] N. J. Higham, Functions of Matrices: Theory and Computation, 2008.

[6] E. Defez, L. Jódar, Some applications of Hermite matrix polynomials series expansions, Journal of Computational and Applied Mathematics 99 (1998) 105–117.

[7] G. H. Golub, C. F. V. Loan, Matrix computations, 2nd Edition, The Johns Hopkins University Press, Baltimore, MD, USA, 1989.

[8] L. Jódar, R. Company, Hermite matrix polynomials and second order matrix differential equations, Journal Approximation Theory Application 12 (2) (1996) 20–30.

[9] N. N. Lebedev, Special Functions and Their Applications, New York, 1972.

[10] P. I. Davies, N. J. Higham, A schur-parlett algorithm for computing matrix functions, SIAM Journal on Matrix Analysis and Applications 25 (2) (2003) 464–485.

[11] N. Dunford, J. Schwartz, Linear Operators, Part I., New York, 1957.

[12] M. S. Paterson, L. Stockmeyer, On the number of nonscalar multiplications necessary to evaluate polynomials, SIAM Journal on Computing 2 (1) (1973) 60–66.

[13] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, ScaLAPACK Users' Guide, SIAM, 1997.

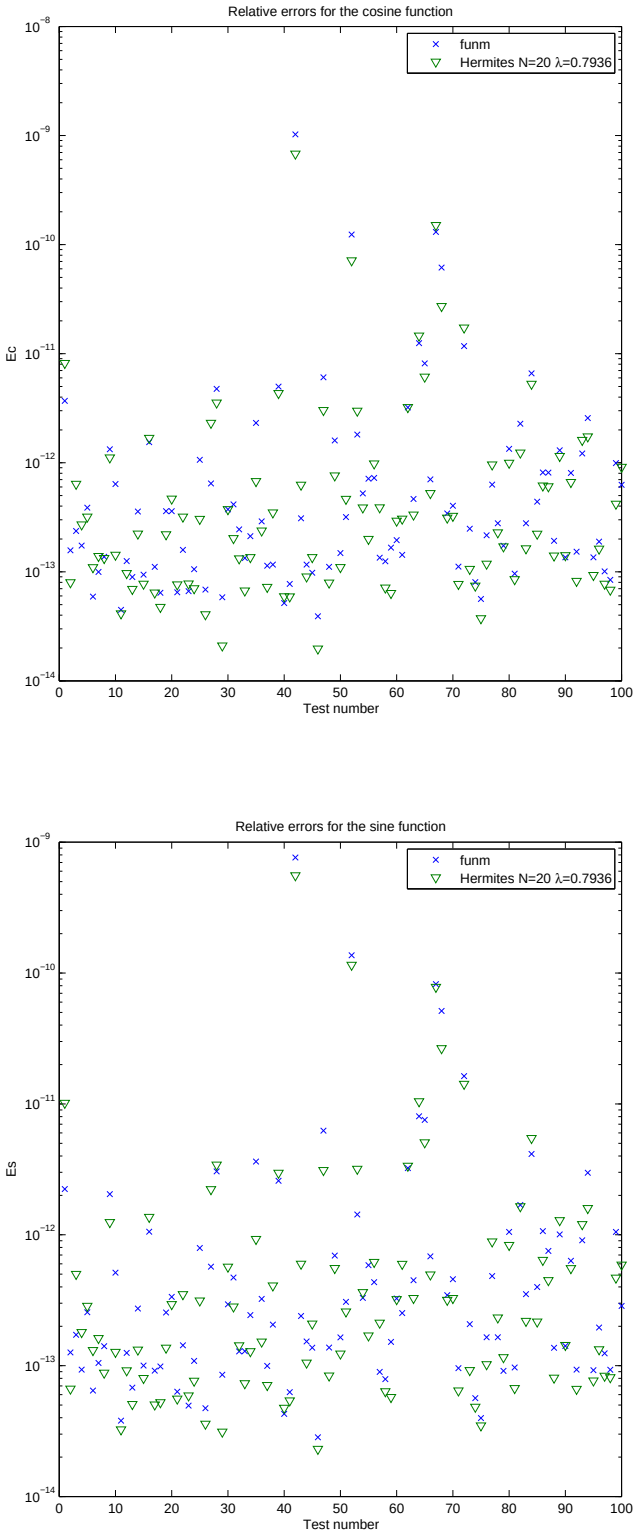



Figure 3: Comparison between the relative errors for cosine and sine computation with $N = 20$ and $\lambda = 0.7936$.

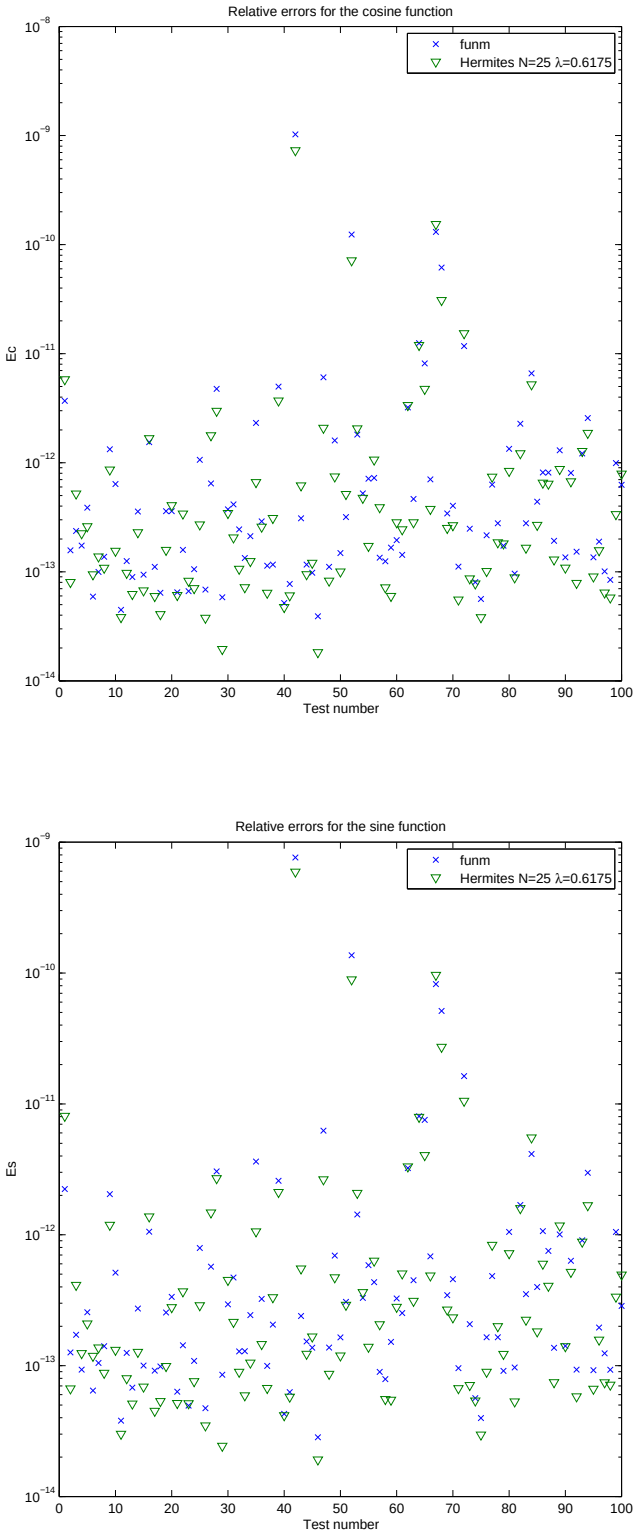



Figure 4: Comparison between the relative errors for cosine and sine computation with $N = 25$ and $\lambda = 0.6175$.

# A Numerical Approximation for Incomplete Second-Order Matrix Models in Engineering*

Emilio Defez†, Michael M. Tung ‡, and Javier Ibáñez§

(†) Instituto de Matemática Multidisciplinar.

Universidad Politécnica de Valencia,Camino de Vera s/n, 46022

Valencia, España.

(§) Instituto de Aplicaciones de las Tecnologías de la

Información y de las Comunicaciones Avanzadas,

UPV, Valencia, España.

December 11, 2008

Matrix initial value problems of the form:

$$\left.\begin{array}{rcl} Y''(x) & = & f(x,Y'(x)) \\ Y(a) & = & Y_0\ ,\ Y'(a)\ =\ Y_1 \end{array}\right\}\ a \leq x \leq b\ , \tag{1}$$

are frequently encountered in different fields of physics and engineering, e.g. problem (1) could be the statement of Newton's second law of motion for a coupled mechanical system in which the forces may depend on time and velocity but not on position. Other problems appear in molecular dynamics, quantum mechanics, or when one uses shooting methods to solve scalar or vectorial problems with boundary values conditions, mainly for scattering problems, [1, 2, 3, 4, 5, 6].

*Acknowledgments. This work has been supported by the *Universidad Politécnica de Valencia* project PAID-06-07/3283.

†e-mail: edefez@imm.upv.es

‡mtung@imm.upv.es

§jjibanez@dsic.upv.es

Problems of type (1) can be written as an extended first-order matrix problem [7]. For this, we consider the matrix-valued function $F : [a,b]\times\mathbb{C}^{2r\times q} \to \mathbb{C}^{2r\times q}$ defined by

$$F(x,W(x)) = \begin{pmatrix} f(x,W_1(x)) \\ W_1(x) \end{pmatrix} \ , \ \text{where } W(x) = \begin{pmatrix} W_1(x) \\ W_2(x) \end{pmatrix} \ ,$$

and the differential problem

$$\left.\begin{array}{rcl} W'(x) & = & F(x,W(x)) \\ W(a) & = & \begin{pmatrix} Y_1 \\ Y_0 \end{pmatrix} \in \mathbb{C}^{2r\times q} \end{array}\right\} \ a \le x \le b \ . \tag{2}$$

Such standard approach, however, involves an increase of the computational cost derived from the increase of the problem dimension.

In the scalar case, cubic splines were used in [8] for the resolution of first-order differential equations, obtaining approximations that, among other advantages, were of class $C^1$ in the interval $[a,b]$. These splines are easy to compute and produce an approximation error of only $O(h^4)$. Recently, this method has been used in the resolution of other scalar problems as discussed in [9, 10, 11], for vector problems (see [12]), linear matrix problems (see [13]), even first-order matrix differential equations [14, 19] and extended to particular second-order matrix case [18]. The present work extends this powerful scheme to the resolution of matrix problems of type (1).

Throughout this work, we will denote by $\mathbb{C}^{p\times q}$ the set of the rectangular $p\times q$ complex matrices, and $\|.\|$ denotes any subordinate matrix norm.

## Construction of the method

Let us consider the initial value problem

$$\left.\begin{array}{rcl} Y''(x) & = & f(x,Y'(x)) \\ Y(a) & = & Y_0 \ , \ Y'(a) \ = \ Y_1 \end{array}\right\} \ a \le x \le b \ , \tag{3}$$

where $Y_0, Y_1, Y(t) \in \mathbb{C}^{r\times q}$, $f : [a,b] \times \mathbb{C}^{r\times q} \to \mathbb{C}^{r\times q}$, $f \in \mathcal{C}^0\,(T)$, with

$$T \ = \ \left\{(x,Y) \ ; \ a \le x \le b \ , \ Y \in \mathbb{C}^{r\times q}\right\} \ , \tag{4}$$

and $f$ fulfills the Lipschitz's condition

$$\|f\,(x,Y_1)\;-\;f\,(x,Y_2)\|\;\leq L\,\|Y_1-Y_2\|\;\;,\;a\leq x\leq b\;,Y_1,Y_2\in\mathbb{C}^{r\times q}\;. \quad (5)$$

Let us consider the partition of the interval $[a,b]$ given by

$$\Delta_{[a,b]}=\{a=x_0\;<\;x_1\;<\;\ldots\;<\;x_n=b\}\;\;,\;x_k=a+kh\;,\;k=0,1,\ldots,n\;, \quad (6)$$

where $h=(b-a)/n$ and $n$ being a positive integer.

We will construct in each subinterval $[a+kh,a+(k+1)h]$ a matrix-cubic spline approximating the solution of problem (3). For the first interval $[a,a+h]$, we consider that the spline is defined by

$$S\Big|_{[a,a+h]}(x)\;=\;Y(a)\;+\;Y'(a)(x-a)\;+\;\frac{1}{2!}Y''(a)(x-a)^2\;+\;\frac{1}{3!}A_0(x-a)^3\;, \quad (7)$$

where $A_0\in\mathbb{C}^{r\times q}$ is a matrix parameter to be determined. With this definition, it is straightforward to check:

$$S\Big|_{[a,a+h]}(a)=Y(a)\;,\quad S'\Big|_{[a,a+h]}(a)=Y'(a)\;,\quad S''\Big|_{[a,a+h]}(a)=Y''(a)=f(a,S'\Big|_{[a,a+h]}(a))\;,$$

and thus, (7) satisfies problem (3) at $x=a$. To fully determine the spline we still must obtain $A_0$. By imposing that (7) is a solution of problem (3) in $x=a+h$:

$$S''\Big|_{[a,a+h]}(a+h)\;=\;f\left(a+h,S'\Big|_{[a,a+h]}(a+h)\right)\;, \quad (8)$$

from (8) we obtain the matrix equation with only one unknown matrix $A_0$:

$$A_0\;=\;\frac{1}{h}\left[f\left(a+h,Y'(a)+Y''(a)h+\frac{1}{2}A_0h^2\right)-Y''(a)\right]\;. \quad (9)$$

Assuming that the matrix equation (9) has a unique solution $A_0$, the spline is totally determined in the interval $[a,a+h]$.

Now, in the interval $[a+h,a+2h]$, the matrix-cubic spline is defined by:

$$S\big|_{[a+h,a+2h]}(x) = S\big|_{[a,a+h]}(a+h)+S'\big|_{[a,a+h]}(a+h)(x-(a+h)) + \frac{1}{2!}S''\big|_{[a,a+h]}(a+h)(x-(a+h))^2+\frac{1}{3!}A_1(x-(a+h))^3 \ . \quad (10)$$

Defined by (10), $S(x)$ is of class $\mathcal{C}^2([a,a+h]\cup[a+h,a+2h])$, and all of the coefficients of the spline $S\big|_{[a+h,a+2h]}(x)$ are determined with the exception of $A_1 \in \mathbb{C}^{r\times q}$.

By construction, spline (10) satisfies the differential equation (3) for $x = a+h$. We can obtain $A_1$ by requiring that the differential equation (3) holds at $x = a+2h$:

$$S''\big|_{[a+h,a+2h]}(a+2h) = f\left(a+2h, S'\big|_{[a+h,a+2h]}(a+2h)\right) \ . \quad (11)$$

By using (11), we obtain the matrix equation with only one unknown matrix $A_1$:

$$A_1 = \frac{1}{h}\left[f\left(a+2h, S'\big|_{[a,a+h]}(a+h)+S''\big|_{[a,a+h]}(a+h)h+\frac{1}{2}A_1h^2\right)-S''\big|_{[a,a+h]}(a+h)\right] \ . \quad (12)$$

Let us assume that the matrix equation (12) has only the solution $A_1$. This way the spline is totally determined in the interval $[a+h, a+2h]$.

Iterating this process, we are in the position to construct the matrix-cubic spline taking
$[a+kh, a+(k+1)h]$ as the next subinterval, and we define the corresponding spline as

$$S\big|_{[a+kh,a+(k+1)h]}(x) = \beta_k(x) + \frac{1}{3!}A_k(x-(a+kh))^3 \ , \quad (13)$$

where

$$\beta_k(x) = \sum_{l=0}^{2}\frac{1}{l!}S^{(l)}\big|_{[a+(k-1)h,a+kh]}(a+kh)(x-(a+kh))^l \ . \quad (14)$$

With this definition, the matrix-cubic spline is $S(x) \in \mathcal{C}^2 \left( \bigcup_{j=0}^{k} [a+jh, a+(j+1)h] \right)$ and fulfills the differential equation (3) at $x = a + kh$. As an additional requirement, we assume that $S(x)$ satisfies the differential equation (3) at $x = a + (k+1)h$:

$$S''_{\big|_{[a+kh,a+(k+1)h]}}(a+(k+1)h) \;=\; f\left(a+(k+1)h, S'_{\big|_{[a+kh,a+(k+1)h]}}(a+(k+1)h)\right) . \tag{15}$$

By employing (15), we obtain the equation with the unknown matrix $A_k$:

$$A_k = \frac{1}{h}\left[ f\left(a+(k+1)h, \beta'_k(a+(k+1)h) + \frac{1}{2}A_k h^2\right) - \beta''_k(a+(k+1)h)\right] . \tag{16}$$

Note that this matrix equation (16) is analogous to equations (9) and (12), when taking $k = 0$ and $k = 1$, respectively. We will show that these equations have a unique solution using a fixed-point argument.

For a fixed $h$, we will consider the matrix function of matrix variable $g : \mathbb{C}^{r\times q} \mapsto \mathbb{C}^{r\times q}$ defined by

$$g(T) = \frac{1}{h}\left[ f\left(a+(k+1)h, \beta'_k(a+(k+1)h) + \frac{1}{2}Th^2\right) - \beta''_k(a+(k+1)h)\right] . \tag{17}$$

Relation (16) holds if and only if $A_k = g(A_k)$, that is, if $A_k$ is a fixed point for function $g(T)$.

Applying the Lipschitz's conditions (5), it follows that

$$\|g(T_1) - g(T_2)\| \leq \frac{Lh}{2}\|T_1 - T_2\|$$

Taking $h < 2/L$ then $Lh/2 < 1$ and $g(T)$ yields a contractive matrix function, which guarantees that equation (16) has unique solutions $A_k$ for $k = 0, 1, \ldots, n-1$. Hence, the matrix-cubic spline is completely determined. Taking components and working in a similar way to the proof of Theorem 5 of [8], the following result has been established:

**Theorem 1** *Let $L$ be the Lipschitz constant defined by (5). If the step size is chosen $h < 2/L$, then there exists a matrix-cubic spline $S(x)$ for each subinterval $[a+kh, a+(k+1)h]$, $k = 0, 1, \ldots, n-1$. If $f \in C^1(T)$, then $\|Y(x) - S(x)\|$ is at least of global order $O(h^2)$ $\forall x \in [a, b]$, where $Y(x)$ is the theoretical solution of* (3).

## Algorithm

The following algorithm is designed to compute the approximate solution of (3) by means of matrix-cubic splines in the interval $[a, b]$ with a global error of the order $O(h^2)$.

- **STEP 1:** Take $n > \dfrac{L(b-a)}{2}$, $h = (b-a)/n$ and the partition $\Delta_{[a,\ b]}$ given by (6).
- **STEP 2:** For $k = 0$, solve the matrix equation (9). Calculate $S_{\big|_{[a,a+h]}}(x)$ defined in (7).
- **STEP 3:** For $k = 1, \ldots, n-1$, solve the matrix equation (16). Calculate $S_{\big|_{[a+kh,a+(k+1)h]}}(x)$ defined in (13).

Depending on the function $f$, matrix equations (9) and (16) can be solved explicitly (see [15]) or by using the iterative method ([16, 17]):

$T^s_{l+1} = g(T^s_l)$ where $T^s_0$ is an arbitrary matrix in $\mathbb{C}^{r\times q}$ for $s = 0, 1, \ldots, n-1$, and $g(T)$ is given by (17).

## Example: A non-linear vector system.

We consider the following non-linear vector differential system:

$$\left.\begin{array}{rcll} y_1''(x) & = & 1 - \cos(x) + \sin(y_2'(x)) + \cos(y_2'(x)) & \\ y_2''(x) & = & 0 & \\ y_1(0) & = & 1, \qquad y_2(0) = 0, & \\ y_1'(0) & = & 0, \qquad y_2'(0) = \pi & \end{array}\right\} \; 0 \le x \le 1. \qquad (18)$$

It is easy to verify that this problem has the exact solution $y_1(x) = \cos(x), y_2(x) = \pi x$. Thus, we can compare our numerical estimates with this solution in order to obtain the exact errors of the approximation.

The system Eq. (18) can be recast in the more compact form

$$\begin{aligned} Y''(x) &= F(x, Y'), \quad Y(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix} \\ Y(0) &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad Y'(0) = \begin{pmatrix} 0 \\ \pi \end{pmatrix}, \end{aligned} \tag{19}$$

where

$$F(x, Y') = \begin{pmatrix} 1 - \cos(x) + \sin(y_2'(x)) + \cos(y_2'(x)) \\ 0 \end{pmatrix}. \tag{20}$$

Thus, we obtain

$$Y''(0) = F(0, Y'(0)) = \begin{pmatrix} -1 \\ 0 \end{pmatrix},$$

and it is not difficult to show that $F(x, Y')$, given by Eq. (20), fulfills the global Lipschitz condition:

$$\|F(x, Y_1) - F(x, Y_2)\| \leq 2 \|Y_1 - Y_2\| \ , \ 0 \leq x \leq 1 \ , Y_1, Y_2 \in \mathbb{R}^2. \tag{21}$$

Next, we determine $L$ calculated by Eq. (5) as $L = 2$. Theorem 1 implies that we need to take $h < 1$, so here we choose $h = 0.1$. To attack the emerging algebraic equations, we have used *Mathematica*[1] Most of the results are generated using its `FindRoot` function. Table 1 summarizes all the numerical estimates, which have been rounded to the fourth relevant digit. In each interval, we evaluated the difference between the estimates of our numerical approach and the exact solution, and then took the Frobenius norm of this difference. The corresponding errors are indicated in the last column.

[1] *Mathematica* is a computer algebra and programming system developed by Wolfram Research, Inc. First released in 1988, *Mathematica* has had a profound effect on the way computers are used in technical and other fields.

| Interval | Approximation | Error |
|---|---|---|
| $[0, 0.1]$ | $\begin{pmatrix} 1 - 0.5x^2 + 0.0083264x^3 \\ 3.1416x \end{pmatrix}$ | $4.1611 \times 10^{-6}$ |
| $[0.1, 0.2]$ | $\begin{pmatrix} 0.99998 + 0.00049709x - 0.50497x^2 + 0.024896x^3 \\ 3.1416x \end{pmatrix}$ | $1.6603 \times 10^{-5}$ |
| $[0.2, 0.3]$ | $\begin{pmatrix} 0.99985 + 0.0024556x - 0.51476x^2 + 0.041217x^3 \\ 3.1416x \end{pmatrix}$ | $3.72009 \times 10^{-5}$ |
| $[0.3, 0.4]$ | $\begin{pmatrix} 0.99942 + 0.0067510x - 0.52908x^2 + 0.057126x^3 \\ 3.1416x \end{pmatrix}$ | $6.57496 \times 10^{-5}$ |
| $[0.4, 0.5]$ | $\begin{pmatrix} 0.99844 + 0.014113x - 0.54749x^2 + 0.072464x^3 \\ 3.1416x \end{pmatrix}$ | $1.01964 \times 10^{-4}$ |
| $[0.5, 0.6]$ | $\begin{pmatrix} 0.99661 + 0.025074x - 0.56941x^2 + 0.087078x^3 \\ 3.1416x \end{pmatrix}$ | $1.45481 \times 10^{-4}$ |
| $[0.6, 0.7]$ | $\begin{pmatrix} 0.9936 + 0.0399x - 0.5941x^2 + 0.1008x^3 \\ 3.1416x \end{pmatrix}$ | $1.95867 \times 10^{-4}$ |
| $[0.7, 0.8]$ | $\begin{pmatrix} 0.9893 + 0.0586x - 0.6209x^2 + 0.1136x^3 \\ 3.1416x \end{pmatrix}$ | $2.52618 \times 10^{-4}$ |
| $[0.8, 0.9]$ | $\begin{pmatrix} 0.9833 + 0.0809x - 0.6487x^2 + 0.1252x^3 \\ 3.1416x \end{pmatrix}$ | $3.15167 \times 10^{-4}$ |
| $[0.9, 1.0]$ | $\begin{pmatrix} 0.9758 + 0.1061x - 0.6767x^2 + 0.1355x^3 \\ 3.1416x \end{pmatrix}$ | $3.8289 \times 10^{-4}$ |

Table 1: Approximation for the vector differential system Eq. (18) in the interval $[0, 1]$ with step size $h = 0.1$.

## Conclusions.

In this work, we have presented a novel method for the numerical treatment of second-order differential matrix systems of the type $Y''(x) = f(x, Y'(x)), x \in [a, b]$, as they are frequently encountered in engineering modeling. Our approach is a generalization of previously developed methods employing matrix-cubic splines for lower-order equations.

There are several important advantages of our proposed method. Firstly, the algorithm is straightforward to implement on numerical and symbolical computer systems or by even using suitable low-level programming languages. Secondly, in the case of second-order differential matrix systems, our method does not require to disentangle the system at hand and reduce it to a higher dimensional system of lower order, commonly practice in problems of this kind. This reduction would only come at the price of increasing computational cost. Thirdly, all spline solutions are by construction already continuous in the interval under consideration. An explicit numerical example have tested the method and have shown that global errors is, at least, only of the order $O(h^2)$. By adapting the step size to a particular problem, in principle, any desired accuracy can be reached.

With these benefits, it is hoped that our approach provides an alternative method to existing ones and may open up new avenues to the numerical integration of second-order models in practical applications.

# References

[1] **P. Marzulli**, Global error estimates for the standard parallel shooting method. *J. Comput. Appl. Math.* 34, pp. 233–241, 1991.

[2] **J.M. Ortega**, *Numerical Analysis: A second Course*. Academic Press, New York, 1972.

[3] **B.W. Shore**, Comparison of matrix methods to the radial Schrödinger eigenvalue equation: The Morse potential. *J. Chemical Physics* 59 (12), pp. 6450–6463, 1972.

[4] **C. Froese**, Numerical solutions of the Hartree-Fock equations. *Can. J. Phys.* 41, pp. 1895–1910, 1963.

[5] **J.R. Claeyssen, G. Canahualpa, C. Jung**, A direct approach to second-order matrix non-classical vibrating equations. *Appl. Numer. Math.* 30 , pp. 65-78, 1999.

[6] **J. F. Zhang**, Optimal control for mechanical vibration systems based on second-order matrix equations. *Mechanical Systems and Signal Processing* 16(1), pp. 61-67, 2002.

[7] **E.A. Coddington, N. Levinson**, *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955.

[8] **F.R. Loscalzo, T.D. Talbot**, Spline function approximations for solutions of ordinary differential equations. *SIAM J. Numer. Anal.*, 4(3), pp. 433–445, 1967.

[9] **E.A. Al-Said**, The use of cubic splines in the numerical solution of a system of second-order boundary value problems. *Comput. Math. Appl.* 42, pp. 861–869, 2001.

[10] **M.K. Kadalbajoo, K.C. Patidar**, Numerical solution of singularly perturbed two-point boundary value problems by spline in tension. *Appl. Math. Comput.*, 131, pp. 299–320, 2002.

[11] **E.A. Al-Said, M.A. Noor**, Cubic splines method for a system of third-order boundary value problems. *Appl. Math. Comput.*, 142, pp. 195–204, 2003.

[12] **G. Micula, A. Revnic**, An implicit numerical spline method for systems for ODEs. *Appl. Math. Comput. 111*, pp. 121–132, 2000.

[13] **E. Defez, L. Soler, A. Hervás, C. Santamaría.** Numerical solutions of matrix differential models using cubic matrix splines. *Comput. Math. Appl.* 50, pp. 693–699, 2005.

[14] **E. Defez, L. Soler, A. Hervás, M.M. Tung.** Numerical solutions of matrix differential models using cubic matrix splines II. *Math. Comput. Modelling*, 46, pp. 657–669, 2007.

[15] **P. Lancaster**, Explicit solutions of linear matrix equations. *SIAM Review* 12, pp. 544–566, 1970.

[16] **P.T. Boggs**, The solution of nonlinear systems of equations by A-stable integration techniques. *SIAM J. Numer. Anal.* 8 (4), pp. 767–785, 1971.

[17] **J.M. Ortega, W.C. Rheinboldt**, *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, 1972.

[18] **A. Borhanifar, R. Abazari**, Numerical Solution of Second-Order Matrix Differential Models Using Cubic Matrix Splines. *Applied Mathematical Sciences* 1 (59), pp. 2927–2937, 2007.

[19] **GUO Qing-wei, Wang Fang**, Approximate solution of matrix differential equations by quartic matrix spline. *Journal of Hefei University of Technology (China)* 30 (11), pp. 1537–1541, 2007.

# Modelling nitrogen dynamics in a citrus orchard

W.A. Contreras[a], A.L. Lidón[b], D. Ginestar[c], R. Bru[c]

[a] Department of Mathematics, Faculty of basic sciences,
University of Pamplona, Km 1 way Bucaramanga, Colombia.

[b] Departamento de Química,
Universidad Politécnica de Valencia.
Camino de Vera, 14, 46022, Valencia, Spain.

[c] Instituto de Matemática Multidisciplinar.
Universidad Politécnica de Valencia.
Camí de Vera, 14, 46022. Valencia. Spain.

December 11, 2008

## 1 Introduction

Groundwater pollution by nitrate is a serious problem in the European Union and in many developed countries. Several models simulating nitrogen transformations and transport in the soil-plant system have been developed and tested over the past years. One of these models is the code LEACHM, developed to simulate water and solute transport in unsaturated or partially saturated soils [1].

In order to facilitate the recomendations for the agricultural soil management it is interesting to develop simple models with a small number of parameters, which provide a fast response. In this paper, we present a simplified model for the nitrogen dynamics in unsaturated soil adapted to citrus orchards. This model is based on the fundamental mechanisms of the processes taking place in the soil and it is coupled with a simplified water trans-

port model, which uses a small amount of parameters and its use requires small experience on numerical modelling.

## 2 Soil water capacity model

Some of the processes described in the soil nitrogen cycle depend on the soil water content and its movement. We have considered a tipping bucket model based on the balance of water in the root profile, that assumes that soil acts as a water reservoir with a storage capacity. Also, it is assumed that plant can not extract water from soil below a certain level and that the evapotranspiration varies with the soil water content in a well-defined pattern [2].

Two models of this kind have been studied. A first model where the soil column is divided into on compartment or layer and a second one, where the soil column has been divided into three layers. These models assume no lateral entry or outlet of water to or from soil (except the runoff) and carries out daily water balances.

## 3 Model for nitrogen in the soil

The processes that involve transformations of soil organic matter, take into account the cycles of carbon and nitrogen, which are closely coupled [3]. For this reason, the modelling of nitrogen on the soil column should take into account the processes involved in organic carbon. To represent these processes, a compartmental system consisting of five compartments, the litter, the humus, the biomas, the amoniumm and the nitrate has been considered.

The carbon balance equation for the litter compartment is given by

$$\frac{dC_l}{dt} = ADD + BD - DEC_l \tag{1}$$

where the term $ADD$ is the plant residues input in the system, the term $BD = k_d C_b$ represents the rate at which carbon returns to the litter compartment due to the death of microbial biomass, $k_d$ is the rate of return and $C_b$ is the carbon concentration in the biomass compartment. Finally, $DEC_l = [\varphi f_d(s) k_l C_b] C_l$ represents the carbon output due to microbial decomposition, where the coefficient $\varphi$ is a non-dimensional factor that accounts for a possible reduction of the decomposition, the constant $k_l$ defines the rate

of decomposition for the litter compartment and $C_l$ is the carbon concentration in the litter compartment. The term, $f_d(s)$, describes soil moisture effects on decomposition [3].

The differential equation that represents the dynamics of nitrogen in litter is

$$\frac{dN_l}{dt} = \frac{ADD}{(C/N)_{add}} + \frac{BD}{(C/N)_b} - \frac{DEC_l}{(C/N)_l} \tag{2}$$

The balance equation for carbon in the humus compartment is

$$\frac{dC_h}{dt} = r_h DEC_l - DEC_h \tag{3}$$

where $r_h$ is the isohumic coefficient. The term $DEC_h = [\varphi f_d(s) k_h C_b] C_h$ represents the humus decomposition, where $k_h$ is the rate of mineralization of the humus, and $C_h$ is the carbon concentration in the humus compartment.

The balance equation for nitrogen is

$$\frac{dN_h}{dt} = r_h \frac{DEC_l}{(C/N)_h} - \frac{DEC_h}{(C/N)_h} \tag{4}$$

The carbon balance in the biomass compartment is given by

$$\frac{dC_b}{dt} = (1 - r_h - r_r)\, DEC_l + (1 - r_r) DEC_h - BD \tag{5}$$

where the constant $r_r$ $(0 \leq r_r \leq 1 - r_h)$ defines the fraction of decomposed organic carbon that goes into respiration ($CO_2$ production).

The balance equation for nitrogen in the biomass compartment is

$$\frac{dN_b}{dt} = \left(1 - r_h \frac{(C/N)_l}{(C/N)_h}\right) \frac{DEC_l}{(C/N)_l} + \frac{DEC_h}{(C/N)_h} - \frac{BD}{(C/N)_b} - \Phi \tag{6}$$

where the term $\Phi$ takes into account the contribution due to either the net mineralization or the immobilization. This term can take values positive or negative in relation to the difference between the rate of gross mineralization and the total rate of immobilization of $NH_4^+$, $IMM^+$ and $NO_3^-$, $IMM^-$.

In the case where the nitrogen supply from immobilization is not enough to ensure a constant $(C/N)_b$, the rates of decomposition are reduced to below their potential values by means of the parameter $\varphi$, [3].. Since the immobilization rate may be limited, mainly by insufficient mineral nitrogen, it is assumed that, maximum level of inmovilization.

The mineral nitrogen in the soil can be modelled by the balance of ammonium and nitrate as,

$$\frac{dN^+}{dt} = AB^+ + MIN - IMM^+ - NIT - LE^+ - UP^+ \tag{7}$$

and

$$\frac{dN^-}{dt} = AB^- + R^- + P^- + NIT - IMM^- - LE^- - UP^- \tag{8}$$

where, $AB^{\pm}$ represents the input fertilizer of ammonium or nitrate, respectively, $R^-$ is the contribution of nitrate with water irrigation, $P^-$ is the contribution nitrate with rainfall, $LE^{\pm}$ represents the output of the system by leaching and $UP^{\pm}$ the output by plant uptake.

The nitrification can be modelled by first order kinetics of the form

$$NIT = f_n(s) k_n C_b N^+ \tag{9}$$

where $k_n$ is the nitrification rate and $f_n(s)$ is a non dimensional term that describes soil moisture effects on nitrification [3].

Ammonium and nitrate are dissolved in soil solution, therefore the main transport mechanism is convection, where the movement of the solute is produced by the bulk water movement through the soil. The plant uptake is also considered.

# 4 Simulation Results

To solve numerically the differential equations representing the different compartments of the system a first order finite differences explicit scheme is used. As the dynamics of the process is smooth, such a scheme will suffice to study the carbon-nitrogen cycle in the soil.

The experimental data used for model calibration were obtained in a citrus orchard (plot Cuñat), located near Valencia city in Spain. The period 31/05/1991 - 10/04/1992, was studied and experimental data for water and nitrogen were measured.

The results obtained with the proposed models of one and three layers, are compared with real measured data and with simulations made with the code LEACHM [1].

The results for the depth water in the profile, obtained with the soil water capacity models with one and three layers and the results for nitrate nitrogen in the soil are shown in Figure 1.

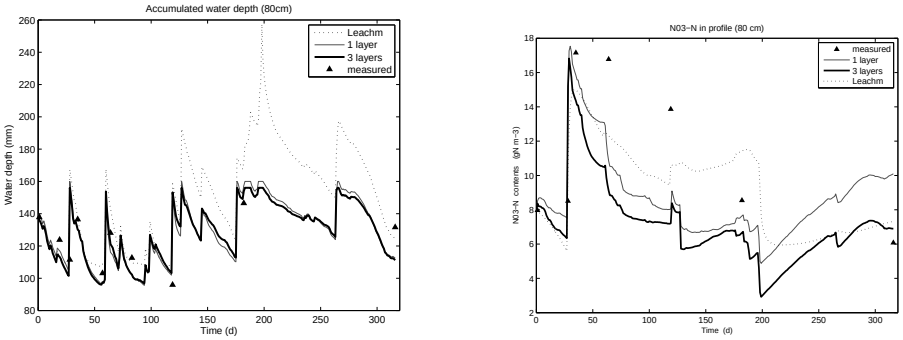



Figure 1: Water depth and soil nitrate contents in soil profile.

# 5 Conclusions

A simple model has been develop to calculate the content of soil nitrogen and nitrate leaching in a citrus orchard. This model has been compared with experimental data obtained in a citrus plot and with the results obtained with LEACHM model. The results show that the model proposed for one and three layers, gives values similar to those obtained with LEACHM. Differences in the models were basically due to the equations used for soil water movement. In general terms, the three layer model performed better than the one layer model, and can be easily used to predict the nitrogen dynamics in citus orchard soil profile with reasonable accuracy.

# References

[1] R.J. Wagenet, J.L. Hutson, *LEACHM: Leaching Estimation and Chemistry Model: A process based model of water and solute movement, transformations, plant uptake and chemical reactions in the unsaturated zone.* Cornell University (1989).

[2] A. Lidon, C. Ramos, A. Rodrigo, Comparison of drainage estimation methods in irrigated citrus orchards. *Irrig. Sci.*, 19 25-36 (1999).

[3] A. Porporato, P. D'Odorico, F. Laio, and I. Rodríguez-Iturbe, Hydrologic control on soil carbon and nitrogen cycles. I. Modeling scheme. *Advances in water resources*, 26(1), 45-58 (2003).

# Collocation methods for the neutron diffusion equation based on a continuos basis of polynomials

S. González-Pintor[a], D. Ginestar[b], G. Verdú[a]

[a] Departamento de Ingeniería Química y Nuclear,
Universidad Politécnica de Valencia.
Camino de Vera, 14, 46022, Valencia, Spain.

[b] Instituto de Matemática Multidisciplinar.
Universidad Politécnica de Valencia.
Camí de Vera, 14, 46022. Valencia. Spain.

December 11, 2008

## 1 Introduction

We will consider the *Lambda modes problem* [1], which for reactors with 1D geometry of length $L_r$, in the approximation of one group of energy,

$$J = -D\frac{d\Phi}{dx} \ , \quad -\frac{d}{dx}\left(D\frac{d\Phi}{dx}\right) + \Sigma_a\Phi = \frac{1}{\lambda}\nu\Sigma_f\Phi \ , \tag{1}$$

where $D(x)$ is the diffusion coefficient, $\Sigma_a(x)$ is the absorption cross section, $\Sigma_f(x)$ is the fission cross section, $\nu$ is the average number of neutrons produced in each fission, $J(x)$ is the neutron current, $\lambda$ is the eigenvalue of the problem and $\Phi(x)$ its corresponding eigenvector. We will use general boundary conditions of the form

$$\alpha^-\Phi(0) + \beta^- D(0)\frac{d\Phi}{dx}(0) = 0 \ , \quad \alpha^+\Phi\left(L_r\right) + \beta^+ D\left(L_r\right)\frac{d\Phi}{dx}\left(L_r\right) = 0 \ , \tag{2}$$

where $\alpha^{\pm}$, $\beta^{\pm}$ are constants defining different albedo, zero flux or zero current conditions.

Realistic problems are dealt with several groups of energy, but the genereralization of the formulation of the methods presented here to more groups of energy is simple.

# 2 Collocation methods

To discretize the neutron diffusion equation (1), the first step is to divide the domain defining the reactor, $\Omega$, into a set of elements $\Omega_e$, $1 \leq e \leq N$, defined by the different materials considered in the reactor.

To develop the collocation methods, each element $\Omega_e = [x_{e-1}, x_e]$ of the reactor is transformed into the reference element $\Omega_{REF} = [0, 1]$ by means of a suitable change of variables, and it is assumed that the solution for the neutron flux over each element $\Omega_e$, can be expanded as

$$\Phi_e(u) = \sum_{i=0}^{K} \Phi_{e,i}\, p_i(u) \ , \ 1 \leq e \leq N \ , \tag{3}$$

being $\{p_i(u)\}_{i=0}^{K}$ a basis of polynomials defined over the reference element $\Omega_{REF}$.

## 2.1 Polynomial basis

The basis of polynomials that will be used for the reference element, $\Omega_{REF}$, is the following one,

$$\begin{aligned} p_0(u) &= 1-u \ , \quad p_1(u) = u \ , \\ p_i(u) &= (1-u)u\, P_{i-2}^{1,1}(2u-1)2^{\frac{3}{2}} \ , \quad 2 \leq i \leq K \ , \end{aligned}$$

where $P_j^{1,1}(x)$ is the Jacobi polynomial of degree $j$ defined in $[-1, 1]$. We will assume that these polynomials vanish out of $\Omega_{REF}$.

The coefficients in expansion (3) are selected in such a way that neutronic flux is continuos in the domain defining the reactor. Continuity conditions for the neutron flux, are fulfilled if the coefficients of the expansions satisfy

$$\Phi_{e,1} = \Phi_{e+1,0} \ , \quad 1 \leq e \leq N-1 \ . \tag{4}$$

The unknowns to be determined by the different collocation methods are the coefficients $\Phi_{e,i}$, of expansions (3), which will named *local* of the method. As the relations (4) hold, the *local* unknowns are not independent. Eliminating the dependent unknowns we obtain the *global* coefficients.

## 2.2 Continuous Pseudospectral Method (CPM)

Any method to discretize the neutron diffusion equation must assure the continuity of the neutron flux and current. The continuity of the neutron flux is assured by the relation (4). To impose the continuity of the neutron current in the different elements $\Omega_e$, of the reactor domain, we impose, in terms of the variable $u$, the following conditions

$$-\frac{D_e}{\Delta x_e}\sum_{i=0}^{K}\Phi_{e,i}\frac{dp_i}{du}(1) = -\frac{D_{e+1}}{\Delta x_e}\sum_{i=0}^{K}\Phi_{e+1,i}\frac{dp_i}{du}(0)\ , \quad 1 \le e \le N-1\ , \tag{5}$$

where $\Delta x_e = x_e - x_{e-1}$.

At the reactor boundaries we will impose the boundary conditions (2) that, in terms of the variable $u$, can be written as

$$\alpha^{-}\Phi_{1,0}+\beta^{-}\frac{D_1}{\Delta x_1}\sum_{i=0}^{K}\Phi_{1,i}\frac{dp_i}{du}(0) = 0\ , \quad \alpha^{+}\Phi_{N,1}+\beta^{+}\frac{D_N}{\Delta x_N}\sum_{i=0}^{K}\Phi_{N,i}\frac{dp_i}{du}(1) = 0\ . \tag{6}$$

Finally, to approximate equation (1) over each element $\Omega_e$ of the mesh in terms of the variable $u$ of the reference domain $\Omega_{REF}$ and the local unknowns, we consider moment-like equations of the form

$$\begin{aligned} &-\frac{D_e}{(\Delta x_e)^2}\sum_{j=0}^{K}\Phi_{e,j}\int_0^1 \frac{d^2p_j}{du^2}(u)p_i(u)\,du + \Sigma_{a,e}\sum_{j=0}^{K}\Phi_{e,j}\int_0^1 p_j(u)p_i(u)\,du \\ &= \frac{1}{\lambda}\nu\Sigma_{f,e}\sum_{j=0}^{K}\Phi_{e,j}\int_0^1 p_j(u)p_i(u)\,\mathrm{d}u\ ,\ 1 \le e \le N,\ 0 \le i \le K-2\ , \end{aligned} \tag{7}$$

where $D_e$, $\Sigma_{a,e}$ and $\nu\Sigma_{f,e}$ are the macroscopic cross sections, considered constant over each element $\Omega_e$.

To optimize the method, these equations are considered in terms of the *global* unknowns.

### 2.3 Point-wise Collocation Method (PCM)

In this method, the continuity of neutron flux and current together with the boundary conditions (2) are imposed in the same way as it is done in the Continuous Pseudospectral Method (CPM).

To complete the set of relations needed to determine the algebraic problem, we impose that the neutronic flux for each element $\Omega_e$, (3), must satisfy the neutron diffusion equation, (1), on a set of collocation points over the reactor domain. This set of collocation points will be the Gauss-Legendre quadrature points in $[-1,1]$, $\{\xi_i\}_{i=0}^{K}$, applied to each element $\Omega_e$. These relations together with the continuity conditions for the neutronic current (5) and the boundary conditions (6), all them considered in terms of the *global* unknowns, define the algebraic approximation of the Lambda modes problem obtained with the PCM method.

### 2.4 Spectral Element Method (SEM)

Together with the spectral methods exposed above, we will consider a finite element method [3]. Let us consider the following functional

$$\begin{aligned} \mathcal{F}(\Phi) &= \int_\Omega D\left(\frac{d\Phi}{dx}(x)\right)^2 dx + \int_\Omega \Sigma_a \Phi^2(x) dx \\ &- \int_\Omega \frac{1}{\lambda}\nu\Sigma_f \Phi^2(x)\, dx + \frac{\alpha^+}{\beta^+}\Phi^2\left(L_r\right) - \frac{\alpha^-}{\beta^-}\Phi^2\left(0\right) \ , \end{aligned} \tag{8}$$

with $\Phi(x)$ belonging to a suitable Sobolev space. $\Phi(x)$ is a stationary point of this functional if, and only if, the following conditions are satisfied: The function $\Phi(x)$ holds the neutron diffusion equation (1). The neutron current is continuous. The function $\Phi(x)$ should satisfies the boundary conditions (2).

Using this result, the Spectral Element Method proposed is based on obtaining an approximation for the stationary point of functional (8). It is assumed that the neutron flux in each element can be expressed in terms of the *global* unknowns, and the stationary point of the functional is obtained making zero the derivatives with respect to these, $\Phi_i$.

## 3 Numerical results

The spectral methods for the neutron diffusion equation presented above, have been implemented into a computer code written in FORTRAN 77, which

solves the resultant algebraic eigenvalue problem for an arbitrary approximation degree, $K$, in the polynomial expansion of the neutron flux.

## 3.1 Homogeneous eigenvalue problem

We consider a homogeneous slab of length 2 cm [2], in the approximation of one group of energy and vacuum boundary conditions, that is, $\alpha^- = 1$, $\alpha^+ = 1$, $\beta^- = 2$ and $\beta^+ = -2$. The nuclear cross sections for this problem are: $D = \frac{1}{3}$, $\Sigma_a = 0.1$, and $\nu\Sigma_f = 0.25$.

The first 2 dominant eigenvalues for this problem obtained from the analytical solution [2] are $\lambda_1 = 0.587489$ and $\lambda_2 = 0.149135$. In Table 1, we show the results obtained for these first 2 dominant eigenvalules using the spectral methods CPM, PCM and SEM for different values of $K$ in the neutron flux expansion.

Table 1: Results for $\lambda_1$ and $\lambda_2$ for the homogeneous eigenvalue problem.

| | *CPM* | | *PCM* | | *SEM* | |
|---|---|---|---|---|---|---|
| $K$ | $\lambda_1(k_{\text{eff}})$ | $\lambda_2$ | $\lambda_1(k_{\text{eff}})$ | $\lambda_2$ | $\lambda_1(k_{\text{eff}})$ | $\lambda_2$ |
| 4 | 0.587484 | 0.141509 | 0.587725 | 0.141509 | 0.587489 | 0.148478 |
| 5 | 0.587484 | 0.149100 | 0.587484 | 0.149677 | 0.587489 | 0.149134 |
| 6 | 0.587489 | 0.149100 | 0.587489 | 0.149100 | 0.587489 | 0.149134 |
| 7 | 0.587489 | 0.149135 | 0.587489 | 0.149136 | 0.587489 | 0.149135 |

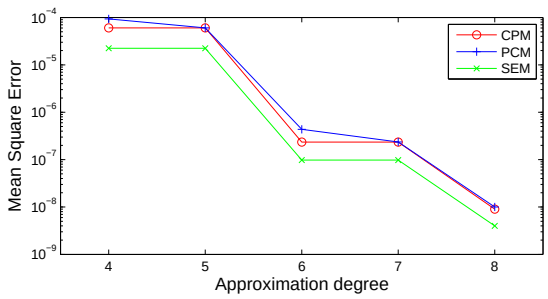
In Figure 1 we show in a logarithmic scale the mean square error for different flux approximation degrees, $K$, for the computed eigenvector corresponding to the eigenvalues $\lambda_1$.

## 3.2 1D typical BWR reactor

Now we consider a more realistic problem, that consist of computing the dominant Lambda modes of a typical 1D BWR reactor. This problem is formulated in the approximation of two groups of energy.

We use zero-flux boundary conditions, i.e., the values $\beta^- = 0$ and $\beta^+ = 0$, that implies $\Phi_{1,2}(x_0) = \Phi_{1,2}(x_N) = 0$. Following the same methodology as the one presented above for the mono energetic approximation, we have developed the spectral methods CPM, PCM and SEM for this problem.

Figure 1: Error for the first eigenvector.




The reactor is divided into 25 fuel assemblies of lenght of 15.24 cm and 2 reflector nodes of 30.48 cm at the boundary.

Table 2 shows the results for the first 2 dominant eigenvalues $\lambda_1$ and $\lambda_2$ obtained with the different spectral methods using different approximation degrees, $K$, in the flux expansion. In this case, we use as a reference the value for the dominant eigenvalues $\lambda_1 = k_{\text{eff}} = 1.005234$ computed with the code KTRAC [5].

Table 2: 1D BWR results for $\lambda_1(k_{\text{eff}})$ and $\lambda_2$.

| | *CPM* | | *PCM* | | *SEM* | |
|---|---|---|---|---|---|---|
| $K$ | $\lambda_1(k_{\text{eff}})$ | $\lambda_2$ | $\lambda_1(k_{\text{eff}})$ | $\lambda_2$ | $\lambda_1(k_{\text{eff}})$ | $\lambda_2$ |
| 4 | 1.005232 | 0.994348 | 1.005233 | 0.994352 | 1.005235 | 0.994351 |
| 5 | 1.005234 | 0.994346 | 1.005233 | 0.994345 | 1.005235 | 0.994349 |
| 6 | 1.005234 | 0.994348 | 1.005234 | 0.994348 | 1.005234 | 0.994349 |
| 7 | 1.005234 | 0.994349 | 1.005234 | 0.994349 | 1.005234 | 0.994349 |

# 4 Conclusions

We have compared three collocation methods for the Lambda modes problem for reactors in 1D geometry based on a continuous basis of polynomials, with the aim of generalizing the method to be able to study reactors in a 3D geometry using a triangular mesh.

To test the performance of the methods, we have considered two benchmark problems, obtaining the SEM method the best results in the the calculations of both the eigenvalues and the eigenvectors.

## Acknowledgments

This work has been partially supported by the Spanish Ministerio de Educación y Ciencia under projects ENE2005-09219-C02-01 and MTM2007-64477-AR07.

## References

[1] A. F. Henry, Nuclear-Reactor Analysis, The MIT Press (1982) Cambridge, MA.

[2] M. Capilla, C.F. Talavera, D. Ginestar, G. Verdú, A nodal collocation method for the calculation of the lambda modes of the PL equations, Ann. Nucl. Energy 32 (2005) 1825–1853.

[3] A. Hébert, Application of a Dual Variational Formulation to Finite Element Reactor Calculations, Ann. Nucl. Energy 20 (1993) 823–845, 1993.

[4] G. Verdú, R. Miró, D. Ginestar, V. Vidal, The implicit restarted Arnoldi method, an efficient alternative to solve the neutron diffusion equation, Annals of Nuclear Energy, 26 (1999) 579-593.

[5] D.W. Niggs, One Dimensional Nodal neutronics Routines for TRAC-BD1 Thermal-Hydraulics Program, EGG-PBS-6379, EG&G (1984) Idaho.

# Analysis of the prevalence of the diabetes using a dynamic model *

Carmen Coll, Alicia Herrero, Elena Sánchez, Néstor Thome †

Instituto de Matemática Multidisciplinar,

Universidad Politécnica de Valencia, Valencia, España

December 11, 2008

In the last years different studies revealing the increase of population who suffers the diabetes disease have appeared. This chronic disease, which nowadays has no cure, consists of a disorder of the metabolism that carries that the level of glucose in blood is higher than the normal one.

In general, two types of diabetes can appear: the type I, where the problem is that the pancreas does not produce or produces less insulin than the body needs for the glucose going on to the cells, and the diabetes of type II, where the problem is that the cells do not answer correctly to the insulin produced. This last type of diabetes is very common and appears mainly in people of mature age.

On the other hand, the high level of glucose in blood has many dangerous effects in the human body, for example it can cause amputations of low members, blindness, cardiac diseases, kidney diseases and even the death. In spite of the advances in the study of diabetes an important factor is to know a priori the prevalence of the above mentioned disease throughout the years.

The knowledge about the evolution of the diabetes disease in the population can determine the election of certain strategies to prevent or reduce the risk of complications of the disease. This fact motivates the use of mathematical models, which allow us to analyze the prevalence of the diabetes within a relatively long period of time. Several works on estimations of the prevalence of the diabetes have been published by different authors [4, 5, 6, 7].

*This paper has been partially supported by DGI grant MTM2007-64477

†mccoll, aherrero, esanchezj, njthome@mat.upv.es

To elaborate our model we have taken into account the number of people diagnosed with diabetes, those who have not been diagnosed by this disease and those who die every year. General factors of risk, as age, sex and race, which can have influence in the prevalence of the diabetes have also been considered in the development of the model. There have not been included the women who have suffered diabetes only during the period of pregnancy.

The mathematical model we will use involves a difference equation, which represents the dynamic of the process and whose coefficients are rectangular matrices. So, the Moore-Penrose inverse will play a very important role in the obtaining of the solution of the model.

Before approaching the model, let us point out some useful properties of the Moore-Penrose inverse. We will denote for $A^\dagger$ the Moore-Penrose inverse of a matrix $A \in \mathbb{R}^{m\times n}$, then the matrix $A^\dagger$ satisfies the following properties

$$AA^\dagger A = A,\ A^\dagger AA^\dagger = A^\dagger,\ (AA^\dagger)^T = AA^\dagger,\ (A^\dagger A)^T = A^\dagger A,$$

where $(A^\dagger A)^T$ denotes the transpose matrix of the matrix $A^\dagger A$. When the matrix $A$ has full row rank it is possible to obtain the expression of its Moore-Penrose inverse using the equality $A^\dagger = A^T(AA^T)^{-1}$. Furthermore, along this work we will use the following notation. Given a collection of matrices $\{A(k) \in \mathbb{R}^{n\times n},\ k \in \mathbb{Z}\}$, from now on, we will denote for $\Phi_A(k, k_0),\ k \geq k_0$, the function

$$\Phi_A(k, k_0) = \begin{cases} I & k = k_0 \\ A(k-1)A(k-2)\dots A(k_0) & k > k_0. \end{cases} \tag{1}$$

Then, more precisely, in this work we consider a mathematical model that will allow us to analyze the evolution of the diabetes in certain populations structured by ages. Firstly, we will define the following vectors in order to construct the model.

- The population vector $p(k) \in \mathbb{R}^{2(m+1)}$, which will be understood as the vector that gives us the information about the number of people in the year $k$ where we are applying the model. It includes the number of people who does not suffer the disease of diabetes, $p_i^{nd}(k)$, and those who are affected by the above mentioned disease, $p_i^d(k)$, in the year $k$. The low subscript $i$ represents the age. To simplify the model, we are going to consider ages from 0 to $m$, taking $m$ as a sufficiently large number which allows us to reject the population older than $m$. Under

all this considerations, the population vector can be represented in the following form

$$p(k) = \left[ \begin{array}{cccccccc} p_0^{nd}(k) & p_0^{d}(k) & \dots & p_i^{nd}(k) & p_i^{d}(k) & \dots & p_m^{nd}(k) & p_m^{d}(k) \end{array} \right].$$

- The migration vector $pm(k) \in \mathbb{R}^{2(m+1)}$, which will be understood as the vector that provides information about the net migration in the year $k$.
- The vector of birth $b(k) = \left[ \begin{array}{cccc} b_1(k) & 0 & \dots & 0 \end{array} \right] \in \mathbb{R}^{2(m+1)}$, where the entry $b_1(k) \in \mathbb{R}^2$ represents the number of births in the year $k$.

Next, in order to establish the model, we observe that the population vector in the year $k+1$ not only depends on the population vector in the year $k$ but also on the births in the year $k+1$ and on the migration of the same year. Thus, a relation between the population vectors in two consecutive years $k$ and $k+1$ exists. In fact, for every age $i$, the following matrices

$$M_i(k) = \left( \begin{array}{cc} \alpha_i(k) & \beta_i \\ 0 & \gamma_i(k) \end{array} \right)$$

give this relation. The entries of the matrices $M_i(k)$ represent the following probabilities: $\alpha_i(k)$ is the probability of not having diabetes at the age $i+1$ not having had it at the age $i$, $\beta_i$ represents the probability of having diabetes at the age $i+1$ not having had it at the age $i$, and finally $\gamma_i(k)$ indicates the probability of having diabetes at the age $i+1$ having been a diabetic at the age $i$. Note that $\gamma_i(k)$ and $\alpha_i(k)$ depend on the instant $k$ ought to we have considered the rates of mortality for the people that was suffering and was not suffering diabetes at this instant. Then, the formulation of the dynamic process is represented by the following difference equation

$$p^T(k+1) = M^T(k)p^T(k) + b^T(k+1) + pm^T(k+1) \tag{2}$$

where

$$M(k) = \left( \begin{array}{ccccc} O & M_0(k) & O & \cdots & O \\ O & O & M_1(k) & \cdots & O \\ O & O & O & \cdots & O \\ \vdots & \vdots & \vdots & & \vdots \\ O & O & O & \cdots & M_{m-1}(k) \\ O & O & O & \cdots & M_m(k) \end{array} \right),$$

is a block matrix.

Notice that the obtained structured model can be used for different cases. For example, it is possible to analyze separately the information for men or women, for different races and even for certain subgroups of risk. This can be done easily making on an estimation of the corresponding parameters from the known information.

Restructuring the equation (2), which represents our model, we can observe that it is possible to interpret the system that represents the prevalence of the diabetes as a singular autonomous system in discrete time, but, unlike the studies presented in ([1, 3]), the matrices of the coefficients are rectangular.

In this stage, one of the main aims is the development of the theoretical basis that allows us to obtain the solution of this type of systems. Consider a singular system in discrete time of the form

$$Ex(k+1) = A(k)x(k), \tag{3}$$

where the coefficient matrices are represented by the following expressions

$$E = \left[\begin{array}{ccc} I & -I & -I \end{array}\right] \in \mathbb{R}^{n\times 3n}, \qquad A(k) = \left[\begin{array}{ccc} M^T(k) & O & O \end{array}\right] \in \mathbb{R}^{n\times 3n},$$

the state vector is given by the population, migration and birth vectors as follows

$$x(k) = \left[\begin{array}{c} p^T(k) \\ b^T(k) \\ pm^T(k) \end{array}\right]$$

and $n = 2(m+1)$. It is easy to prove that $\text{rank}(E) = \text{rank}(E - A(k)) = n$. Moreover, taking into account the properties of the Moore-Penrose inverse matrix we have that

$$(E - A(k))^\dagger = (E - A(k))^T((E - A(k))(E - A(k))^T)^{-1}. \tag{4}$$

On the one hand, if we define the following matrices of size $n \times n$

$$\hat{E}(k) = E(E - A(k))^\dagger \qquad \text{and} \qquad \hat{\text{A}}(\text{k}) = \text{A(k)}(\text{E} - \text{A(k)})^\dagger, \tag{5}$$

its properties will allow us to obtain directly the solution of the system. Concretely, the matrices $\hat{A}(k)$ and $\hat{E}(k)$, defined by the expression (5), satisfy that $\hat{A}(k) = \hat{E}(k) - I$. Then, we can easily prove that both matrices commute.

On the other hand, using (4) we can define the projector $P(k) = (E - A(k))^{\dagger}(E - A(k))$. This projector allows us to make the change of variable $x(k) = P(k)z(k)$ in the dynamic system (3) obtaining a new system of the form

$$EP(k+1)z(k+1) = A(k)P(k)z(k).$$

Now, using the definition of the matrices $\hat{E}(k)$, $\hat{A}(k)$, and of the projector $P(k)$, we can obtain the associated dynamic system given for

$$\hat{E}(k)\hat{x}(k+1) = \hat{A}(k)\hat{x}(k), \tag{6}$$

where the state vector is defined by $\hat{x}(k) = (E - A(k))z(k)$. Moreover, due to the structure of the matrices $E$ and $A(k)$, it is easy to see that $\text{rank}(\hat{E}(k)) = n$. Then, we can obtain a new dynamic system of the form

$$\hat{x}(k+1) = Q(k)\hat{x}(k),$$

where the matrix $Q(k) = \hat{E}(k)^{-1}\hat{A}(k)$. Here, it is easy to see that if we consider an admissible initial vector $\hat{x}(0)$, the above system has a solution in the form $\hat{x}(k) = \Phi_Q(k,0)\hat{x}(0)$, being $\Phi_Q(k,0)$ a matrix defined using the expression (1). Therefore, we have found the solution of the initial model.

Summarizing, in this work a structured mathematical model has been considered, since the coefficient matrices depend on parameters, which will be determine in every case. This model has been interpreted as a dynamic singular system in discrete time where the coefficient matrices are rectangular and depend on the instant that we are considering. The solution of the model has been obtained from the properties of the matrices constructed along the process of resolution and involves the Moore-Penrose inverse matrix.

The validity of the model must be compared with experimental data. However, the obtained result can be used for different cases only changing the parameters that appear in the solution as well as the initial condition, which has to be determined from the information corresponding to every situation.

# References

[1] S. L. Campbell: *Singular systems of differential equations,* Pitman Advanced Publishing Program, (San Francisco, 1980).

[2] L. Dai: *Singular control systems*, Lecture notes in Control and Inform. Sci. 118, Springer-Verlag, (Berlín, 1989).

[3] T. Kaczorek: *Linear Control Systems,* Wiley and Sons, (New York, 1992).

[4] A.F. Amos, D.J. McCarty, P. Zimmet, The rising global burden of diabetes and its complications: estimates and projections to the year 2010, 1997, Diabet Med 14 (Suppl. 5), S1–S85.

[5] A. Honeycutt, J.P. Boyle, K.R. Broglio, T.J. Thompson, T.J. Hoerger, L.S. Geiss, K.M. Venkat, A dynamic Markov model for forecasting diabetes prevalence in the United State through 2050, 2003, Helath Care Management Science 6, 155-164.

[6] H. King, R.E. Aubert, W.H. Herman, Global burden of diabetes, 1995–2025: prevalence,numerical estimates, and projections, 1998, Diabetes Care 21, 1414–1431.

[7] S. Wild, G. Rogilc, A. Green, R. Sicree, H. King, Global prevalence of diabetes, 2004, Diabetes Care 27 (5), 1047-1053.

# Solving random discrete models arising in long-time medicine treatment strategies *

G. Calbo , J.-C. Cortés , L. Jódar †

Instituto de Matemática Multidisciplinar

Universidad Politécnica de Valencia,

Edificio 8G, 2ª, P.O. Box 22012, Valencia, España

December 11, 2008

In recent years, it has become apparent that medical, physical, natural, chemical, economical or engineering systems, classically modeled by deterministic differential equations, can be more satisfactorily represented by certain random counterparts if uncertain effects in the considered phenomena as well as measuring devices are to be taken into account. As a result, an important number of recent published papers have been devoted to present mathematical models based on differential equations containing uncertainty into their formulation. The most part of these contributions has been focused to introduce uncertainty into these models by means the important stochastic process namely white noise which is gaussian and stationary. This type of equations which are driven by white noise and interpreted mathematically as Ito equations are referred in literature as stochastic differential equations [1, 2]. On the other hand, the term random differential equations is reserved for differential equations that introduce randomness by another kind of stochastic processes [3]. Recent interesting contributions in both random and stochastic differential equations line can be found in [4] and [5]. Despite of evident applicability from deterministic framework, the corresponding stochastic or random discrete models require still attention because research's efforts have been focused toward the development of continuous models.

In this talk, we consider random coupled difference systems of the form

$$\vec{X}_{n+1} = \mathbf{A}\vec{X}_n + \vec{B}_n \quad , \quad n = 1, 2, \ldots, \tag{0.1}$$

for a given initial condition $\vec{X}_0$. In (0.1) the coefficient $\mathbf{A}$ is a nonzero deterministic matrix in $\mathbb{R}^{r\times r}$, the input term $\vec{B}_n$ is a vectorial sequence of random variables

*This work has been partially supported by the Spanish M.C.Y.T. and FEDER grant TRA2007–68006–C02–02 and Generalitat Valenciana Grant GVPRE/2008/092

†e-mails: gcalbo@imm.upv.es , jccortes@imm.upv.es , ljodar@imm.upv.es

of size $r \times 1$ and the seed $\vec{X}_0$ is a random vector of the same size.

Random linear matrix difference equations of type (0.1) are of great importance in analyzing dynamical systems involving dynamic states as compartmental systems which are widespread in Medicine as well as Biology [6, 7, 8]. These type of models have been successfully applied in studying the evolution of a concentration of drug between different organs of the human body, [9, 10]. The formulation of these type of systems is based up in mass and energy balance considerations of compartments which interchange substance by means intercompartmental flow laws. However, above references are only devoted to study models of type (0.1) as well as its continuous counterparts in the deterministic framework despite initial concentrations of substance (drugs, medicament, particles in blood, etc.) into the organs and intercompartmental flows are seldom known in a deterministic way in practice. A more realistic approach is considered in this paper, where in a first stage, one considers that initial condition and input-term are modeled by means random variables and stochastic discrete processes, respectively. From a medical practical point of view, the solution of this kind of models allows to take control about the level of drug into each compartment at every period $n$ taking advantage about the best treatment strategy to be selected for patients. Besides of controlling level drug at each period there are others aspects that deserve to be considered in practice. For instance, in aggressive treatments one requires to assure that any concentration of drugs has been disappeared at the end of the treatment, that is, in long-time. In addition, it is frequent in medicine dealing with the dosification of a drug, to guarantee that this dose is bounded below by an efficiency level and upper bounded by a toxicity level, so it is interesting to obtain conditions which guarantee that such dose be bounded. Others interesting questions related to these medical models has been considered in recent contributions, but only in the deterministic scenario [11, 12]. On the other hand, the study of random problems of type (0.1) provides a first stage to consider nonlinear random models by means linearization techniques.

The scalar random variables as well as the stochastic processes involved into (0.1) when $r = 1$ belong to the so-called second order. That is, for an underlying probability space $(\Omega, \mathcal{F}, P)$, the application $X : \Omega \to \mathbb{R}$ is a second order real random variable (2-r.v.) if it satisfies that

$$\mathrm{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f_X(x) dx < \infty, \tag{0.2}$$

where $f_X(x)$ denotes the probability density function of $X$ and $\mathrm{E}[\cdot]$ is the expectation operator. The set of all 2-r.v.'s defined on $(\Omega, \mathcal{F}, P)$ is denoted by $L_2$ and endowed with the norm $\|X\| = \left(\mathrm{E}\left[X^2\right]\right)^{1/2}$ has a Banach space structure. An important fact is that the 2-norm $\|\cdot\|$ in $L_2$ does not provide a Banach algebra

structure, i.e., it is not submultiplicative because the property $\|XY\| \leq \|X\| \, \|Y\|$ does not hold. In practice, this inconvenient causes the introduction of non-trivial extra hypotheses in order get suitable inequalities whose are required in the involved reasoning and computations. With respect the vectorial framework, given a positive integer $r$, a second order random vector of size $r$ is a vector $\vec{X} = \left(X^1, \ldots, X^r\right)^T$ whose entries $X^i$ lie in $L_2$ for $1 \leq i \leq r$ (here $T$ denotes the transposed of a vector or matrix). The set $L_2^r$ of all these vectors with the norm

$$\left\|\vec{X}\right\|_r = \max_{1 \leq i \leq r} \left\|X^i\right\|, \tag{0.3}$$

provides a Banach structure to $L_2^r$. From (0.3) it is evident that mean square convergence in $L_2^r$ is equivalent to the componentwise mean square convergence defined as follows: a sequence $\{X_n : n \geq 0\}$ of 2-r.v.'s is mean square convergent to the 2-r.v. $X$ if and only if $\|X_n - X\| \to 0$ as $n \to +\infty$. Given $\{X_n : n \geq 0\}$ a sequence of random variables in $L_2$ m.s. convergent to $X$, this type of stochastic convergence has the following nice properties (see [3, p.88])

$$\mathrm{E}\left[X_n\right] \xrightarrow[n \to \infty]{} \mathrm{E}\left[X\right] \quad , \quad \mathrm{Var}\left[X_n\right] \xrightarrow[n \to \infty]{} \mathrm{Var}\left[X\right]. \tag{0.4}$$

Firstly, we will obtain the explicit solution of the initial value problem (0.1)

$$\vec{X}_n = \mathbf{A}^n \vec{X}_0 + \sum_{j=0}^{n-1} \mathbf{A}^{n-j-1} \vec{B}_j, \quad n = 0, 1, \ldots, \tag{0.5}$$

for a given initial condition $\vec{X}_0$. Then, if $\rho\left(\cdot\right)$ denotes the spectral radius of a matrix (see, [15]), under hypotheses

$$\rho\left(\mathbf{A}\right) < 1\,, \tag{0.6}$$

and

$$\sum_{n=0}^{\infty} \left\|\vec{B}_n\right\|_r = \sum_{n=0}^{\infty} \max_{1 \leq i \leq r} \left( \left( \mathrm{E}\left[ \left(B_n^i\right)^2 \right] \right)^{\frac{1}{2}} \right) < +\infty\,, \tag{0.7}$$

the mean square convergence of (0.1) will be established.

From a practical medical point of view, it is important to assure that level of concentration of a strong drug will disappear of the human body finally in the long-time once treatment has been finished. However, the above result provides sufficient conditions for m.s. convergence of (0.5) but it does not precise the m.s. limit value. Next, we will exhibit that (0.7) and $\|\mathbf{A}\|_\infty < 1$ (see, [13]) are sufficient conditions under which discrete stochastic process (0.5) converges at the null random vector of size $r$.

On the other hand, one will exhibit different situations where whether above hypothesis (0.6) or (0.7) does not hold, but m.s. convergence of $\vec{X}_n$ can be established. The first one will be when $\mathbf{A} = \mathbf{I}$, being $\mathbf{I}$ the identity matrix of size $r$ (in this case, hypothesis (0.6) does not hold) and in the second one will consider that $\mathbf{A}$ is a matrix of class $\mathfrak{M}$ (see, [15]) such that $\rho(\mathbf{A}) = 1$ and the initial condition is the null random vector of size $r$ (this situation takes place often in medicine applications, for instance, when a patient initialize a treatment) and moreover hypothesis (0.7) holds.

After providing sufficient conditions for assuring m.s. convergence, we will establish conditions on coefficients $\mathbf{A}$ and $\vec{B}_n$ in order to guarantee that solution of (0.1) is not mean square unstable, in the sense that, it remains m.s. bounded or may be m.s. convergent, when $n$ goes to infinity. For this goal, we will take advantage from a majorant deterministic scalar difference equation associated to (0.1), that is, a scalar equation which solution $y_n$ is always greater than the $r$-norm of the solution $\vec{X}_n$ of (0.1). Then imposing the well-known boundedness conditions on $y_n$, it will be assured the corresponding property for $\vec{X}_n$. In this way, we will establish that under condition $\|\mathbf{A}\|_\infty < 1$ and

$$\left\{\vec{B}_n : \, n \geq 0\right\} \quad \text{is m.s. uniformly bounded :} \quad \left\|\vec{B}_n\right\|_r < M_B, \quad \forall\, n \geq 0, \tag{0.8}$$

the vector stochastic discrete process solution $\vec{X}_n$ of the problem (0.1) is mean square bounded as follows

$$\left\|\vec{X}_n\right\|_r \leq \max\left(\left\|\vec{X}_0\right\|_r, \frac{M_B}{1 - \|\mathbf{A}\|_\infty}\right), \tag{0.9}$$

where $M_B$ is given by (0.8).

It is frequent in medicine dealing with the dosification of a drug, to guarantee that this dose is bounded below by an efficiency level and upper bounded by a toxicity level, so it is interesting to obtain conditions which guarantee that such dose be bounded. In this sense we will establish that if $\mathbf{A}$ is of class $\mathfrak{M}$ such that $\rho(\mathbf{A}) = 1$ and the partial sums of input term $\vec{B}_n$ are bounded then for every initial condition $\vec{X}_0$ the discrete stochastic process solution $\vec{X}_n$ is m.s. bounded.

On the other hand, once the discrete stochastic process solution as well as sufficient conditions for its m.s. convergence has been established, we will deal with computing its main statistic properties such that mean and variance functions. For the scalar framework, that is, taking $r = 1$, the expectation is given by

$$\mathrm{E}\left[X_n\right] = a^n \mathrm{E}\left[X_0\right] + \sum_{j=0}^{n-1} a^{n-j-1} \mathrm{E}\left[B_j\right], \quad n = 0, 1, \ldots, \tag{0.10}$$

for a prefixed initial condition $\mathrm{E}[X_0]$. On the other hand, the variance expression is given by

$$\mathrm{Var}[X_n] = a^{2n}\left\{\mathrm{Var}[X_0] + \sum_{j=0}^{n-1}\sum_{k=0}^{n-1} a^{-(j+k+2)}\mathrm{Cov}[B_j, B_k]\right\}, \tag{0.11}$$

whenever pairwise independence hypothesis between initial condition $X_0$ and input term $\{B_n : n \geq 0\}$ holds.

Finally, we will show several illustrative examples arising in long-time medicine treatment. Next, we show the framework and main computations involved of the first one.

**Example 0.1** *Let us consider a random linear long-time medicine treatment strategy by means the following discrete model*

$$X_{n+1} = a\,X_n + \frac{1}{n+1}\left(b + \cos\left((n+1)U\right)\right), \quad n = 0, 1, 2, \ldots \tag{0.12}$$

*where $X_n$ denotes the concentration of drug contained into the body at period $n$. Under standard behaviour and from individual characteristics (like weight, height, age, sex, etc), it is realistic to assume that the delivery rate of drug is known with certainty. Let us suppose, that it is $a = 1/2$, between consecutive periods. On the other hand, medical treatment for chronic illness considers an input of medicament at each period, which it is decreasing but it has a constant term $b$ and, another one that it is oscillatory, namely given by the sequence $(1/(n+1))\,(2 + \cos((n+1)U)) \geq 0$, for each $n = 0, 1, 2, \ldots$, where $U$ is a uniform r.v. on the interval $[0, \pi]$, i.e., $U \sim Un([0, \pi])$. Also, one assumes that the initial concentration of drug in the patient is unknown in a deterministic way, but it is modelled through an exponential 2-r.v. of parameter $\lambda = 1$, that is, $X_0 \sim Exp(\lambda = 1)$, which it is independent of r.v. $U$.*
*From (0.5), the concentration of drug at each period is given by*

$$X_n = \left(\frac{1}{2}\right)^n X_0 + \sum_{j=0}^{n-1}\left(\frac{1}{2}\right)^{n-j-1}\frac{1}{j+1}\left(2 + \cos\left((j+1)U\right)\right), \quad n = 0, 1, 2, \ldots \tag{0.13}$$

*From a medical point of view, it is important to assure in the treatment of strong illness that, concentration of invasive medicament vanishes of the organs when treatment has been finished. Regarding its asymptotic behaviour, note that conditions (0.6) and (0.7)*

$$|a| = \frac{1}{2} < 1 \quad , \quad \sum_{n=0}^{\infty} E\left[(B_n)^2\right] = \sum_{n=0}^{\infty}\frac{9}{2(1+n)^2} = \frac{3\pi^2}{6} < +\infty,$$

*hold. Therefore* $\underset{n\to\infty}{\text{l.i.m.}}\, X_n = 0$.
*In order to provide statistical information about the discrete process solution*

*(0.13), we compute the expectation of random input*

$$E[B_n] = \frac{2}{1+n}\,, \quad n = 0, 1, \ldots,$$

*as well as its covariance*

$$\begin{aligned} Cov[B_j, B_k] &= E[B_j\, B_k] - E[B_j]\; E[B_k] \\ &= \frac{1}{(j+1)\,(k+1)} \{E[\cos((j+1)\,U)\; \cos((k+1)\,U)] \\ &+ \; 2\,E[\cos((k+1)\,U)] + 2\,E[\cos((j+1)\,U)]\} \\ &= \begin{cases} 0 & \textit{if} \quad j \neq k \\ \dfrac{1}{2\,(j+1)^2} & \textit{if} \quad j = k\,, \end{cases} \end{aligned}$$

*where we have used that* $E[\cos((k+1)U)] = 0$ *and*

$$E[\cos((j+1)U)\cos((k+1)U)] = \begin{cases} 0 & \textit{if} \quad j \neq k \\ \dfrac{1}{2} & \textit{if} \quad j = k \end{cases} \qquad j, k = 0, 1, \ldots$$

*Then by (0.10), (0.11) and taking into account that* $E[X_0] = 1$ *and* $Var[X_0] = 1$ *one gets*

$$E[X_n] = \left(\frac{1}{2}\right)^n \left(1 + 4\sum_{j=0}^{n-1} \frac{2^j}{1+j}\right), \quad n = 0, 1, \ldots \tag{0.14}$$

*and*

$$Var[X_n] = \left(\frac{1}{4}\right)^n \left(1 + 2\sum_{j=0}^{n-1} \frac{4^j}{(1+j)^2}\right), \quad n = 0, 1, \ldots \tag{0.15}$$

*Finally, it is worthwhile to point out that according with property (0.4) one gets*

$$E[X_n] \xrightarrow[n\to\infty]{} 0 \quad , \quad Var[X_n] \xrightarrow[n\to\infty]{} 0.$$

# References


[1] D. Henderson and P. Plaschko, *Stochastic Differential Equations in Science and Engineering*, World Scientific, Singapore (2006).

[2] E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin (1992).

[3] T.T. Soong, *Random Differential Equations in Science and Engineering*, Academic Press, New York (1973).

[4] M. El-Tawil, W. El-Tahan, A. Hussein A proposed technique of SFEM on solving ordinary random differential equation, *Appl. Math. Computat.* **161** 35–47 (2005).

[5] M. El-Tawil, The approximate solutions of some stochastic differential equations using transformations, *Appl. Math. Computat.* **164** 167–178 (2005).

[6] D.H. Anderson, *Compartmental Modeling and Tracer Kinetics*, Lecture Notes in Biomathematics, vol. 50, Springer-Verlag, Berlin (1983).

[7] J.A. Jacquez, *Compartmental Analysis in Biology and Medicine*, 2nd. ed., University of Michigan Press, Michigan (1985).

[8] D.S. Berstein, D.C. Hyland, Compartmental modeling and second-moment analysis of state space systems, *SIAM J. Matrix Anal. Appl.*, **14** 880–901 (1993).

[9] M. Gibaldi, D. Perrier, *Pharmacokinetics*, Marcel Dekker, New York (1975).

[10] K. Godfrey, *Compartmental Models and Their Applications*, Marcel Dekker, New York (1975).

[11] W.M. Haddad, V. Chellaboina, E. August, Stability and dissipativity for discrete-time nonnegative and compartmental dynamical system, Proc. IEEE Conference on Decision and Control (Florida) 4236–4241 (2001).

[12] V. Chellaboina, W.M. Haddad, J.M. Bailey, J. Ramakrishnan, On the absence of oscillations in compartmental dynamics systems, Proc. IEEE Conference on Decision and Control (Nevada), 1663–1668 (2002).

[13] G. Golub and C.F. Van Loan, *Matrix Computations*, Third Ed., The Johns Hopkins University Press, Baltimore MD (1996).

[14] G. Calbo, J.C. Cortés, L. Jódar, Random analytic solution of coupled differential models with uncertain initial condition and source term, *Computers Math. Applic.* **56** 785–798 (2008).

[15] J.M. Ortega, *Numerical Analysis. A Second Course*, SIAM, Philadelphia (1990).

[16] M. Loève, *Teoría de la Probabilidad*, Tecnos, Madrid (1976). (*In Spanish*).

[17] J.L. Doob, *Stochastic Processes*, John Wiley & Sons, New York (1953).

[18] B. Kegan and R.W. West, Modeling the simple epidemic with deterministic differential equations and random initial conditions, *Math. Biosci.* **195** 179–193 (2005).

[19] J.D. Hamilton, *Time Series Analysis*, Princeton University Press, New Jersey (1994).

# Quantitative patterns in ecological trophic networks *

Cristina Jordán and Juan R. Torregrosa

Instituto de Matemática Multidisciplinar,

Universidad Politécnica de Valencia,

Camino de Vera, s/n,

46022 Valencia, España

cjordan@mat.upv.es / jrtorre@mat.upv.es

December 11, 2008

## 1 Previous considerations

The network trophic theory tries to explain the working of basic ecosystems by studying the trophic connection of species. It has been proved that trophic networks of very divers ecosystems such as oceans, deserts or lakes, present regularities in their structure, which suggests that there are some mechanisms common to all these types of trophic networks (see [1] and [2]). A basic model, which helps to understand some of the rules that might be performed in ecosystems is the *Trophic Cascade model*. The main rules of this model are the unidirectionality of preying from carnivores to herbivorous and from herbivorous to primary producers and the exigency of a trophic link from consumers due to their heterotrophic nature (viable ecological trophic networks). More specifically:

1. Level one of the network must be made up of at least one trophic specie, that can be prey but no predator (for example the herbivorous),

*This research was supported by Ministerio de Ciencia y Tecnología MTM2007-64477 and CGL2006-02891/BOS

2. The following levels will only have species that are predator of some prey of the immediately lower level. Then it is not allowed:

   **a)** Cannibalism (to eat members of its own colony),

   **b)** Cycles, that is, if $A$ eats B, and $B$ eats $C$, this one can no eats $A$,

   **c)** Omnivory, that is, a predator can not eat of several lower levels.

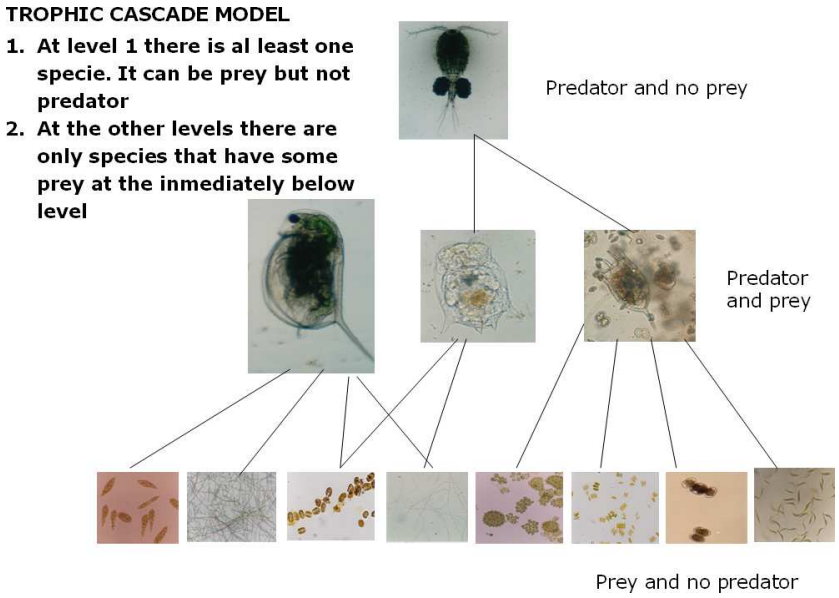
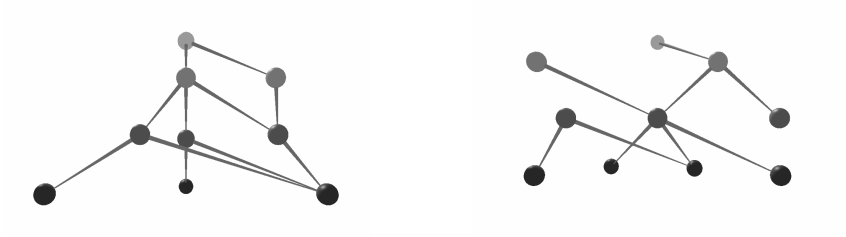



Figure 1: Species of a network of microscopic organisms of the water of a lake: microalgas, herbivorous and carnivores

So, in order to modelize the above mentioned situation, we coin the concept *viable network of n levels*, as that network of $n$ levels where the number of species at the first level and the number of links between the two first levels are nonzero, that there is not loops, if there is not a link between the specie $e_{ij}$ of level $i$ and any specie of level $i-1$, then there is not a link between $e_{ij}$ and any other specie of level $i+1$, and besides there is only links between levels $i+1$ and $i$, $i = 1, 2, \ldots$ As we will see, all these conditions reduce the number of networks that we must consider. Figure 2 shows examples of viable and non-viable networks.

In the present work we develop and implement an algorithm that, given a maximum number of species at each level, provides:



Viable network Non-viable network

Figure 2: Examples

1. The number of all the possible networks, if we consider no restriction
2. A classification of all the networks from the number of links between levels and the number of possibilities of viable networks in each case

The obtained results allow us to assure, in each case, if the real networks observed occur more frequently than if their was at random and to conclude that their relations are the most stable.

# 2 Algorithm

The general process of the developed program consists of finding out the links distribution possibilities among the different trophic levels. After one of these possibilities is chosen, the algorithm considers iteratively, beginning by $l = 1$, the different combinations of links between levels $l$ and $l + 1$. The process just keep the viable networks. A running example with five trophic levels is shown. Implementation has been made with the symbolic/numerical software MatLab 7.3.

The program is formed by some files .m that are recursively called among them. The first of these files, called *root*, is the boot of the program. It asks us as an enter parameter, the number of links of the trophic networks we are interested in. It also asks for the maximum number of species at each level, and name them with the letters of the alphabet, $a, b, c, \ldots$. The program calcules now the cartesian products of the sets of species at levels $l$ and $l+1$, $l = 1, 2, \ldots, 4$, and it builds a new chain by concatenating these products. In a trophic network each of its links has associated one of the elements of

this cartesian product, element that we will use to name it. At this moment, we have a table in which all the possible networks, classified by the number of links, are gathered. Specifically, if we consider the network of row $i$ of the table, column $j$ of this row gives the number of links between levels $j$ and $j+1$. The last element of row $i$ gives the number of networks with the mentioned characteristics, if we do not take into account any restriction.

We must now to select the type of network that we are going to analyze. The study is different when we consider the links between the two first levels than when we consider that which are between levels $l$ and $l+1$ with $l > 1$, (see Figure 3). Therefore we define a variable $h$, initialized as $h = 1$, that will can be upgraded when we change of level. We find now a call to the procedure *Combinaciones*. Depending on the value of $h$ (if $h$ is 1 or greater than 1), the program selects between calling to procedure *ElegirComp* or to *ElegirComo*, where it must choose again, in this case between *Com#aristf* or *Com#aristif* respectively, where the symbol # means the number of links chosen.

The elements of the cartesian product corresponding to the studied levels are now considered. If $h = 1$ and the analyzed network has $m$ links between the first and second level we consider, one by one, all the combinations of $m$ elements taken from the cartesian product. We construct chains of length $m$ with the second elements of each of these $m$ pairs and define a vector $V$ in which components we will store the sets of elements that form the chains and a meter vector that will show how many times each of them has appeared. Let $\alpha$ be one of those sets. By calling to procedure *finales1y2*, $\alpha$ is compared with the elements of chains of $V$. If they match, the corresponding component of meter vector is increased in 1. Else, $\alpha$ is added to $V$, and in the same way a new component of the meter vector is added and initialized with 1. When all the combinations have been analyzed, $h$ is increased. Now come back to the procedure *Combinaciones*, after to *ElegirComo* and then to *Com#aristif*. Again, the cartesian product corresponding to the levels that we are analyzing is considered and the different combinations of # pairs that can be formed are picked one by one.

Let $C$ be one of these combinations. First, we construct a chain $\beta$ with the first # elements of the pairs of $C$. We compare $\beta$ with the set of end elements obtained in vector $V$ of the last iteration. If it is not a subset of any of these final sets it is rejected. Else, we construct a new chain $\alpha$ with the second elements of $C$ and we proceed as case $h = 1$, that is, we define a new vector $V'$ with an associated meter vector and analyze and modify

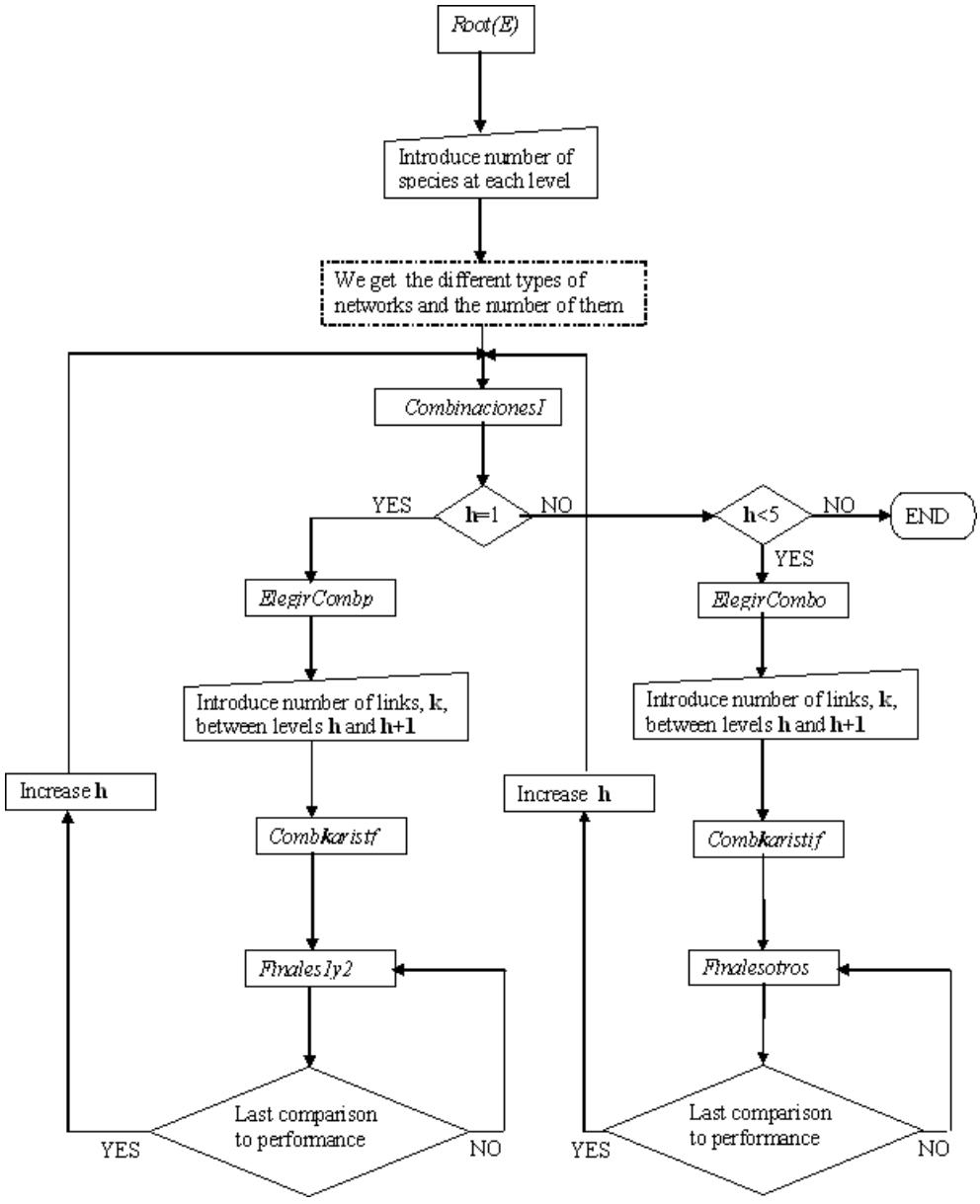



Figure 3: Diagram of flow

adequately $V'$ and its meter vector, according to $\alpha$ is or not subset of some of the elements of de $V'$. All the cases studied, $h$ is again increased and we come back to procedure *Combinaciones*.

# References

[1] J.E. Cohen, F. Briand, and C. M. Newman, *Community food webs: Data and Theory*, Biomathematics, Vol. 20, Springer-Verlag, Berlin (1990).

[2] D.B. Stouffer, J. Camacho, R. Guimera, C.A. Ng and L.A.N. Amaral, Quantitative patterns in the structure of model and empirical food webs, *Ecology* **86** (5) 1301-1311 (2005).

[3] C. Jordán, J. Larrosa, J. R. Torregrosa and C. Rojo, ¿Cuántas comunidades ecológicas podríamos observar? *Proceedings of XI Conferencia española y Primer Encuentro Iberoamericano de Biometría* , 407-409 (2007).

[4] C. Rojo, M.A. Rodrigo and M. Álvarez-Cobelas, Plankton diversity in the outcome of an assembly process, *Verh. Internat. Verein. Limnol.* **29** (3) 1906-1908 (2006).

# Handling occlusion in stereo tracking

Eduardo Parrilla, Jaime Riera,
Juan-R. Torregrosa, José-L. Hueso *
Instituto de Matemática Multidisciplinar,
Universidad Politécnica de Valencia,
Camino de Vera s/n, 46022 Valencia, España

December 11, 2008

## 1 Introduction

In previous works [1], we have studied a system that combines stereoscopic vision [2] and optical flow algorithms [3] for object tracking in a three-dimensional space. One of the most important problems of this technique is that this method is not able to handle the occlusion of the moving objects. For solving this occlusion problem, we propose the use of adaptive filters [4, 5] and neural networks [6] to predict the expected 3D velocities of the objects. The use of adaptative filters and neural networks for handling occlusion has been tested successfully in optical flow algorithms for 2D object tracking [7].

*e-mails: edparber@fis.upv.es, jriera@fis.upv.es, jrtorre@mat.upv.es, jl-hueso@mat.upv.es

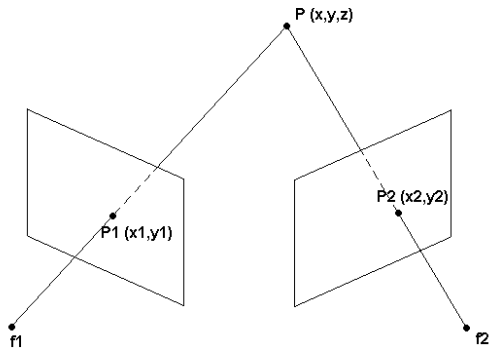



Figure 1: Triangulation principle

# 2 Optical flow

Optical flow is an approximation to the 2-d motion field of an image sequence, that is a projection of the 3-d velocities of surface points onto the imaging surface. Optical flow computation is based on supposing that intensity variations of the image only depend on displacements of the objects, without considering other causes [8]. There are many techniques that consider different restrictions to calculate the optical flow. One of these techniques was proposed by Lucas and Kanade in 1981 [9] and has been succesfully proved in different applications [10, 11].

# 3 Stereoscopic vision

Stereo vision [12] consists in, given a point $P$ from which we know their projections $P_1$ and $P_2$ in the images of a pair of cameras, calculating its position in the space by using the triangulation principle (Fig. 1).

The main problem of stereo vision is to find the correspondences between the points of the different images (disparity), i.e, given the projection $P_1$ in the left image, to find what point of the right image corresponds to the projection $P_2$, to situate in the space the point $P$. We have programmed a

correlation search method that has provided satisfactory results.

# 4 Adaptive filters

An adaptive filter is a filter whose transfer function varies with time, as a function of the input signal. Filter coefficients are unknown when the filter is designed, and they are calculated and updated according to an optimizing algorithm. In this article, we study a two-dimensional recursive least squares (RLS) filter.

## 4.1 Two-dimensional adaptive filter

We use adaptive filters to predict instantaneous velocities in order to compare them with the velocities obtained by using optical flow and stereo vision. For the component $v_y$, we use a simple adaptive filter because this component is the same in the two video sequences. On the other hand, there are two components $v_{Lx}$ and $v_{Rx}$ that correspond to the velocities in the left and right images. In this case, we will adapt the RLS filter to operate with two signals simultaneously.

We start from two signals $x_{Ln}$ and $x_{Rn}$, $n = 0, \ldots, N_T$, and we assume for the signals an AR($p$) model

$$\begin{aligned} x_{Lk} &= -(a_1)_k\, x_{Lk-1} - (a_2)_k\, x_{Rk-1} - \cdots \\ &\quad \cdots - (a_{2p-1})_k\, x_{Lk-p} - (a_{2p})_k\, x_{Rk-p} + \varepsilon_k \ , \\ x_{Rk} &= -(b_1)_k\, x_{Lk-1} - (b_2)_k\, x_{Rk-1} - \cdots \\ &\quad \cdots - (b_{2p-1})_k\, x_{Lk-p} - (b_{2p})_k\, x_{Rk-p} + \varepsilon'_k \ , \end{aligned} \tag{1}$$

In this case, $x_{Lk}$ and $x_{Rk}$ will depend on the previous samples of both signals.

If we apply the RLS filter theory [13] to these systems of equations, we will obtain an adaptive filter for two correlated signals. In this way, we can

calculate the filter coefficients by using a recursion and estimate the following sample of the signals

$$\begin{aligned} x_{LN+1} &= c_{N+1}^T a_{N+1} \ , \\ x_{RN+1} &= c_{N+1}^T b_{N+1} \ , \end{aligned} \tag{2}$$

where

$$c_{N+1}^T = [x_{LN}, x_{RN}, x_{LN-1}, x_{RN-1}, \cdots, x_{LN-p+1}, x_{RN-p+1}] \ .$$

# 5 Neural networks

Another method that we have used to predict the velocities and handle occlusion is the use of neural networks. They are a form of multiprocessor computer system with simple processing elements, a high degree of interconnection and an adaptive interaction between elements. In this work, we concentrate on a Multilayer Perceptron (MLP) network [6].

## 5.1 Multilayer Perceptron

A multilayer perceptron (MLP) is a network of simple neurons called perceptrons. The perceptron is a type of artificial neural network introduced by Frank Rosenblatt in 1958 [14]. A typical multilayer perceptron is formed by an input layer, one or more hidden layers, and an output layer. In this case, we use a network with some inputs and two outputs. Inputs are composed of samples of the signals of the left and right images and the two outputs are the prediction of the next left and right samples. The MLP is trained by supervised learning using the back-propagation algorithm [15].

# 6 Handling occlusion

In order to predict the velocities of the objects, we compare the difference between the estimate velocity and the velocity calculated by means of the optical flow with a certain tolerance value *tol*. Tolerance values are a critical factor for the correct operation of the proposed algorithms. In a previous work [7], we have demonstrated that an optimum value of tolerance is $tol_{L,RN} = k\,|v_{L,RN-1}|$.

# 7 Numerical results

The two methods explained above have been used to analyze synthetic and real video sequences (Fig. 2). In all the examples, we have used windows of $7 \times 7$ pixels to calculate optical flow and disparity. In the different sequences that have been analyzed we have used a value of $k = 0.75$ for the tolerance.

For the adaptative filter, we have used an order filter of 2 and a forgetting factor of $\lambda = 0.99$. In the case of the MLP network, we have used a network with $N_{ent} = 2 \times 7$ inputs and $N_{oc} = 5$ neurons in the hidden layer.

In all the examples, objects are successfully tracked along all the frames, there is not any error when they disappear behind the obstacles and their trajectory is predicted correctly .

# 8 Conclusions

In this article, we have studied a system to track objects in the three-dimensional space by combining optical flow and stereo vision.

We have analyzed the occlusion problem in object tracking in a stereo video sequence. We have proposed the use of adaptive filters and neural networks to predict the velocity of the object in the left and right sequences. We have adapted a RLS filter to work with two correlated signals simultaneously.

Finally, we have shown two examples that verify the efficiency of the algo-

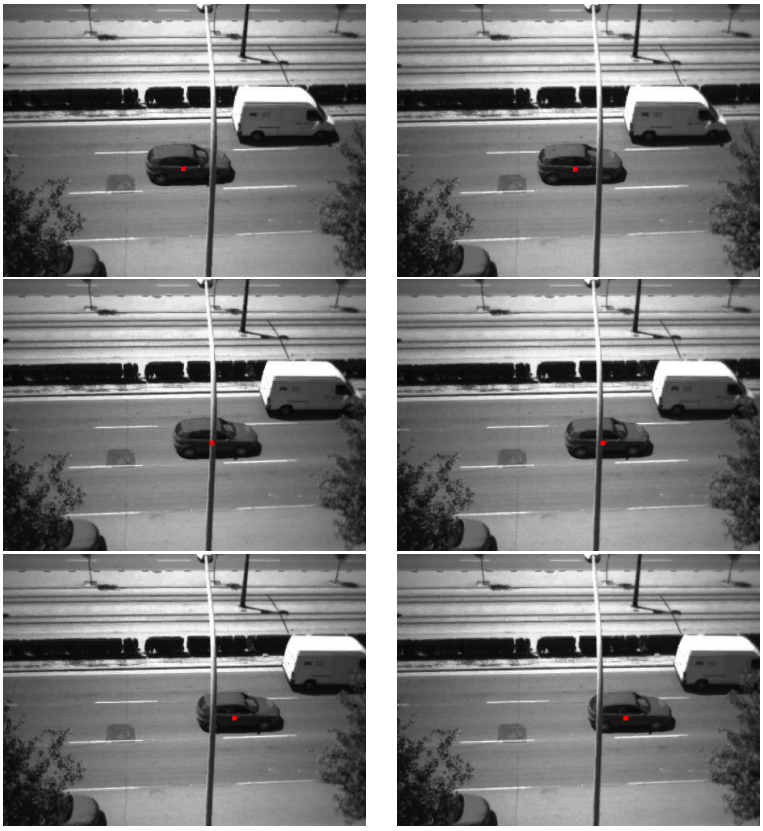


Figure 2: RLS tracking in a real sequence

rithms.

# References

[1] E. Parrilla, J. Riera, M.H. Gimnez, J.R. Torregrosa, J.L. Hueso, Cálculo de velocidades mediante un sistema estereoscópico y algoritmos de flujo óptico, Congreso de Ecuaciones Diferenciales y Aplicaciones (2005).

[2] S.D. Cochran, G. Medioni, 3-D Surface Description from Binocular Stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence, **14**(10) (1992), 981-994.

[3] E. Trucco, K. Plakas, Video Tracking: A Concise Survey, IEEE Journal of Oceanic Engineering **31**(2) (2006), 520-529.

[4] S. Haykin, Adaptive Filter Theory, Prentice-Hall, Englewood Cliffs NJ (1986).

[5] L. Ljung, System Identification. Theory for the User, Prentice-Hall, Upper Saddle River NJ. (1999).

[6] R. Rojas, Neural Networks. A systematic introduction, Springer (1996).

[7] E. Parrilla, D.Ginestar, J.L. Hueso, J. Riera, J.R. Torregrosa, Handling occlusion in optical flow algorithms for object tracking, Computers and Mathematics with Applications **56**(3) (2008) 733-742.

[8] B.K.P. Horn, B.G. Schunck, Determining optical flow,Artificial Intelligence **17**(1-3) (1981) 185-203.

[9] B.D. Lukas, T. Kanade, An iterative image registration technique with an application to stereovision, Proceedings of Imaging Understanding Workshop (1981) 121-130.

[10] F. Bourel, C.C. Chibelushi, A.A. Low, Recognition of facial expressions in the presence of occlusion, Proceedings of the Twelfth British Machine Vision Conference, **1** (2001) 213-222.

[11] E. Parrilla, J. Riera, M.H. Giménez, J.R. Torregrosa and J.L. Hueso, Vehicle tracking in urban routes by means of Lucas&Kanade algorithm, Proceedings of the International Conference on Computational and Mathematical Methods in Science and Engineering (2005) 438-445.

[12] S.D. Cochran, Surface description from binocular stereo, Dissertation presented in partial fulfillment of the requirements for the degree Doctor of Philosophy **1** (1990).

[13] K. Ogata, Discrete-Time Control Systems, Prentice Hall, Upper Saddle River, NJ (1987).

[14] F. Rosenblatt, The Perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review **65** (1958) 386-408.

[15] D. Rumelhart, J. McClelland, Parallel Distributed Processing, MIT Press, Cambridge, MA (1986).

# Numerical integration of differential Riccati equations arising in boundary value problems*

S. Blanes, E. Ponsoda†

Instituto de Matemática Multidisciplinar

Universidad Politécnica de Valencia

46022 Valencia. España.

December 11, 2008

## 1 Introduction

Let us consider the two-point boundary value problem defined by the linear differential equation

$$\mathbf{y'} = \begin{bmatrix} \mathbf{y'}_1 \\ \mathbf{y'}_2 \end{bmatrix} = \begin{bmatrix} A(t) & B(t) \\ C(t) & D(t) \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \qquad 0 \le t \le T \tag{1}$$

with separated boundary conditions

$$[\, K_{11} \;\; K_{12} \,] \begin{bmatrix} \mathbf{y}_1(0) \\ \mathbf{y}_2(0) \end{bmatrix} = \gamma_1, \qquad [\, K_{21} \;\; K_{22} \,] \begin{bmatrix} \mathbf{y}_1(T) \\ \mathbf{y}_2(T) \end{bmatrix} = \gamma_2 \,. \tag{2}$$

Here, $A \in \mathbb{C}^{q\times q}$, $B \in \mathbb{C}^{q\times p}$, $C \in \mathbb{C}^{p\times q}$, $D \in \mathbb{C}^{p\times p}$ are continuous matrix valued functions, whereas $\mathbf{y}_1$, $\gamma_2 \in \mathbb{C}^p$, $\mathbf{y}_2$, $\gamma_1 \in \mathbb{C}^q$ and the matrices $K_{ij}$ have appropriate dimensions.

We next introduce the time-dependent change of variables $\mathbf{y} = Z(t)\,\mathbf{w}$, with

$$Z(t) = \begin{bmatrix} I_q & 0 \\ X(t) & I_p \end{bmatrix},$$

*This work has been partially supported by the Universidad Politcnica de Valencia through project 20070307.

†serblaza, eponsoda@imm.upv.es

and choose the matrix $X \in \mathbb{C}^{p\times q}$ so as to ensure that in the new variables $\mathbf{w} = Z^{-1}(t)\mathbf{y}$ the system assume the partly decoupled structure

$$\mathbf{w'} = \begin{bmatrix} \mathbf{w'}_1 \\ \mathbf{w'}_2 \end{bmatrix} = \begin{bmatrix} A+BX & B \\ 0 & D-XB \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}, \tag{3}$$

together with the corresponding boundary conditions for $\mathbf{w}$, see [3] for details. It turns out that this is possible if and only if $X(t)$ satisfies the so-called differential Riccati equation

$$X'(t) = C(t) + D(t)\,X(t) - X\,A(t) - X(t)\,B(t)\,X(t), \qquad X(0) = X_0 \tag{4}$$

where the unknown $X(t)$ lies in $\mathbb{C}^{p\times q}$ is a continuous matrix valued function. By requiring

$$X_0 = -K_{12}^{-1}\,K_{11}\,, \tag{5}$$

the boundary conditions (2) also decouple as

$$[\,0 \;\; K_{12}\,]\,\mathbf{w}(0) = \gamma_1\,, \qquad [\,K_{21}+K_{22}X(T) \;\; K_{22}]\,\mathbf{w}(T) = \gamma_2\,. \tag{6}$$

Here we assume that $K_{12}$ is invertible. In this way, the original boundary value problem can be solved as follows (see [3]):

**(I)** solve equation (4) with initial condition (5) from $t = 0$ to $t = T$.

**(II)** Solve the $\mathbf{w}_2$-equation in (3) and (6), also from zero to $T$;

$$\mathbf{w'}_2 = (D(t) - X(t)B(t))\mathbf{w}_2, \qquad \mathbf{w}_2(0) = K_{12}^{-1}\gamma_1. \tag{7}$$

**(III)** Solve the $\mathbf{w}_1$-equation in (3) from $t = T$ to $t = 0$.

$$\mathbf{w'}_1 = (A(t) + B(t)X(t))\mathbf{w}_1 + B(t)\mathbf{w}_2(t), \tag{8}$$

with initial conditions $\mathbf{w}_1(T) = (\gamma_2 - K_{22}\mathbf{w}_2(T))\,(K_{21} + K_{22}X(T))^{-1}$.

**(IV)** To recover $\mathbf{y}(t) = Z(t)\,\mathbf{w}(t)$.

In other words, the solution of the original two-point boundary value problem can be obtained by solving a sequence of three different initial value problems, one of which involves the nonlinear equation (4).

In most applications, the solution of the Riccati differential equations appear as coefficient matrix functions inside other differential equations. These

other equations have also to be solved in order to find the solution to the full problem. In this sense, it is important to develop efficient numerical integrators for the Riccati equation, but in a more wide sense since the performance will depend on the numerical solution of the full system. For this reason, depending on the problem, a numerical integrator for the Riccati differential equation has to take into account the following aspects:

- Accuracy of the solution in the sense that the solution $X(t)$ is part of another differential equation and has to be computed with sufficient accuracy.
- Preservation of the qualitative properties of the numerical solution for $X(t)$ to preserve the structure of the remaining equations to be solved.
- To have good accuracy and stability to allow to use relatively large time-steps for solving e.g. stiff or oscillatory problems.
- Storage requirements is also an important question. In fact, frequently the numerical solution $X(t)$ as well as the matrices $A$, $B$, $C$, $D$ have to computed and stored at some instants to be used later for solving the remaining equations of the problem.

For these reasons, the most appropriate method for solving a given Riccati differential equation will depend on the particular problem in which it appears. In this work we study the Riccati equation appearing in boundary value problems.

## 2 The numerical integration

To design an efficient numerical integrators requires, however, to take into account some important computational aspects.

- First, to solve equation (4) we have to compute $A(t), B(t), C(t), D(t)$ at a given mesh, say $t = \tau_1, \tau_2, \ldots, \tau_m$. Some of the computed matrices $A(\tau_i)$, $B(\tau_i)$, $D(\tau_i)$ can also be used in the numerical solutions of equations (7) and (8), so these vales have then to be stored.
- As a result, we obtain numerical approximations to $X(t)$ at the end of each time step, $X_0, X_1, X_2, \ldots, X_N$. If we use a constant time step, $h$, then $X_n \simeq X(t_n)$ with $t_n = t_0 + nh$ and $T = Nh$. These values

have to be computed and stored to be used in equations (7) and (8). Note that if a method of order $p$ in the time step is used, then $X_1 = X(h) + \mathcal{O}(h^{p+1})$ but $X_N = X(T) + \mathcal{O}(h^p)$, so the accuracy deteriorates and the error propagation is very important to be considered when choosing an appropriate numerical integrator.

- We have to design efficient algorithms to numerically solve equation (7) to obtain $\mathbf{w}_{2,0}, \mathbf{w}_{2,1}, \ldots, \mathbf{w}_{2,M}$ using a time step $H \geq h$ such that $T = MH$, where $\mathbf{w}_{2,i} \simeq \mathbf{w}_2(t_0 + iH)$.

- Next, we have to solve the linear non-homogeneous equation (8) making use of the stored values $A(t)$, $B(t)$, $C(t)$, $D(t)$ at the mesh points $t = \tau_1, \tau_2, \ldots, \tau_m$ in tandem with the approximate values $X_1, X_2, \ldots, X_N$ at the mesh points $t_n = t_0 + nh$ and $\mathbf{w}_{2,0}, \mathbf{w}_{2,1}, \ldots, \mathbf{w}_{2,M}$ at the mesh points $T_n = t_0 + nH$. This procedure gives us $\mathbf{w}_{1,0}, \mathbf{w}_{1,1}, \ldots, \mathbf{w}_{2,L}$ using a time step $K \geq h, H$ such that $T = KL$, where $\mathbf{w}_{1,j} \simeq \mathbf{w}_1(t_0 + jK)$.

- Finally, we have to obtain approximations to the solution $\mathbf{y}(t) = Z(t)\,\mathbf{w}(t)$ from $X_i, \mathbf{w}_{1,j}, \mathbf{w}_{2,k}$ obtained at different mesh points.

Lie group integrators, like methods based on Magnus [7] or Fer [4] expansions for example, have shown during the last decade to be highly efficient for the numerical integration of linear differential equations since they give qualitatively correct and qualitatively accurate results. Then it seems that steps (I) and (III) can be solved using Lie group integrators. It could be perhaps more surprising that these algorithms can indeed be used to integrate the Riccati equation in step (II). The idea is then to carry the numerical solution of all equations using Lie group integrators in the most appropriate way in order to design efficient numerical algorithms.

It seems clear that the bottle neck both in the computational cost and the accuracy achieved remains in the numerical solution of the Riccati equation.

In order to apply Lie group methods to solve numerically the Riccati equation, we first apply the transformation

$$X(t) \;=\; V(t)\,W^{-1}(t),$$

with $V \in \mathbb{C}^{p\times q}$, $W \in \mathbb{C}^{q\times q}$ and $V(0) = X_0, W(0) = I_q$, in the region where $W(t)$ is invertible. Then equation (4) is equivalent to the linear system

$$Y' \;=\; S(t)\,Y(t)\,, \qquad Y(0) \;=\; \left[\begin{array}{c} I_q \\ X_0 \end{array}\right], \tag{9}$$

with

$$Y(t) = \begin{bmatrix} W(t) \\ V(t) \end{bmatrix}, \qquad S(t) = \begin{bmatrix} A(t) & B(t) \\ C(t) & D(t) \end{bmatrix},$$

see [6] for details. If we denote $Z(t) = \Phi(t,0)Z_0$ with $\Phi$ the fundamental solution given by

$$\Phi(t,0) = \begin{bmatrix} \alpha(t) & \beta(t) \\ \gamma(t) & \delta(t) \end{bmatrix},$$

then $W(t) = \alpha(t) + \beta(t)X_0,\ V(t) = \gamma(t) + \delta(t)X_0$ and then

$$X(t) = (\gamma(t) + \delta(t)X_0)\,(\alpha(t) + \beta(t)X_0)^{-1}.$$

Lie group integrators for linear problems can be applied here, see [1], [2] or [5]. Apparently, this system is similar to (1), but now we are dealing with an initial value problem and $Y$ is a matrix instead of a vector.

## 2.1 An explicit symmetric second order Lie group integrator

We present a second order Lie group integrator which should suffice for most problems when only relatively low accuracy is needed and large time steps are desired due to storage requirements.

This method will be the starting point to build higher order methods and will help us to better understand the main difficulties for this purpose.

It is well known that equation (9) can be solved by

$$\Psi(t_n + h, t_n) \equiv \exp\left(\int_{t_n}^{t_n+h} S(t)\,dt\right) = \Phi(t_n + h, t_n) + \mathcal{O}(h^3) \tag{10}$$

so

$$\begin{bmatrix} W_{n+1} \\ V_{n+1} \end{bmatrix} = \Psi(t_n + h, t_n) \begin{bmatrix} W_n \\ V_n \end{bmatrix} \quad \Rightarrow \quad X_{n+1} = V_{n+1}\,W_{n+1}^{-1}$$

which is an approximation to $X(t_n + h)$. We can approximate the integral with a second order quadrature rule like the midpoint or trapezoidal rule. However, for this problem it seems more appropriate to consider the trapezoidal rule

$$\int_{t_n}^{t_n+h} S(t)\,dt = \frac{h}{2}(S(t_n + h) + S(t_n)) + \mathcal{O}(h^3).$$

In this way, the same evaluations of the matrix functions $A(t_n), B(t_n), D(t_n)$ and the computed $X_n = X(t_n) + \mathcal{O}(h^3)$ can be reused for the numerical solution of (7) and (8).

For the numerical solution of (7) we can use the same method

$$\mathbf{w}_{2,n+1} = e^{\Theta_n}\, \mathbf{w}_{2,n}$$

where

$$\Theta_n \equiv \frac{h}{2}(D_{n+1}+D_n-X_{n+1}B_{n+1}-X_nB_n) = \int_{t_n}^{t_n+h} (D(t)-X(t)B(t))\, dt+\mathcal{O}(h^3).$$

The numerical solution of (8) requires some caution. If $\Phi(t,0)$ denotes the fundamental matrix solution of the homogeneous equation $\mathbf{w'}_1 = (A(t)+B(t)X(t))\mathbf{w}_1$ then the solution of the non-homogeneous equation is

$$\mathbf{w}_1(t) = \Phi(t,0)\mathbf{w}_1 + \int_0^t \Phi(t,\tau)B(\tau)\mathbf{w}_2(\tau)\, d\tau.$$

Fortunately, this integral can be approximated by the trapezoidal rule making use of the already computed functions

$$\int_{t_n}^{t_{n+1}} \Phi(t_{n+1},\tau)B(\tau)\mathbf{w}_2(\tau)\, d\tau = \frac{h}{2}\left(B_{n+1}\mathbf{w}_{2,n+1} + \Phi(t_{n+1},t_n)B_n\mathbf{w}_{2,n}\right).$$

and then

$$\mathbf{w}_{1,n+1} = \Phi(t_n+h,t_n)\left(\mathbf{w}_{1,n} + \frac{h}{2}B_n\mathbf{w}_{2,n}\right) + \frac{h}{2}B_n\mathbf{w}_{2,n}$$

However, we must keep in mind that this equation is solved backward in time, so the recursive scheme is

$$\mathbf{w}_{1,n} = \Phi(t_{n+1}-h,t_{n+1})\left(\mathbf{w}_{1,n+1} - \frac{h}{2}B_{n+1}\mathbf{w}_{2,n+1}\right) - \frac{h}{2}B_n\mathbf{w}_{2,n}$$

This illustrates the time-symmetry of the method if we approximate $\Phi$ by an scheme that also satisfies $\Phi(t_n,t_{n+1}) = \Phi(t_{n+1},t_n)^{-1}$, as it is the case if we consider

$$\Phi(t_{n+1},t_n) = e^{\Lambda}$$

with

$$\Lambda = \frac{h}{2}\left(A_{n+1} + A_n + B_n X_n + B_{n+1} X_{n+1}\right)$$

Finally, the solution given by

$$\mathbf{y}(t) = (\mathbf{y}_1(t), \mathbf{y}_2(t)) = (\mathbf{w}_1(t), X(t)\mathbf{w}_1(t) + \mathbf{w}_2(t))$$

is approximated at the mesh points $t_i = t_0 + ih,\ i = 0, 1, \ldots, N$ by

$$\mathbf{y}_{1,n} = \mathbf{w}_{1,n}, \qquad \mathbf{y}_{2,n} = X_n \mathbf{w}_{1,n} + \mathbf{w}_{2,n}, \quad n = 0, 1, \ldots, N. \tag{11}$$

We present an algorithm for the numerical integration of the linear boundary value problem (1) and (2), using $N$ steps of length $h = T/N$ (we consider $t_0 = 0$) with the time-symmetric Lie group integrator presented in (10) to (11):

**Algorithm**

$h = T/N$
**do** $i = 0, N$
$A_i = A(ih),\ B_i = B(ih),\ D_i = D(ih),$
**enddo**
$X_0 = -K_{12}^{-1} K_{11}$
$W = I,\ V = X_0$
**do** $i = 1, N$
$C_i = C(ih)$
$(W, V)^T = \exp\left(\frac{h}{2}\left(A_i + A_{i-1}, B_i + B_{i-1}; C_i + C_{i-1}, D_i + D_{i-1}\right)\right)(W, V)^T$
$X_i = VW^{-1}$
**enddo**
$\mathbf{w}_{2,0} = K_{12}^{-1}\gamma_1$
**do** $i = 1, N$
$\mathbf{w}_{2,i} = \exp\left(\frac{h}{2}\left(D_{i-1} + D_i - X_{i-1}B_{i-1} - X_i B_i\right)\right)\mathbf{w}_{2,i-1}$
**enddo**
$\mathbf{w}_{1,0} = (\gamma_2 - K_{22}\mathbf{w}_{2,N})\,(K_{21} + K_{22}X_N)^{-1}$
**do** $i = N - 1, 0$
$\mathbf{w}_{1,i}$
$= \exp\left(-\frac{h}{2}\left(A_{i+1} + A_i - B_{i+1}X_{i+1} + B_i X_i\right)\right)\left(\mathbf{w}_{1,i+1} - \frac{h}{2}B_{i+1}X_{i+1}\right) - \frac{h}{2}B_i X_i$
**enddo**
**do** $i = 0, N$
$\mathbf{y}_i = (\mathbf{w}_{1,i},\ X_i\mathbf{w}_{1,i} + \mathbf{w}_{2,i})$
**enddo**

The performance of this algorithms depends on the rate between the computational cost and the accurate. In this sense, note that about the computational cost we need

- to compute and store $A$, $B$, $C$, $D$, $X$, $\mathbf{w}_1, \mathbf{w}_2$ at $N+1$ points.
- To compute $N+2$ inverse matrices.
- To compute $N$ exponentials of matrices of dimension $p \times p$, $q \times q$ and $(p+q) \times (p+q)$, respectively (or their actions on vectors or other matrices).

and about the accuracy

- we have that $X_1 = X(h) + \mathcal{O}(h^3)$ but, since $T = Nh$ then $X_N = X(NH) + \mathcal{O}(h^2)$, and similarly for $\mathbf{w}_{2,i}$.
- For the equation of $\mathbf{w}_1$, we have as initial conditions $\mathbf{w}_{1,N} = \mathbf{w}_1(NH) + \mathcal{O}(h^2)$ to start the backward numerical integration, which can lead to an important error accumulation. Since time-symmetry is preserved, we expect this trouble to be diminished.
- Exponential methods usually lead to considerably more accurate results than polynomial or rational approximations (like Runge-Kutta methods). They can efficiently deal with moderately stiff problems (while being explicit) as well as with oscillatory problems.

# References

[1] Blanes, S., Casas, F., Ros, J., High order optimized geometric integrators for linear differential equations, *BIT* **42** , pp. 262-284. (2002)

[2] Blanes, S., Jódar, L., Ponsoda, E., Approximate solutions with a priori error bounds for continuous coefficient matrix Riccati equations, Math. Comp. Modelling, **31**, pp. 1-15. (2000)

[3] Dieci, L., Numerical integration of the differential Riccati equation and some related issues. *SIAM J. Numer. Anal.*, **29**, 781-815, (1992).

[4] Fer, F., Résolution de l'équation matricielle $U' = pU$ par produit infini d'exponentielles matricielles, *Bull. Classe Sci. Acad. Roy. Bel.*, **44**, 818-829. (1958)

[5] Iserles, A., Munthe-Kaas, H. Z., Nørsett, S. P., Zanna, A., Lie group methods, *Acta Numerica* **9**, pp. 215-365. (2000)

[6] Jódar, L., and Ponsoda, E., Non-autonomous Riccati-type matrix differential equations: existence interval, construction of continuous numerical solutions and error bounds. *IMA J. Num. Anal.*, **15**, 61-74. (1995).

[7] Magnus, W., On the exponential solution of differential equations for a linear operator, *Commun. Pure Appl. Math.* **7** , pp. 649-673. (1954)

# A compartmental model for nitrogen partitioning in evergreen trees *

Rafael Cantó†, Beatriz Ricarte‡, Ana M. Urbano§

Instituto de Matemática Multidisciplinar,

Universidad Politécnica de Valencia,

Camino de Vera s/n, 46022 Valencia, España

December 11, 2008

## 1 Introduction

It is important to study how to improve the efficiency of the nitrogen fertilization because it is a cultural practice that directly influences on the profitability of the cultures [1], in addition to be the main source of contamination of underground waters in farms, mostly of Citrus trees [2].

The objectives of the different works carried out up to now have been centered in physiological aspects of the plant. But, all this information can be used to approach the problem from a mathematical point of view [3, 4, 5]. Hence, a mathematical expression that simulates the dynamics of nitrogen in Citrus trees may predict the behavior of nitrogen after being absorbed from the irrigation water, which lets a better adjustment of the present programs of fertilization.

In principle, the compartmental model we describe in this work could be used to model dynamics of any nutrient, or at least with similar mobility as

*Supported by the Spanish DGI grant DGI MTM2007-64477 and the UPV under its research program.

†e-mail: rcanto@mat.upv.es

‡e-mail: bearibe@mat.upv.es

§e-mail: amurbano@mat.upv.es

nitrogen, in evergreen tress. Nevertheless, parameter estimation and later model validation have been done in citrus trees.

# 2 Mathematical modelling

## 2.1 Experimental data

The experimental data came from seasonal absorption studies made on five-year-old Valencia Late orange trees [*Citrus sinensis* (L.) Osbeck] grafted on citrange Troyer and grown in sand-filled containers by hydroponic culture. These trees were marked applying nutritious solution enriched with $^{15}$N at different moments of the annual cycle of development. Immediately after every marked period, same trees were pull out and separated into their different organs. At each organs dry matter, mass uptake of labelled $^{15}$N and the total nitrogen content of each tissue was measured.

## 2.2 Compartmental model

The compartment distribution follows physical and physiological criterion leading to the compartmental system formed by six compartments shown in Figure 1: roots (denoted by R), old branches and trunk (OBT), old leaves (OL), young branches (YB) and young leaves (YL) from the different flushes in that year, respectively, and flower organs and fruits (FF). Arrows, denotes by $D$, flows of nitrogen between compartments, except for the first one, which represent the amount of nitrogen daily applied. This general diagram of Figure 1 changes through the annual growth cycle. At the beginning of the year, the system consists on the three compartments corresponded to old organs. In spring, first growth flush and flowering happen appearing the next three compartments, completing all the diagram. After, in summer and autumn, two more flushes may take place but these new organs also belong to compartments 4 and 5, i.e., young branches and leaves. Next year, the diagram will be the same, since organs from the previous flushes are now considered as old organs. We begin with a three compartment system but the following compartments will appear with the next flushes, and so on, year after year. That is, it is a periodic compartmental system of period 365 days.

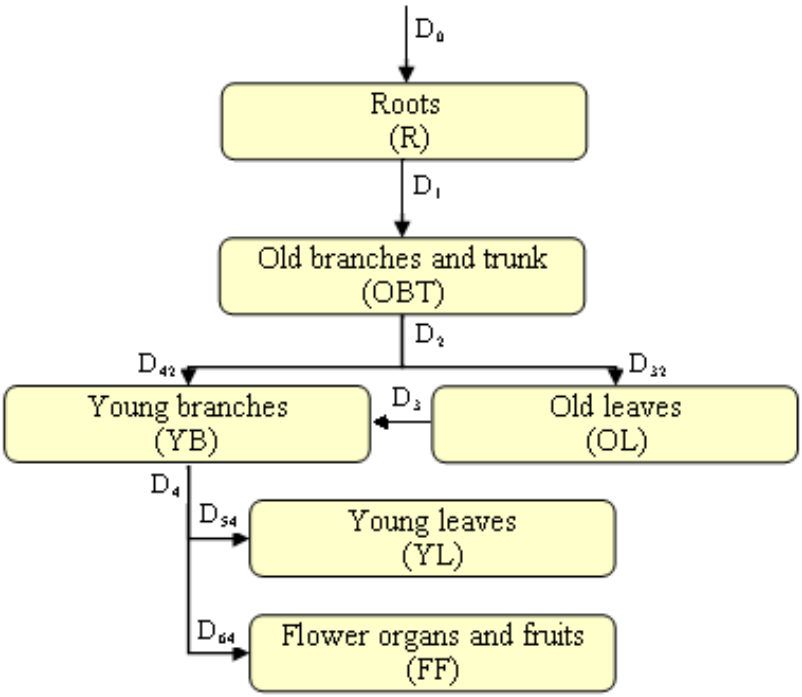



Figure 1: Structure of the partitioning model showing the source-sink relationships

## 2.3 Model equation

The model considers each tissue type, that is, each organ, as source and sink for nitrogen and uses flow equations to simulate the dynamics of nitrogen partitioning during a year. At each instant $t$, we consider that the net outflow of nitrogen in milligrams per day (mg day$^{-1}$) from a source tissue to sink organs depends on the concentration in that compartment and it is proportional to the sum of the rate of dry matter increment in new growing sink compartments (young branches, young leaves, flower organs and fruits) and the total absorption in old sink compartments (old branches and trunk, old leaves), i.e.,

$$D_i(t) = K_i(t)\frac{Ni(t)}{P_i(t)}\left[\sum_{j>i}(Abs_j(t)) + \sum_{k>i}\left(\omega_k\frac{dP_k(t)}{dt}\right)\right] = D_{ji}(t) + D_{ki}(t) \tag{1}$$

for $i = 1, 2, 3, 4$, $j = 2, 3$, $k = 4, 5, 6$, where $K_i(t)$ is a dimensionless tissue specific transport coefficient, $N_i(t)$ the nitrogen content in the organ $i$ measured in milligrams (mg), $P_i(t)$ its dry matter measured in grams (g), $Abs_j(t)$ the nitrogen absorption evolution (in g day$^{-1}$) of the organ $j$ and $\omega_k$ is a dimensionless coefficient that reflects the relationship between the increase in dry matter of the corresponding organ and its paper as a sink. The variable nitrogen content of each organ reflects its source capacity.

The change in nitrogen content is calculated using a mass balance equation taking into account the compartment distribution in Figure 1 and that $D_0$ is the total nitrogen taken up by the roots, i.e.,

$$\frac{dNi(t)}{dt} = \sum_{i>j}(\alpha_{ij}D_{ij}(t)) + \sum_{i>k}(\alpha_{ik}D_{ik}(t)) - D_i(t) \tag{2}$$

with the operators $\alpha$ equal to 1 when the corresponding organs are connected and equal to 0 in the opposite case. Equations (1) and (2) lead to a differential system that can be written in matricial form as

$$\dot{N}(t) = A(t)N(t) + B(t)u(t) \tag{3}$$

where the state matrix $A(t)$ is a time variant matrix given by

$$A(t) = \begin{bmatrix} -D_1(t) & 0 & 0 & 0 & 0 & 0 \\ D_1(t) & -D_2(t) & 0 & 0 & 0 & 0 \\ 0 & D_{32}(t) & -D_3(t) & 0 & 0 & 0 \\ 0 & D_{42}(t) & D_3(t) & -D_4(t) & 0 & 0 \\ 0 & 0 & 0 & D_{54}(t) & 0 & 0 \\ 0 & 0 & 0 & D_{64}(t) & 0 & 0 \end{bmatrix}$$

whereas the control matrix $B(t)$ is the time invariant matrix $B = [1\ 0\ \ldots\ 0]^T$ because fertilizer is always absorbed by the roots, that is, the first compartment, and $u(t) = D_0(t)$.

# 3 Results and discussion

The goodness of fits were assessed using the coefficient of determination $R^2$ between the experimental and predicted data. As a summary, Table 1 shows the obtained results, where $R^2$ denotes the coefficient of determination for 5 years old trees (used for parameter model estimation), whereas $R^2_{val}$ denotes the coefficient of determination for 2 years old trees (used for validation). Except for roots, note that the model reflects the real situation. But we

| | R | OBT | OL | YB | YL | FF |
|---|---|---|---|---|---|---|
| $R^2$ | 0.5564 | 0.9571 | 0.9160 | 0.9456 | 0.9799 | 0.9694 |
| $R^2_{val}$ | 0.4450 | 0.9833 | 0.3993 | 0.9573 | 0.9936 | 0.9763 |

Table 1: Comparison between experimental data and the values estimated by the model

are not worried about that because we have in mind to improve in the near future the usefulness of the model adding the soil compartment, which is sure to influence on the root compartment. The worse result obtain in old leaves may be due to this compartment behaves different according to the age.

The aim of the model is to obtain information about the nitrogen dynamics in the tree, which can be achieved by simulation. That is, to change the control we apply to the model and check the variable behavior.

Let us see an example. Consider the five-year-old Valencia Late orange trees. In Figure 2 we can see what would happen if they absorbed the

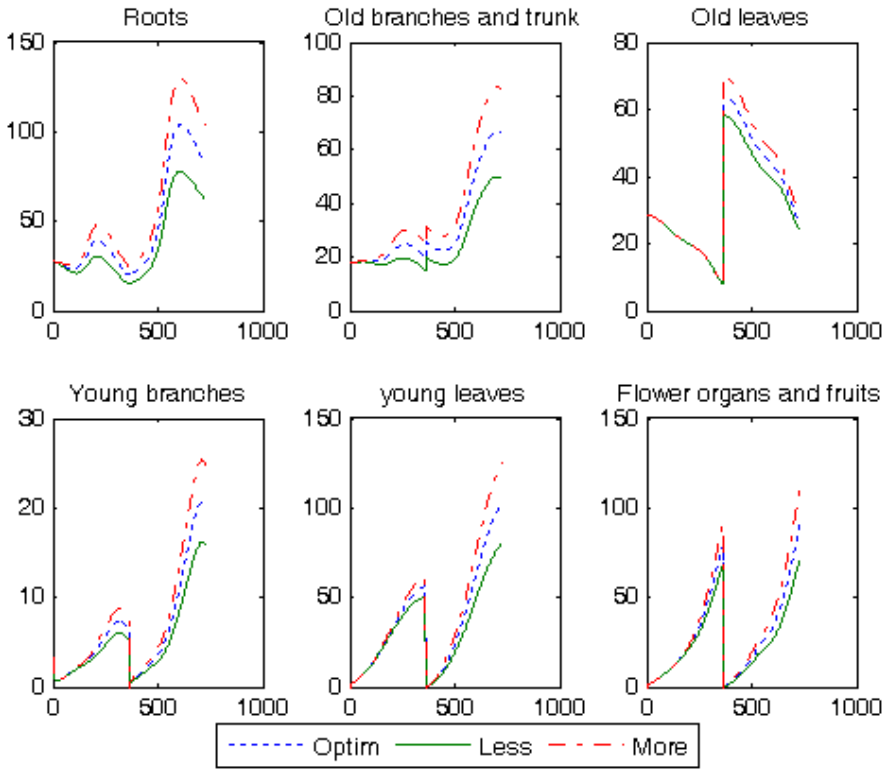



Figure 2: Nitrogen content (g) in the different organs estimated from the model (-) compared with experimental data

optimum amount of nitrogen ($D_0$), 75% or 1.25% of the optimum rate in each time during two years.

# 4 Conclusion

Nutrient dynamics in trees can be represented by a mathematical model. In particular, the dynamics of nitrogen absorption, distribution and translocation in citrus trees can be modelled by a continuous-time periodic compartmental system of one year periodicity. This mathematical model may clarify the behavior of nitrogen when it is absorbed by the tree so it could be used to improve the fertilizer criteria applied nowadays.

# References

[1] H. D. Chapman, The mineral nutrition of citrus, In *The citrus industry*, (Edited by W. Reuther, L. D. Barchelor and M. J. Webber), pp. 127-289. Univ. Calif. Div. Agr. Sci., Bekeley, California, (1968).

[2] F. T. Bingham, S. David and E. Shade, Water relations, salt balance and nitrate leaching losses of a 960 acre citrus watershed. *Soil Sci.* **112** (6) 410-418 (1971).

[3] R. Bru, R. Cantó and B. Ricarte, Modelling nitrogen dynamics in citrus tree. *Mathematical and Computer Modelling* **38** 975-987 (2003).

[4] R. Habib, P. Millard and M. F. Proe, Modelling the Seasonal Nitrogen Partitioning in Young Sycamore (Acer pseudoplatanus) Trees in Relation to Nitrogen Supply. *Annals of Botany* **71** 453-459 (1993).

[5] R. Habib and P. Monestiez, Modelling of dynamics of nitrogen partitioning in a young fruit tree during the exponential growth stage I: Modelling and fitting of parameters. *Agronomie* **7** (6) 401-408 (1987).

# A modified CE-SE method for solving advection-diffusion problems*

R. Company, E. Ponsoda, J.-V. Romero and M.-D. Roselló †

Instituto de Matemática Multidisciplinar,

Universidad Politécnica de Valencia,

Camino de Vera s/n, Edificio 8G, $2^o$,

46022 Valencia,España

December 11, 2008

## 1 Introduction

Let us consider the advection-diffusion equation with constant coefficients

$$\left.\begin{array}{c} \dfrac{\partial}{\partial t}u(x,t) \,+\, a\,\dfrac{\partial}{\partial x}u(x,t) \,-\, \mu\,\dfrac{\partial^2}{\partial x^2}u(x,t) \,=\, 0\,, \\ (x,t) \in \mathbb{R} \times [0\,,+\infty[\;; \quad \mu \geq 0\,, \end{array}\right\} \tag{1}$$

under the initial condition

$$u(x,0) \,=\, f(x)\,. \tag{2}$$

The advection-diffusion equation appears frequently in many physical and technological models, for example, in the evaluation of the heating trough radiations of microwaves [1], in the study of the transmission of flows in

*This work has been partially supported by the Generalitat Valenciana grants GV/2007/009 and GVPRE/2008/092, the Spanish M.C.Y.T. and FEDER grant TRA2007–68006–C02–02, and Universidad Politécnica de Valencia 2007-09, grant 4611.

†e-mails: rcompany, eponsoda, jvromero, drosello @imm.upv.es

industrial tubes [2] and in problems of heat transmission in solids, see [3] or [4] and the references therein.

In [5], the space-time conservation element and solution element (CE-SE) method is applied to problem (1)–(2), and the resultant $a - \mu$ scheme presents many advantages such as behavior and stability if it is compared with other numerical methods, like finite differences or finite elements, for example. Furthermore, the integral form of the method exploits the physical properties of conservation of flow, unlike the differential form. Also, this explicit scheme evaluates the variable and its derivative simultaneously in each knot of the partitioned domain, see [6].

In the standard CE-SE numerical scheme, the space-time domain is partitioned in a grid such that the knots $(j, n)$ are obtained for $n = 0,\ 1/2,\ 1,\ 3/2, \ldots$ and, for each $n$, the spatial knot is given by $j = n, n\pm1,\ n\pm2,\ n\pm3, \ldots$ see [6] for details. Then, the solution element $SE(j, n)$ is defined as the space-time region enclosed inside the rhombus centered in $(x_j, t^n)$ and whose diagonals are $\Delta t$ and $\Delta x$, see [4]. Using the standard CE-SE method, in each solution element, we define $U(x, t; j, n)$ as the first order Taylor approximation of $U(x, t)$, centered at $(x_j, t^n)$, with coefficients $U_j^n$, $(U_x)_j^n$, $(U_t)_j^n$. Coefficients $U_j^n$, $(U_x)_j^n$, $(U_t)_j^n$ are constants to be determined into $SE(j, n)$, see [4]. In this way, the conservation elements, $CE_{\pm}(j, n)$, are defined as in [6] to evaluate these constants.

In this paper, a new numerical method for solving the advection-diffusion equation, based in the CE-SE method is developed in section 2. This method increases the precision using a second order Taylor expansion in the spatial variable. A better approximate numerical solution of system (1)–(2) is obtained and it is validated with an analytical solution.

## 2 The numerical scheme

In a similar way that in the $a - \mu$ scheme, [5], the space-time domain is partitioned in a grid such that the knots $(j, n)$ are obtained for $n = 0,\ 1/2,\ 1,\ 3/2, \ldots$ and, for each $n$, we take $j = n, n \pm 1,\ n \pm 2,\ n \pm 3, \ldots$. Then, we define the solution element $SE(j, n)$ as the space-time region enclosed inside the rhombus centered in $(x_j, t^n)$ and whose diagonals are $\Delta t$ and $\Delta x$, see [4] for details.

As in the usual first order CE-SE numerical scheme, in each solution

element we define

$$\left.\begin{array}{c} U\left(x,t;j,n\right)=U_j^n+\left(U_x\right)_j^n\left(x-x_j\right)+\left(U_t\right)_j^n\left(t-t^n\right)+\dfrac{1}{2}\left(U_{xx}\right)_j^n\left(x-x_j\right)^2\ , \\ \\ \forall(x,t)\in SE(j,n)\ , \end{array}\right\} \tag{3}$$

where $U_j^n$, $\left(U_x\right)_j^n$, $\left(U_t\right)_j^n$ and $\left(U_{xx}\right)_j{}^n$ are constants to be determined into each $SE(j,n)$. Now, note that $U\left(x,t;j,n\right)$ has been developed in Taylor's series up to the second order in the spatial variable. We have increased the precision in this variable in view that in the equation (1), the second derivative with respect to $x$ appears.

To calculate these constants the so called elements of conservation $CE(j,n)$ are in use, which exploit the conservation of the flow in certain space-time regions that later we will describe. It is for it that turns out to be important the integral formulation of problem (1), whose proof can be found in [4].

In this way, in the solution elements we define

$$\left.\begin{array}{c} G\left(x,t;j,n\right)=\left(G_1(j,n),G_2(j,n)\right) \\ \\ =\left(-U(x,t;j,n)\,,\,aU(x,t;j,n)-\mu\left(U_x\right)_j^n-\mu\left(U_{xx}\right)_j^n\left(x-x_j\right)\right)\ , \\ \\ \forall(x,t)\in SE(j,n)\,. \end{array}\right\} \tag{4}$$

In the usual first order CE-SE method, only two constants are to be determined, then two conservation elements are necessary. In the new approximation using the second order Taylor expansion (3), four constants have to be determined, $U_j^n$, $\left(U_t\right)_j^n$, $\left(U_x\right)_j^n$ and $\left(U_{xx}\right)_j^n$, then we need to introduce two news conservation elements. In this sense, we define the conservation elements $CE_i(j,n)\,,\ 1\leq i\leq 4\,.$ $CE_1(j,n)$ and $CE_2(j,n)$ are the rectangular regions such that the knot $(x_j,t^n)$ is located in the right top corner, and whose sides have a length $\Delta x/2,\ \Delta t/2$ for $CE_1(j,n)$ and $\Delta x/4,\ \Delta t/2$ for $CE_2(j,n)$. The elements $CE_3(j,n)$ and $CE_4(j,n)$ are defined in the same way, but the knot $(x_j,t^n)$ is located now in the left top corner. Note that $CE_1(j,n)$ and $CE_4(j,n)$ are the same elements $CE_-(j,n)$ and $CE_+(j,n)$, respectively, from the usual first order CE-SE method.

In order to calculate the constants $U_j^n$, $\left(U_t\right)_j^n$, $\left(U_x\right)_j^n$ and $\left(U_{xx}\right)_j^n$ in (3), we use the following approximation of the integral equation

$$F_i(j,n)=\oint_{S(CE_i(j,n))}G\left(x,t;j,n\right)dr$$

$$= \oint_{S(CE_i(j,n))} G_1(j,n)\,dx + G_2(j,n)\,dt = 0\ ,$$

where $CE_i(j,n)$ is the conservation element and $G\,(x,t;j,n)$ is given by (4).

Note that for evaluate the integral in the first conservation element, from $(x_j, t^n)$ to $(x_{j-1/2}, t^n)$ the variable $t$ does not change, and the solution element is the $SE(j,n)$, but from $(x_{j-1/2}, t^n)$ to $(x_{j-1/2}, t^{n-1/2})$ the solution element is $SE(j-1/2, n-1/2)$ and now, the variable $x$ is constant. Thus, the first integral needs to compute four different integrals

$$F_1 = \int_{x_j}^{x_{j-1/2}} G_1(j,n)dx + \int_{t^n}^{t^{n-1/2}} G_2(j-1/2, n-1/2)dt$$

$$+ \int_{x_{j-1/2}}^{x_j} G_1(j-1/2, n-1/2)dx + \int_{t^{n-1/2}}^{t^n} G_2(j,n)dt\,,$$

where $G_i(j,n)$, $i = 1, 2$, are defined in (4). Similar arguments are necessary to compute the others integrals.

$$F_2 = \int_{x_j}^{x_{j-1/4}} G_1(j,n)dx + \int_{t^n}^{t^{n-1/4}} G_2(j,n)dt + \int_{t^{n-1/4}}^{t^{n-1/2}} G_2(j-1/2, n-1/2)dt$$

$$+ \int_{x_{j-1/4}}^{x_j} G_1(j-1/2, n-1/2)dx + \int_{t^{n-1/2}}^{t^n} G_2(j,n)dt\,,$$

$$F_3 = \int_{t^n}^{t^{n-1/2}} G_2(j,n)dt + \int_{x_j}^{x_{j+1/4}} G_1(j+1/2, n-1/2)dx$$

$$+ \int_{t^{n-1/2}}^{t^{n-1/4}} G_2(j+1/2, n-1/2)dt + \int_{t^{n-1/4}}^{t^n} G_2(j,n)dt + \int_{x_{j+1/4}}^{x_j} G_1(j,n)dx,$$

$$F_4 = \int_{t^n}^{t^{n-1/2}} G_2(j,n)dt + \int_{x_j}^{x_{j+1/2}} G_1(j+1/2, n-1/2)dx$$

$$+ \int_{t^{n-1/2}}^{t^n} G_2(j+1/2, n-1/2)dt + \int_{x_{j+1/2}}^{x_j} G_1(j,n)dx\,.$$

We define

$$q(j,n) = \begin{bmatrix} U_j^n \\ \frac{1}{2!}\frac{\Delta x}{2}\,(U_x)_j^n \\ \frac{1}{3!}\left(\frac{\Delta x}{2}\right)^2 (U_{xx})_j^n \\ \frac{1}{4!}\left(\frac{\Delta x}{2}\right)^3 (U_t)_j^n \end{bmatrix}, \tag{5}$$

and the initial condition

$$q(j,0) = \begin{bmatrix} u(x_j,0) \\ \frac{\Delta x}{4}\frac{\partial}{\partial x}u(x_j,0) \\ \frac{\Delta x}{24}\frac{\partial^2}{\partial x^2}u(x_j,0) \\ \frac{\Delta x}{192}\frac{\partial}{\partial t}u(x_j,0) \end{bmatrix}.$$

From equations $F_i(j,n) = 0$, $1 \le i \le 4$, and expressing them in the matrix form, we have

$$q(j,n+1) = Q_+^2\; q(j-1,n) + \left[Q_+Q_- + Q_-Q_+\right]\; q(j,n) + Q_-^2\; q(j+1,n),$$

where $q(j,n)$ is given by (5) and $Q_+$ and $Q_-$ are the $4 \times 4$ matrix

$$Q_+ = \frac{1}{2\nu^2 - 3\,(2 + 5\xi + 2\xi^2)} \begin{bmatrix} q_{+1} & q_{+2} & q_{+3} & q_{+4} \end{bmatrix}$$

and

$$Q_- = \frac{1}{2\nu^2 - 3\,(2 + 5\xi + 2\xi^2)} \begin{bmatrix} q_{-1} & q_{-2} & q_{-3} & q_{-4} \end{bmatrix},$$

where

$$\nu = a\frac{\Delta t}{\Delta x}, \qquad \xi = 4\mu\frac{\Delta t}{(\Delta x)^2},$$

and $q_{\pm i} \in \mathbb{R}^4$, $1 \le i \le 4$, are given in terms of $a$, $\mu$, $\nu$ and $\xi$.

# 3 Numerical example

In order to test the proposed scheme we consider the following problem which analytical solution is well known. We choose the one-dimensional advection-diffusion equation (1) of a Gaussian pulse of unit height, centered at $x_0 = 1$ in a region bounded by $0 \leq x \leq 9$. The exact solution of this problem is given by [7]

$$u_a(x,t) = \frac{1}{\sqrt{4t+1}} \exp\left(-\frac{(x-x_0-at)^2}{\mu(4t+1)}\right). \tag{6}$$

The initial condition is given by

$$u(x,0) = \exp\left(-\frac{(x-x_0)^2}{\mu}\right)$$

and the boundary condition at the two ends at any time $t$ is obtained by substituting $x = 0$ and $x = 9$, respectively, in (6). The values of the parameters used are $\nu = 0.005\,\mathrm{m}^2/\mathrm{s}$ and $a = 0.8\,\mathrm{m/s}$.

In this section we compare the numerical method presented in this article with the original CE-SE method. The implementation of both methods has been realized using the program *Mathematica* version 6.0 [8].

In table 1 we compare the errors of two methods (CE-SE standard and the developed second order CE-SE method) with $\nu = 0.8$ and $\Delta x = 0.4$.

| | CE-SE | Second order CE-SE |
|---|---|---|
| $\Delta x = 0.04$ | 0.0027 | 0.0014 |
| $\Delta x = 0.02$ | 0.0011 | 0.00033 |

*Table 1: Errors in $t = 5$, $x = 5$*

In this table we have presented the results at time $t = 5$, for a mesh size in the spatial variable $\Delta x = 0.4$ and $\Delta x = 0.2$. If we reduce the mesh size, we can observe that the errors in both methods they diminish, having the new second order CE-SE method a better behavior that the standard CE-SE method method. If we analyze different times, the behavior of both methods is similar to the presented one in table 1.

Summarizing, we can observe that the results obtained with the modified CE-SE method improves the one obtained with the usual CE-SE scheme.

# 4 Conclusions

With regard to the traditional $a-\mu$ method, in this work we have increased the Taylor's expansion up to the second order in the spatial variable in each solution element, in view to obtain a higher precision in the results. This motivates the search of new conservations elements because we need to evaluate more constants that in the usual first order expansion.

Taking into account that in each conservation element we have different conservation elements, we solve the four integrals and, in this way we determine the approximate solution of problem (1)–(2) in each knot of the partitioned domain.

In order to evaluate if the modified CE-SE method that we propose in this work improves the usual $a-\mu$ scheme, an illustrative example, which analytical solution is known, is proposed. Then we compare the results obtained with both methods and with the exact solution that is known. We conclude that, our method improves the obtained results.

# References

[1] A.C. Metaxas, R.J. Meredith, *Industrial Microwave Heating*. Peter Peregrinus, (1983).

[2] T.S. Chua, P.M. Dew, The design of a variable step integrator for the simulation of gas transmission networks. *Int. J. Numer. Meths. in Engieneering*, Vol. 20, 1797–1813, (1984).

[3] H.S. Carslaw, J.C. Jaeger, *Conduction of Heat in Solids*. Oxford Univ. Press, Oxford, (1995).

[4] E. Ponsoda, E. Defez, M.-D. Roselló, J.-V. Romero, A stable numerical method for solving for solving variable coefficient advection-diffusion models, *Comp. Math. Applic.* **56**, 754–768, (2008).

[5] X.Y. Wang, C.Y. Chow, S.C. Chang, *Application of the Space-Time Conservation Element and Solution Element Method to One-Dimensional Advection-Difussion Problems*, NASA-TM-1999-209068, (1999).

[6] S.C. Chang, The method of space-time conservation element and solution element. A new approach for solving the Navier-Stokes and Euler equations. *J. Comput. Phys.* **119**, 295–324, (1995).

[7] H. Karahan, A third-order upwind scheme for the advection-diffusion equation using spreadsheets, *Advances in Engineering Software*, **38**, 688–697, (2007).

[8] S. Wolfram, *The Mathematica Book*, Wolfram Media Inc. 1999.

# Modeling the Evolution of Bladder Carcinoma: a Markovian Approach

Cristina Santamaría[*], Belén García–Mora[†], Gregorio Rubio[‡], Enrique Navarro[§]

Instituto de Matemtica Multidisciplinar,

Universidad Politécnica de Valencia,

Camino de Vera s/n,

46022 Valencia, España.

December 11, 2008

## 1 Introduction

Bladder carcinoma is a highly aggressive neoplasm and the second most common malignancy encountered by urologists. It is the fourth most frequent solid tumor among men and the seventh most frequent among women, with more than 350.000 new cases diagnosed annually worldwide [1]. Fortunately, approximately 80% of patients with newly diagnosed bladder carcinoma present *superficial* transitional cell carcinoma (TCC) which can be managed with transurethral resection ($TUR$) [2], which is a surgical endoscopic technique used to remove the macroscopic tumor from the inner of the bladder. However, more than 50% of the patients will have *recurrences* (reappearance of a new superficial tumor) and the 10-30% of patients will have *progression* to muscle invasive disease [3] which leads to a more aggressive treatment including bladder extirpation.

[*]e-mail: crisanna@imm.upv.es
[†]e-mail: magarmo5@imm.upv.es
[‡]e-mail: grubio@imm.upv.es
[§]e-mail: entorres@imm.upv.es

It is necessary to determine the *risk factors* associated with both *recurrence* and *progression* in order to analyze the trajectory of bladder carcinoma and to establish a model that allows the prediction of the disease process after the *TUR*, because this will provide information about the recurrence–progression process and so the physician will have a more rigorous program about the follow-up of the patient.

In this regard, in survival analysis the Cox proportional model [4] has been mostly used to identify the *prognostic factors* associated to the *time to the first recurrence* of *superficial* TCC in the bladder cancer after *TUR* and an initial treatment [5, 6, 7, 8]. However, there are few studies about the analysis of the *multiple recurrences* (the main characteristic of *superficial* TCC of the bladder) and its associated factors as well as the *progression*. In this sense, we analyzed exhaustively the associated factors to multiple recurrences [9] from extensions of the Cox model ([10, 11, 12]). In order to improve the model we included the progression in our analysis and we distinguished the multiple recurrences as *recurring events* and the progression as a *terminating event* [13] similar to the death. This let us to carry out several inferences in the recurrence–progression process and to determine and to differentiate the *risk factors* associated to multiple recurrences and the *risk factors* associated to progression.

On the other hand, one more step in these studies is the introduction of stochastic processes which allows to apply a dynamic survival approach in the analysis of survival data. In the biostatistical literature Markov models have proven to be a useful model in the study of the evolution of diseases [14, 15, 16]. So in order to improve the predictive model and to specify accurately the prognostic factors associated to both recurrence and progression in superficial vesical carcinoma we focus on the use of a homogeneous time–continuous Markov model. The methodology is complemented with the utilization of phase–type distributions [17]. This type of distribution has shown its utility in queue theory and has the advantage of an algorithmic treatment. The calculation of the mathematical expressions can be presented in a closed form that allows algebraic treatment and the corresponding computational implementation in an appropriate way.

The paper is organized as follows: in section 2 we present the bladder cancer data with the description of transient and absorbent states. In section 3 we describe the homogeneous time–continuous Markov model according to our database, taking into account the phase–type distribution approach. Finally, in section 4 we present the results of our analysis.

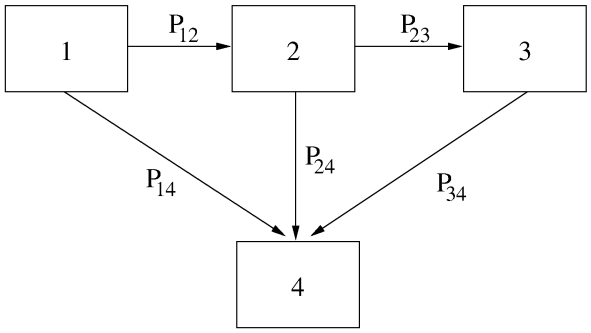



Figure 1: State space and transition probabilities in The Recurrence–Progression Process.

# 2 The Data

The analysis has been carried out from data gathered from the Department of Urology at *La Fe University Hospital* in Valencia (Spain). This database contains detailed information on *superficial* TCC of the bladder and was collected from 380 post-operative patients, between January 1995 and January 2006. All patients presented a primary *superficial* TCC which was removed by means of the *TUR*.

We will distinguish between the *pathological* characteristic (*grade* of the tumor) and *clinical* characteristics (*number* of tumors and *size* of the tumoral mass). *Grade* is categorized from G1 to G3 (from low aggressive to highly aggressive) according to the WHO (World Health Organization) [18]. In our analysis only levels G1 and G2 of *grade* were observed for *superficial* TCC. *Sex* and *age* were collected at the moment of the *TUR*. *Number* was classified in two levels: one and two or more tumors. *Size* has also two levels: minor or equal to 3 cm and more than 3 cm.

Our reference individual is a man with a mean age of 64.63 $\pm$ 12 years, with *clinico–pathological* factors grade G1 and, with only one tumor of size minor or equal to 3 cm. The patient population and an exhaustive descriptive analysis of these *clinico–pathological* factors in bladder tumors is carried out in [9].

In the follow-up period, we take into account up to two recurrences and the possibility of undergoing the terminating progression. So, we distinguish four posible states in patients, three *transient states* and one *absorbent state*.

State *1* (free of disease) is the initial state for all patients after *TUR*, in state *2* the patient suffers a first recurrence as the initial tumor, in state *3* the patient suffers a second recurrence, and the state *4* is the state of progression. We consider the three first states as recurring events and the progression as a terminating event because this includes bladder extirpation and a more aggressive treatment. This last state is in this sense similar to the death. We can observe these transitions and transition probabilities in the Figure 1. In order to prediction purposes we consider the *clinico–pathological* characteristics of the initial tumor for our analysis.

# 3 The model

Let $\{X(t),\, t \geq 0\}$ be a Markov process with a state space $S$, where $X(t),\, t \geq 0$ is the state of the disease process at time $t$. Then, for this model the transition probabilities $p_{ij}(t)$ for $t \geq 0$ are defined

$$p_{ij}(t) = P\{X(t+s) = j | X(s) = i\} = P\{X(t) = j | X(0) = i\}, \qquad (1)$$

where $p_{ij}(t)$ denotes the probability of going from state $i$ to state $j$ in a period of time $t$. As we can see this probability does not depend on $s$. This means that transition $i \to j$ on an interval of length $t$ has the same probability at any time. The transition matrix function is $P(t) = (p_{ij}(t))$. In this model, this matrix is upper triangular, because the only transitions are $i \to j$ for $i \leq j$ in $S$. The transition intensity matrix is $Q = (q_{ij})$, where $q_{ij}$ is the derivative with respect to $t$ of the transition probability function $p_{ij}(t)$ in $t = 0$. This matrix will be also upper triangular. The entry $q_{ij}$ represents the transition intensity of the Markov process between the states $i$, $j$. For this the Kolmogorov forward differential equation in matrix form is $P'(t) = P(t)\, Q$ with initial condition $P(0) = I$ (identity matrix) and so the solution of this matrix expression can be expressed as

$$P(t) = \exp\left(Qt\right) \qquad (2)$$

For the analysis of the recurrence–progression process we have the following state space $S = \{1,\, 2,\, 3,\, 4\}$. Let $X(t),\, t \geq 0$ the state of the process in time $t$. The initial state of the process is *1*, where $X(0) = 1$. As we are interested in the effects of the *clinico–pathological* characteristics on the survival times we introduce covariates in the model to represent them via the transition

intensities. Let $z^T = (z_1, z_2, z_3, z_4)$, with $z^T$ transpose of $z$ and where $z_1 = grade$, $z_2 = number$, $z_3 = size$ and $z_4 = age$. $z_4$ is a continuous variable and the rest are dichotomic that take the value 1 according to the grade G2, two or more tumors and size $>$ 3cm. So the transition intensities will depend on the covariate vector $z$

$$q_{ij}(z) = q_{ij}\exp(z^T\beta_{ij}) \tag{3}$$

where

$$\beta_{ij}^T = (\beta_{ij}^1, \beta_{ij}^2, \beta_{ij}^3, \beta_{ij}^4) \tag{4}$$

is the vector of regression coefficients associated to vector $z$ for the transition $i \to j$, and $q_{ij}$ is the baseline transition intensity between the states $i$ and $j$. These coefficients $(\beta_{ij}^1, \beta_{ij}^2, \beta_{ij}^3, \beta_{ij}^4)$ measure the covariate effects in each transition $i \to j$ of a total of five transitions. So, the transition intensity matrix for this model is

$$Q(z) = \begin{pmatrix} -(q_{12}e^{z^T\beta_{12}} & q_{12}e^{z^T\beta_{12}} & 0 & q_{14}e^{z^T\beta_{14}} \\ +q_{14}e^{z^T\beta_{14}}) & & & \\ 0 & -(q_{23}e^{z^T\beta_{23}} & q_{23}e^{z^T\beta_{23}} & q_{24}e^{z^T\beta_{24}} \\ & +q_{24}e^{z^T\beta_{24}}) & & \\ 0 & 0 & -q_{34}e^{z^T\beta_{34}} & q_{34}e^{z^T\beta_{34}} \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{5}$$

where the generator $Q$ is conservative (file sum= 0).

Now we consider this matrix in the expression (2) and so the transition probability matrix depends on $z$ and will be denoted by $P(t;\, z)$, with entries $p_{ij}(t;\, z)$. So we can calculate the survival probabilities for different groups of patients according to the covariates $z_i$, $i = 1, 2, 3, 4$. The survival functions are expressed in terms of the transition probability functions $p_{ij}(t;\, z)$ from the matrix $P(t;\, z)$ and they are given in terms of the estimated parameters.

We use the maximum likelihood method to obtain the estimated parameters. Let suppose [14] each patient $a$ undergoes $m_a$ transitions at times $0 = t_{a,0} < t_{a,1} < \cdots < t_{a,m_a}$ where the last time can be a progression or censoring. As all patients begin in the state *1* the successive states occupied for each patient $a$ are

$$1 = x_0^a,\, x_1^a,\, x_2^a, \ldots, x_{m_a}^a. \tag{6}$$

So each patient $a$ will contribute in the likelihood function with the following factor

$$\prod_{r=1}^{m_a} p_{x_{r-1}^a,\, x_r^a}(t_{a,\,r} - t_{a,\,r-1};\, z_a), \tag{7}$$

with $z_a$ the covariate vector of the patient $a$. The last factor in this expression will be a survival function if the patient is censored. So the likelihood function for the cohort of $n$ patients is

$$L = \prod_{a=1}^{n} \prod_{r=1}^{m_a} p_{x_{r-1}^a, x_r^a}(t_{a,r} - t_{a,r-1};\, z_a) \tag{8}$$

where the transition probability functions $p_{ij}$, explained above, depend on $q_{ij}$ and the regression coefficients $\beta_{ij}$.

## 3.1 Phase–type distributions

We will use phase–type distributions in order to handle the process up to the progression. This approach has been used in [19]. Let us summarize the basic concepts.

The distribution $F(.)$ on $[0, \infty[$ is a phase–type distribution (PH-distribution) with representation $(\alpha, T)$ if it is the distribution of the time until absorption in a Markov process on the states $\{1, \ldots, m, m+1\}$ with generator

$$\begin{pmatrix} T & T^0 \\ 0 & 0 \end{pmatrix}, \tag{9}$$

with initial probability vector $(\alpha, \alpha_{m+1})$ where $\alpha$ is a row m-vector.

The matrix T of order $m$ is non-singular with negative diagonal entries and non-negative off-diagonal entries and satisfies

$$-Te = T^0 \geq 0, \tag{10}$$

where $e$ denotes a column vector with all components equal to one.

The distribution $F(.)$ is given by

$$F(x) = 1 - \alpha \exp(Tx)e, \qquad x \geq 0 \tag{11}$$

and the survival function by

$$S(t) = \alpha \exp(Tt)e. \tag{12}$$

# 4 Results

We are interested on to know the effects of risk factors that influence on each transition and the survival curves. So we need to estimate the transition intensity values $q_{ij}$ and the regression coefficients $\beta_{ij}$. The parameter estimation is performed using the maximum–likelihood method, but we need previously to take a seed for both types of parameters.

We applied the AG model to our database and according with the obtained baseline hazard function we have $0 < q_{ij} < 0.2$ for our reference individual [9]. So we perform vectors ($q_{12}$, $q_{14}$, $q_{23}$, $q_{24}$, $q_{34}$) with $q_{ij}$ random numbers in the interval [0, 0.2]. We choose the vector that gives a minimum of $-L$, with $L$ the likelihood function. We use this vector as a seed of the Nelder–Mead algorithm [20] implemented in the *fminsearch* Matlab function that gives a local minimum.

The seeds for the estimation of parameters $\beta_{ij}$ are obtained from our results in a previous work [13] by the WLW model [12] (extension of Cox model).

So, we have the following estimated parameters

$$\begin{array}{ll} \hat{q}_{12} = 0.0725 & \hat{\beta}_{12} = (0.604,\ 0.827,\ \ 0.652,\ 0.012) \\ \hat{q}_{14} = 0.0016 & \hat{\beta}_{14} = (2.200,\ 0.169,\ -0.602,\ 0.076) \\ \hat{q}_{23} = 0.0839 & \hat{\beta}_{23} = (0.575,\ 0.921,\ \ 0.940,\ 0.004) \\ \hat{q}_{24} = 0.0103 & \hat{\beta}_{24} = (2.200,\ 0.169,\ -0.602,\ 0.076) \\ \hat{q}_{34} = 0.0274 & \hat{\beta}_{34} = (2.200,\ 0.169,\ -0.602,\ 0.076) \end{array} \tag{13}$$

In a second step from these obtained seeds we use the Nelder–Mead algorithm [20] to guess the estimated parameters in the maximum likelihood estimation. The algorithm did not find any local minimum, so we again generate random vectors around the seeds and finally we obtain an estimation of global minimum of $-L$ in the domain of parameters. The obtained parameters are given in Table 1 and Table 2.

We obtain the estimated transition probability function from a patient of 60 years old, with grade G1, two or more tumors and size $< 3$ cm. Table 3 represents the transition probabilities for this patient in one, two and three years.

From the generator of the process, given by equation 5, it is very easy to obtain the survival function with respect to progression, by means of PH

| | |
|---|---|
| $\hat{q}_{12} = 0.1119$ | s.e.=0.0301 |
| $\hat{q}_{14} = 0.0995$ | s.e.=0.0107 |
| $\hat{q}_{23} = 0.1138$ | s.e.=0.0264 |
| $\hat{q}_{24} = 0.0699$ | s.e.=0.0290 |
| $\hat{q}_{34} = 0.0850$ | s.e.=0.0222 |

Table 1: Estimated baseline transition intensities $\hat{q}_{ij}\,(s.e)$ in each transition.

| Transition | ***grade*** | ***number*** | ***size*** | ***age*** |
|---|---|---|---|---|
| $1 \rightarrow 2$ | 0.6063 (0.0057) | 0.8178 (0.0065) | 0.6504 (0.0058) | 0.0134 (0.0045) |
| $1 \rightarrow 4$ | 2.2012 (0.0061) | 0.1750 (0.0052) | -0.6109 (0.0073) | 0.0722 (0.0062) |
| $2 \rightarrow 3$ | 0.5706 (0.0054) | 0.9191 (0.0059) | 0.9457 (0.0065) | 0.0089 (0.0076) |
| $2 \rightarrow 4$ | 2.1930 (0.0069) | 0.1756 (0.0070) | -0.5994 (0.0051) | 0.0727 (0.0049) |
| $3 \rightarrow 4$ | 2.2047 (0.0052) | 0.1784 (0.0052) | 0.6018 (0.0047) | 0.0837 (0.0052) |

Table 2: Estimated regression coefficients $\hat{\beta}_{ij}\,(s.e)$ for each transition.

| Transition | ***1 year*** | ***2 years*** | ***3 years*** |
|---|---|---|---|
| $1 \rightarrow 2$ | 0.0152 | 0.0225 | 0.0250 |
| $1 \rightarrow 4$ | 0.2796 | 0.4800 | 0.6240 |
| $2 \rightarrow 3$ | 0.0119 | 0.0160 | 0.0162 |
| $2 \rightarrow 4$ | 0.2147 | 0.3858 | 0.5211 |
| $3 \rightarrow 4$ | 0.4295 | 0.6746 | 0.8143 |

Table 3: Transitions probabilities to one, two and three years.

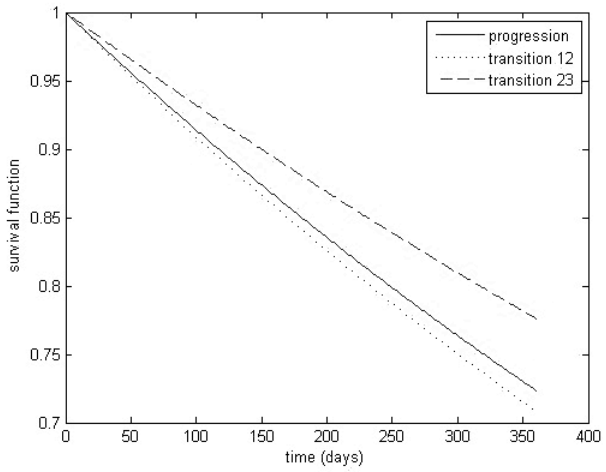



Figure 2: Survival functions in one year period.

distributions. It is simply

$$S(t) = \alpha \exp(Tt)e. \tag{14}$$

We obtain the survival curves (Figure 2)from the transition probabilities and the expression 14, and they have been plotted with the Matlab programme. The curves corresponding to progression and transitions $1 \rightarrow 2$ and $2 \rightarrow 3$ are compared for the first year.

# References

[1] Greenlee, R. T., Hill-Harmon, M. B., Murray, T. and Thun, M. (2001). Cancer Statistics, 2001 *CA Cancer J. Clin*, 51:15–36.

[2] deVere White, R. W., Stapp, E., Murray, T. and Thun, M. (1998). Predicting prognosis in patients with superficial bladder cancer. *Oncology*, 12:1717–1723.

[3] Hölmang, S., Hedelin, H., Anderström, C. and Johansson, S. L. (1995). The relationship among multiple recurrences, progression and prognosis of patients with stage ta and t1 transitional cell cancer of the bladder followed for at least 20 years. *J Urol*, 153:1823–1827.

[4] Cox, D. R. (1972). Regression models and life tables (with discussion). *J Roy Statist Soc*, Series B 34:187–220.

[5] García, B., Rubio, G., Santamaría, C., Pontones, J. L., Vera, C. D. and Jiménez, J. F. (2005). A predictive mathematical model in the recurrence of bladder cancer. *Math Comp Model*, 42:621–634.

[6] Kaasinen, E., Rintala, E., Hellstrom, P., Viitanen, J., Juusela, H. and Rajala, P. (2002). Factors explaining recurrence in patients undergoing chemoimmunotheraphy regimens for frequently recurring superficial bladder carcinoma. *Eur Urol*, 42:167–74.

[7] Millán-Rodríguez, F., Chéchile-Toniolo, G., Griffiths, D. F. and Matthews, P. N. (2000). Multivariate analysis of the prognostic factors of primary superficial bladder cancer. *Br J Urol*, 163:73–78.

[8] Santamaría, C., García–Mora, B., Rubio, G. and Pontones, J. L. (2008). Modelling the Recurrence of Bladder Cancer. *Acta Appl Math*, 104:91–105.

[9] Santamaría, C. (2006) *Modelización matemática de los factores de riesgo en el carcinoma vesical superficial. Nomogramas de predicción de recaída para el seguimiento individualizado de los pacientes.* PhD thesis. Departamento de Matematica Aplicada. Universidad Politécnica de Valencia.

[10] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann Statis*, 10:1100–20.

[11] Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68:373–389.

[12] Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84:1065–1073.

[13] García–Mora, B., Santamaría, C., Rubio, G. and Pontones, J. L. (2008). Modeling the recurrence–progression process in bladder carcinoma. *Computers and Mathematics with Applications*, 56:619–630.

[14] Pérez–Ocón, R. Ruiz–Castro, J. E. and Gámiz–Pérez M. L. (1998). A multivariate Model to Measure the Effect of Treatments in Survival Studies to Breast Cancer. *Biometrical Journal*, **40**, 6, 703–715.

[15] Asmussen, S. (1997) Phase–type Distributions and Related Point Processes: Fitting an recent advances, in S.R. Chakaravarty & A.S. Alfa (eds.), Matrix–Analytic Methods in Stochastics Models, 137–149, Marcel Decker, New York.

[16] Andersen, P. K., Borgan, T., Gill, T. D. and Kleiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer–Verlag.

[17] Neuts, M. F. (1981). *Matrix Geometric Solutions in Stocastic Models. An Algoritmic Approach*. Johns Hopkins Univ Press.

[18] OMS (1999). *International Classification of Tumours*. World Health Organization, Histological typing of urinary bladder tumours, 2 edicin. Volumen 10, Geneva.

[19] Pérez–Ocón, R. Ruiz–Castro, J. E. (2003). A Multiple–Absorbent Markov Process in Survival Studies: Application to Breast Cancer. *Biometrical Journal*, **45**, 7, 783–797.

[20] Nelder, J. A. and Mead, R. (1987). A simplex method for function minimization. *Computer Journal*, 7:308–313.

# A mathematical model for the open circuit voltage recovery of commercial batteries *

J. L. Guiñón⋆, A. Igual⋆, and N. Thome †

(*) Departamento de Ingeniería Química y Nuclear.

Universidad Politécnica de Valencia. E–46022 Valencia, Spain.

(†) Instituto de Matemática Multidisciplinar

Universidad Politécnica de Valencia. E–46022 Valencia, España.

December 11, 2008

In this paper, approach models for the voltage recovery of batteries are found from experimental data. Then the physical-chemical phenomenon and its dynamical behavior is analyzed. By using these models the evolution of the open circuit voltage recovery of $MnO_2$/Zn commercial batteries (alkaline, superalkaline, and heavy duty, at low states of charge) is studied.

In the last years new types of batteries are appearing. However, the primary alkaline, superalkaline, and heavy duty batteries remain widely used in a remarkable variety of applications. Small sizes such as AA and AAA cells, even demonstrated constant growth over the past decade as result of continuous improvement of their overall cell performance. This gain is related to major improvements in engineering and cell components, which resulted in lower cell impedance and lower environmental impact relative to toxicity, recycling and disposal of the active material used [1, 2].

Open circuit voltage recovery of discharged cells has been subject of several works in the last years. The first works were addressed to establish the discharge mechanism and degradation of the cells. At the discharge of $MnO_2$/Lithium and $MnO_2$/Zn cells it has been established that the difference of behavior of the voltage recovery was greatly affected by the nature of the $MnO_2$ used (chemical manganese dioxide and electrolytic manganese dioxide) [3, 4, 5].

Djordjevic [6] developed the calculated discharge curve algorithm (CDCA) in order to generate the actual discharge curve of alkaline manganese cell LR-20 (VARTA commercial) after a continuously discharge through constant load $10\Omega$, to the cutoff voltage of 0.8 V. In this study the empirical power function $U_{cell}(t) = 1.011(t - 308160)^{0.015}$ was used to fit the open circuit voltage recovery vs time.

*This work was partially supported by DGI grant MTM2007-64477.

†Corresponding Author e-mail: njthome@dma.upv.es

The effect of the additives in the discharge and open circuit potential recovery of Zn hexacyanoferrate prussian blue and Cu hexacyanoferrate Prussian blue, and Ni-Fe cells was studied by Jayalakshmi in [7]. In these works, experimental data were presented without a theoretical approach of the curves obtained. It is to be noted that usually in the studies conducted to date, efforts have been made to correlate the final cell voltage with the degradation mechanisms of the cells. Some papers provide an analysis of open circuit voltage recovery based on the involved electrochemical process [8, 9, 10, 11, 12].

In this paper, a mathematical treatment is proposed for describing the evolution of the open circuit voltage recovery of $MnO_2$/Zn commercial cells, alkaline, superalkaline, and heavy duty, at low states of charge. First, we explain the process for obtaining the experimental data. Then, approach models for recovering of batteries are found. Finally, the physical-chemical phenomenon and its dynamical behavior is analyzed and the process of fitting the curves is shown. The work was restricted to commercial sealed batteries.

Manganese dioxide-zinc commercial cells of AA size of nominal voltage of 1.5 V, also identified as LR6 alkaline, charge 2.50 A-hr, ZR6 superalkaline, charge 2.80 A-hr, and R6 heavy-duty (electrolyte $ZnCl_2$), charge 1 A-hr, were tested at room temperature. The composition and characteristics of the cells are shown in Table 1 where the meaning of (a), (b), and (c) is the following:

(a) The outer case in the cell is zinc alloy.

(b) Rest of components: case, collectors, electrolyte, separator.

(c) Open cell voltage. Nominal voltage = 1.5 V.

Table 1: Composition and characteristics of the cells

| **Characteristics** | **Alkaline** | **Superalkaline** | **Heavy duty** |
|---|---|---|---|
| Total mass cell, g | 23.09 | 23.21 | 17.50 |
| Zn, g | 3.42 | 3.45 | 4.20 (a) |
| (C+$MnO_2$), g | 10.53 | 10.55 | 3.42 |
| $MnO_2$, g | 9.48 | 9.47 | 3.14 |
| Rest (b), g | 9.14 | 9.20 | 9.95 |
| OCV (c), Volts | 1.621 | 1.723 | 1.628 |
| Theoretical capacity, Zn, Amp-hr | 2.80 | 2.83 | 3.44 |
| Theoretical capacity, $MnO_2$, Amp-hr | 2.92 | 2.92 | 0.97 |
| Nominal capacity, Amp-hr | 2.50 | 2.80 | 1.00 |
| Theoretical energy, Zn, Watts-hr | 4.53 | 4.88 | 5.60 (a) |
| Theoretical energy, $MnO_2$, Watts-hr | 4.73 | 5.03 | 1.58 |

The cells were discharged continuously through constant load 10, 4, 2, 1 ohms at room temperature to the cutoff voltage, $V_{cutoff} = 0.2V$, followed by an off-load relaxation of the system i.e., the open-circuit recovery. A computer was used to record the discharge voltage and the open circuit voltage recovery. For the discharge curve the period of reading was $\Delta t = 1$ minute, and for the relaxation curve when the voltage increases sharply the period of reading was $\Delta t = 1$ s until the 1 or 2 hr and after when the curve is stabilized tends to asymptote the period remains 10 or 30 minutes. The last data of open circuit voltage was taken at 24 hr. All the graphics have been obtained by using the MATLAB® language [13].

The alkaline cell presents a slow decreasing of the cell voltage type plateau until about 0.9 V and a sharp falling until 0.2 V. Otherwise, for the heavy duty cell the voltage cell decreases intensely through the all the discharge. Since the ohmic drop at the electrode causes an instantaneous variation in the cell voltage on interruption, these transients are quite steep at their start. Subsequent to the initial rise, the cell voltage increases gradually with time and approaches the equilibrium value asymptotically. For both cells the open circuit voltage represents the thermodynamic cell voltage due to the equilibrium that exits between the actives species unreacted and discharged products given by the Nernst expression.

The process of attainment of equilibrium becomes increasingly sluggish as the state of charge (SOC) of the cell (remaining capacity) decreases. The cell capacity is the integrated current over time, and is determined by the area that lies beneath the curve. It can be seen that at the first discharge, the approach to the final equilibrium voltage for the alkaline cell is more fast than for the heavy duty cell, being the SOC 38 and 18 respectively, as it is shown in table 2. This behavior is probably because of the fact that at low states of charge, the surface of the grains of electrode material is enriched with the discharge products, while active material remains in the interior of the porous electrode. The diffusion and the redistribution of the active material from the interior of the grains are slower with increase in the depth of discharge and, hence, the equilibrium cell voltage will be approached more slowly.

Table 2: Open cell voltage (OCV) and state of charge (SOC) (remaining charge/cell capacity).

| Discharge | Heavy duty | | Alkaline | |
|---|---|---|---|---|
| | OCV (mV) | SOC (%) | OCV (mV) | SOC (%) |
| 0 | 1.628 | 100 | 1.621 | 100 |
| 1 | 1.263 | 18 | 1.16 | 38 |
| 2 | 1.04 | 11 | 1.079 | 15 |
| 3 | 0.902 | 5 | 0.93 | 9 |

The evolution of the open circuit voltage recovery of $MnO_2$/Zn commercial batteries is

studied for the alkaline, superalkaline and heavy duty cases. From the experimental data we can approximate the open circuit voltage recovery of the alkaline cell after of the first discharge through constant load of 4 ohms to the cutoff voltage of about 0.2 V.

A good approximation for the obtained curve is given by the expression

$$V(t) = G(1 - Be^{-At}) \tag{1}$$

where the constant $G$ is known and $A$ and $B$ are parameters which will be found by means of the least-square fitting. In fact, by making some algebraic manipulations and applying logarithms to the equation (1) we get the linear model

$$\tilde{V} = -At + C \tag{2}$$

where

$$\tilde{V} = \log\left(1 - \frac{V}{G}\right), \qquad C = \log(B),$$

denoting the natural logarithm by $\log(.)$. Substituting the data $(t, V)$ (recordered previously in an auxiliary file.m) in the model (2) and using $G = 1169$ one has to solve the normal equations

$$N^T N \begin{bmatrix} -A \\ C \end{bmatrix} = N^T b$$

where $N$ is the $2385 \times 2$ matrix with the values of $t$ in the first column and ones in the second one, and $b$ is the column vector defined as $b = \log(1 - V/1169)$ for the data recordered in mV. The transpose of the matrix $N$ is denoted by $N^T$. Concretely, in this case we have

$$1.0e{+}009 * \begin{bmatrix} 4.524983385 & 0.002845305 \\ 0.002845305 & 0.000002385 \end{bmatrix} \begin{bmatrix} -A \\ C \end{bmatrix} = 1.0e{+}007 * \begin{bmatrix} -1.30854644455222 \\ -0.00085130196299 \end{bmatrix}.$$

Since the column of the matrix $N$ are linearly independent, the only solution of this compatible system is $A = 0.0025911920204$ and $C = -0.47810818379313$. So, the approximated model is given by

$$V(t) = 1169(1 - 0.61996e^{-0.00259t}).$$

The corresponding model for the superalkaline batteries is presented by using an expression as in (1) for the approaching. Following a similar reasoning as before the obtained model in this case is given by

$$V(t) = G(1 - 0.61996e^{-0.00259t}).$$

Finally, we present the model for the heavy duty batteries by using again an expression as in (1) for the corresponding approach. In this case the model has the expression

$$V(t) = G(1 - 0.37809e^{-0.00125t}).$$

The present work describes the discharge curves through constant loads at low states of charge and the evolution of the open circuit voltage recovery of $MnO_2$/Zn commercial cells. Approach models of the voltage recovery have been analyzed for explaining the different behavior of the alkaline, superalkaline, and heavy duty batteries.

# References

[1] G. Scholl, W. Baumann, & A. Muth. European ecolabel for batteries for consumer goods. Ecological Economics Research Institute, Heidelberg, 1997.

[2] C.A.C. Sequeira. *Environmental oriented electrochemistrty*. Elsevier, Amsterdam. 1994.

[3] I. Tanabe, & N. Miyamoto. Proceedings of The Electrochemical Society, 85, 493-524, 1985.

[4] J.C. Nardi. Proceedings of The Electrochemical Society, 85-4, 419-443, 1985.

[5] A. Kozawa. *Baterries* Vol.1 (ed Kordesch). M. Dekker, New York, 1974.

[6] A.B. Djordjevic and D.M. Karanovic. Journal of Powers Sources, **83**, 134-140 (1999).

[7] M. Jayalakshmi and F. Scholz. Journal of Powers Sources, **91**, 217-223 (2000).

[8] M. Durga Prasad and S. Sathynarayana. Journal of Applied Electrochemistry, **17**, 463-472 (1987).

[9] M.S. Suresh, A. Subrahmanyam and K. Usha. Journal of Powers Sources, **56**, 171-178 (1995).

[10] K. Vijayamohanan, A.K. Shukla and S. Sathynarayana. Electrochimica Acta, **36**, 369-380 (1991).

[11] A.K. Sleigh and W.R. Mckinnon. Electrochimica Acta, **35**, 1849-1854 (1990).

[12] K. Vijayamohanan, A.K. Shukla and S. Sathynarayana. Journal of Powers Sources, **21**, 53-57 (1987).

[13] R. Pratap. *Getting Started with MATLAB7*. Oxford University Press, 2006.

# A comparison of ROI-based procedures for progressive transmission of digital images

I. Baeza[*], J.-A. Verdoy[⋆], J. Villanueva-Oller[⋆], R.-J. Villanueva[†].

(*)Instituto de Matemática Multidisciplinar,

Universidad Politécnica de Valencia, 46022, Valencia, España

(†) Escuela de Ingeniería Técnica de Informática de Sistemas (Ces Felipe II)

Aranjuez, Madrid, España

December 11, 2008

## 1 Introduction

The use of digital images has extended to reach most of the aspects of our daily life. We take pictures with our mobile phone or digital camera, if we break the speed limit at the highway the police radar flashes and takes a digital image of our car, the film we see on the cinema, the computed tomography the doctor takes of our broken arm, or the satellite image used for predicting the weather, all them have been taken in digital format. Moreover, all them must be processed, stored, and, eventually, retrieved and displayed.

We will focus now on the aspects of the big-sized images. A good example is the medical imaging. A Computed Tomography (CT) was composed in the '70s, of a couple of images of, say 100 kilobytes, while nowadays a complete 3D ultrasound image of 500 frames of 512x512 pixels sizing more than 500 megabytes is a common practice. Intensive use of CT, magnetic resonances (MR) and ecographies produce an huge output of terabytes, information that must be processed and stored, and later retrieved and displayed when physicians make their diagnosis. Or let us think about aerial, satellite or

[*]Corresponding author: ibaeza@imm.upv.es


Figure 1: Example of big-sized image.

surveillance imaging that handle images which cover at the same time large areas with a great level of detail. In Figure 1 we can find a good example in our reference image. This image has 5000x3120 pixels with 16 bits per pixel of grey levels in origin, with a total of 31200000 bytes, were each pixel covers roughly one centimetre. Sending this image using a modem can take several hours, even if it has been compressed (for example, high quality JPEG compression still requires 5 to 7 Megabytes to allocate the reference image).

## 1.1 Progressive transmission and ROIs

To solve this problem we can use progressive transmission schemes able to provide compression (for storage and transmission) and efficient visualisation (reconstruction). On the one hand, using conventional "sequential" transmission, image can not be shown until complete transmission ends. Even if complete transmission is not required, the obtained image usually is useless until most of the data have been received.

On the other hand, progressive transmission schemes arrange, somehow, the image data in such a way that image can be seen since the very beginning

of transmission and image quality improves as more data are received.

Another interesting aspect of progressive transmission is that great level of effective compression can be achieved indirectly because observer can stop data transmission when he or she feels that enough detail level has been reached or image turns out to be uninteresting [1, 2]. Additionally, if observer is able to interact with transmission process in order to select regions of interest (ROI) when he/she realises that some region of the image is relevant, only information concerning the ROI is sent.

### 1.2 Lossless and lossy

After reception of the complete image data set and later reconstruction, the reconstructed image could be identical to the original one. In this case, we call the process "lossless transmission". Alternatively, the reconstructed image could be not identical to the original image, i.e. part of the original information has been lost and can not be recovered. In this case, we call the process "lossy transmission". Lossy schemes have a great advantage: They need by far less information than lossless schemes to reconstruct the image.

Each method is suitable for a different sort of data:

- Lossless processing is the only one acceptable for data such as text or a database, where any loss of information is noticeable and can lead to corruption of the entire data set.
- Lossy processing is suitable for those sorts of data where certain amount of noise (lack of accuracy) is acceptable, such as images, video, or sound.

## 2 Dicrete Cosine Transform and JPEG

JPEG (Joint Photographic Expert Group) compression standard [3] for images is a well-known format that has been used since the early 90's until now. JPEG defines two basic compression methods, the Baseline method based on a lossy DCT scheme, and a predictive method for lossless compression. We will focus on the Baseline method, as it is the most used and the one that offers better compression.

## 2.1 The DCT encoding

DCT stands for Discrete Cosine Transform, and is related to the Discrete Fourier Transform (DFT). Image is divided into groups of 8x8 pixels. The pixel values are shifted from unsigned integers with range $\left[0, 2^{P-1}\right]$ to signed integers with range $\left[-2^{P-1}, 2^{P-2}\right]$ and DCT is applied for each group. This is done for performance reasons, as a DCT calculation of the complete image requires a huge amount of time. If we name the image as F, and F(u,v) as the pixel (u,v) of the image, the DCT is calculated as

$$F(u,v) = \frac{1}{4} C(u) C(v) \left[ \sum_{x=0}^{7} \sum_{y=0}^{7} f(x,y) \cdot \cos\left(\frac{(2x+1)u\pi}{16}\right) \cdot \cos\left(\frac{(2y+1)u\pi}{16}\right) \right]$$

$$where\ C(u) = C(v) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } u, v = 0; \\ 1 & otherwise \end{cases}$$

After applying this transformation, each 8x8 block is represented as 64 basis-signal amplitudes or "DCT coefficients" [4]. The (0,0) coefficient is called "DC Coefficient" and the remaining 63 are called "AC coefficients". For a typical 8x8 block of a typical image, most of the coefficients are close to zero and can be discarded. Therefore only the most significant ones are stored. The more coefficients we discard, the higher compression rates we reach but also the more original image information we lose.

## 2.2 Quantization

After obtaining the DCT coefficients, they are uniformly quantized using a 64-element Quantization Table. This table is a set of integers in the range [1,255] that specifies the step size of the quantizer for its corresponding DCT coefficient, so we can represent each coefficient with no more precision than it really needed and increase the compression ratio. The quantization table is not unique but depends on how much we want to compress the image and how much quality we can lose. This step is the principal source of loss of information in the DCT-based encoders [4].

Quantization is defined as the division of each DCT coefficient by its corresponding quantized step size, rounded to the nearest integer:

$$F^{Q}(u,v) = INT\left(\frac{F(u,v)}{Q(u,v)}\right)$$

## 2.3 Coefficient encoding

The DC coefficient is stored as the difference from the DC term of the previous block in the encoding order. Then, all the 64 coefficients are ordered following a zigzag sequence, so low-frequency coefficients are stored before high frequency ones.

Once the coefficients have been ordered, they are packed using Huffman or arithmetic coding [4].

## 2.4 Reconstruction

The reconstruction process is straightforward: The coefficients are unpacked, rearranged, de-quantized and the inverse DCT applied to recover the original 8x8 pixel values. Notice that calculations are made using 8x8 pixel blocks due to the high CPU time required for this method.

## 2.5 Progressive transmission using DCT

For the progressive DCT-based mode, the 8x8 blocks are encoded in the same order but in multiple scans through the image, with each scan incorporating additional information. This can be done starting with the More Significant Bit (MSB) of each coefficient in the first scan and adding one or more bits in following scans, until the Least Significant Bit (LSB) is reached, or coding a initial subset of the 64 coefficients, say the first ten, and adding more in the incoming scans.

## 2.6 ROI handling in JPEG

JPEG does not include a specific scheme for ROI encoding and decoding. Of course, many approaches have been developed [6, 7]. Some of them [8] select a ROI from the original image, using more bits for the ROI and discarding bits for the sections out of the ROI. A great increase of compression rate can be achieved, of course, but with the disadvantage that the ROI must be known a priori and it is not possible to choose arbitrary regions of the image once it has been coded.

We propose a new scheme to carry on with the experiment, similar to the JPEG progressive transmission explained before, but after the first set of DCT coefficients of each 8x8 block, we wait until the receiver asks for more

detail in a specific ROI. Then, the 8x8 blocks that are included in that ROI are selected, and additional coefficients are sent just for these blocks. This has several advantages:

- There is no need of ROI definition before the image codification,
- It is possible to choose any arbitrary ROI or select several ROIs to be sent in a single step, and
- There is no need of any sort of recalculation in the sender side.

However, the trade-off is an increase in the calculations needed to perform the reconstruction on the receptor's side, as inverse DCTs must be recalculated repeatedly as the additional image information arrives.

# 3 Wavelets and JPEG2000

JPEG 2000 [9] is a standard image compression created for the same committee that created JPEG, and is intended to replace the JPEG standard [4, 10]. The main differences between JPEG and JPEG 2000 are:

1. The image is divided into so-called tiles. Each tile is a rectangular non-overlapping image region that is to be encoded separately. Images can have just one tile (the entire image itself) or many, but all of them must have the same size (except, maybe, the tiles located in the borders). The advantage is that it is possible to encode different tiles with different quality, or decode only selected tiles.
2. DCT is replaced by Discrete Wavelet Transform (DWT). Specifically, one of two possible wavelets can be used:

- Cohen-Daubechies-Feauveau 9/7 (CDF 9/7) wavelet transform [11]. This is a lossy transformation.
- Cohen-Daubechies-Feauveau 5/3 (CDF 5/3) wavelet transform [11]. This is a lossless transformation.

3. ROI management is supported directly. ROIs are encoded in such a way that they are placed in the data stream first and with a higher detail level, therefore ROIs are decoded before the rest of the image and furthermore they look better.

The standard JPEG2000 scheme which uses ROIs is the so called MAXSHIFT [12, 13]. It is called like this due to the fact that the pixels of the ROI (square or circular) are shifted before applying the DWT.

Nevertheless, the MAXSHIFT approach has some limitations:

- By design, it is assumed that there is only one ROI in the whole image.
- The encoder must know where the ROI is BEFORE starting the encoding.
- DWT interlaces data of the image and this is an important drawback to use JPEG 2000 with ROIs [14].
- For more than the usual 8 bpp colour levels (i.e. 16 bpp graylevel images), shifting the ROI can saturate the whole image.

# 4 SVD for adaptive encoding and ROI transmission

In this section we present an SVD encoding of images that allow the transmission of several ROIs at the same time when the receiver selects them.

First, we propose an algorithm for lossy SVD encoding.

## 4.1 SVD Algorithm 1

The SVD algorithm which allows the ROI transmission can be split in the following steps:

1. Apply SVD to the image.
2. Define the regions to be transmitted as $\{u, \sigma, v\}$ where $\sigma$ is the singular value and $u, v$ their corresponding singular vectors.
3. Compute the first $\sigma^*$ such that $\sqrt{\frac{\sum_{\sigma_i > \sigma^*} \sigma_i^2}{\sum_{\sigma_i} \sigma_i^2}} \leq \varepsilon$, [15] where $\varepsilon$ is a prefixed threshold. The theshold allows to
   (a) remove the smaller singular values because they are assumed to not provide substantial improvement in the image reconstruction

(b) as a practical rule for 2-D images, if $\varepsilon = 0.05$, then matrix B is an acceptable approximation of A [16].

4. Reject the regions with singular values $< \sigma^*$.

5. Quantize the regions to be saved, ordered, in an encoded file.

Once the image is encoded, ROI transmission is based on the following property of SVD: If

$$\{\sigma, (u_1, ..., u_m), (v_1, ..., v_n)\}$$

is a singular value and the associated singular vectors of a 2-D image of size $m \times n$, the product

$$\sigma(u_1, ..., u_m)^T(v_1, ..., v_n)$$

is an approximation to the 2-D image and

$$\sigma(u_s, ..., u_{s+k})^T(v_t, ..., v_{t+l})$$

is an approximation to the ROI of size $k \times l$ with upper left corner coordinates $(s, t)$ .

When the client detains the Progressive Transmission to select the ROIs, control matrices are created both in the server and in the client that take the account of the singular values and singular vector transmitted to avoid the redundancies.

Grouping each singular value with its singular vectors $\{u, \sigma, v\}$ transmitting this groups ordered by singular value, the image can be improved when a new group is transmitted and received.

On the one hand, there are important advantages in this scheme:

- The reconstruction algorithm is pretty fast and has a straightforward implementation $\left\{u, \sigma, v^T\right\}$
- There is no redundancy in data transmission.
- Several (rectangular) ROIs can be selected at the same time if required.
- Once the ROIs are selected and their coordinates transmitted to the server, re-encoding the original image to transmit the ROIs progressively, is NOT required.


Figure 2: Selected ROI of reference image.

- Already transmitted data are different from the new transmitted data to improve the quality of the ROIS. There is not redundancy in data transmission.

On the other hand, a drawback is that the encoded image using SVD increases its size respect to the original size (between 150%-190%) .Quantization is usually required.

# 5 Experiments

Figure 1 is a black and white image of 3000x3120 pixels with 16 bits of grey levels. We will use this image as reference, with the following subimages detailed in Figure 6 selected as ROI.

## 5.1 Experiment I: SVD versus JPEG 2000

This situation is not the usual one for a progressive transmission system, as it requires the a priori knowledge of what are the ROIs, but has been choosen as it is the only one supported by JPEG2000.

The first step is the definition of a ROI containing both the Antenna and the Handbag regions.

Once the ROI has been selected, the image is encoded accordingly to JPEG2000 scheme, using MAXSHIFT for the ROI.


Figure 3: (a) JPEG2000 ROI and (b) SVD ROI at ratio 0.78.

The second step is the application of algorithm 1 (SVD lossy encoding) to the whole image.

We start the progressive transmission of the ROIs using both methods, SVD encoding and MAXSHIFT encoding from ratio 0.01 bits/pixel (19547 bytes) to ratio 0.78 (1521063 bytes) with increments of 0.01 bits/pixel.

The image quality is measured in each reconstruction with Peak Signal-to-Noise Ratio

$$PSNR = 10 \log_{10} \left( \frac{MAX_I^2}{MSE} \right)$$

The MAXSHIFT encoding and the further reconstruction were carried out with Kakadu 6.0 [17].

We stop the process when a reconstruction with PSRN greater than 30dB is reached, as it is considered as "good quality" [18].

As we can see in Figure 8, SVD and JPEG 2000 are comparable at low ratios, with a great advantage for SVD at the very beginning. In the first step of transmission, JPEG 2000 requires quite a big quantity of data, as the ROI to be reconstructed is bigger (Figure 8- a) than the ones of SVD (Figure 8 - b):

After that, as more and more data is used for the reconstruction JPEG 2000 turns out to provide a better PSNR at higher ratios than SVD.

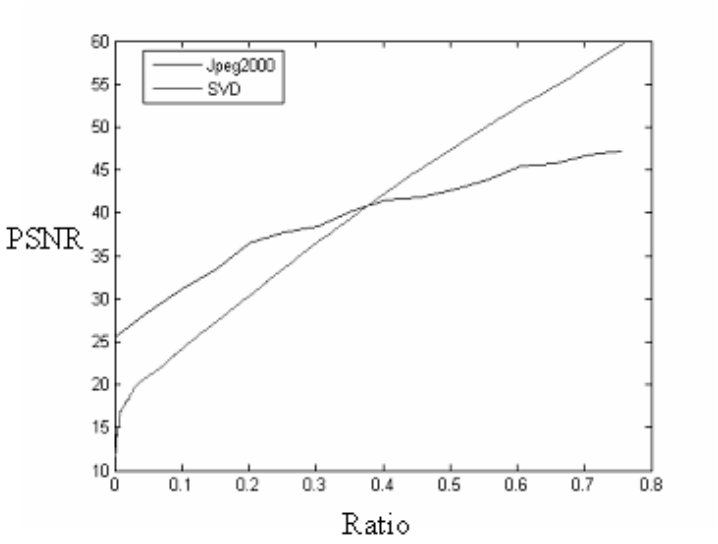



Figure 4: JPEG2000 versus SVD.

## 5.2 Experiment II: SVD versus DCT

This experiment describes a more realistic situation of a real-world progressive transmission. In this case we do not have previous knowledge of about the ROIs to be sent.

At the first step, for DCT we send the first coefficient of each 8x8 block of the image and reconstruct the whole image with the inverse DCT. This produces a PSNR of 22.9 dB.

To compose a similar situation for SVD, we perform a reconstruction of the image using the same number of bits as in DCT method. This produces a PSNR of 21.72 dB.

After the first reconstruction has been received, the client stops the transmission and examines the result. A region of interest emerges and as a result the client selects the ROI. Both methods resume the transmission sending only data corresponding to the ROIs.

- With DCT

A coefficient of each $8x8$ blocks that contain the ROI are transmitted. After each transmission, the ROI is improved using the inverse DCT and PSNR is computed.

(a) DCT first (b) SVD first

Figure 5: DCT and SVD first step.

- With SVD

Regions of singular values with their singular vectors $(\sigma, u, v)$ are sent with the same number of bits to match each transmission performed in DCT method. Then, PSNR is computed.

We repeat the transmission steps until all the 64 coefficients of each block that contains the ROI have been sent for DCT. At the same time, we improve the SVD reconstruction transmitting the same amount of data groups as in DCT method.

As we can see in Figure 8, DCT has a better performance than SVD. However, it should be taken into account the pixelation effect when few data have been transmited and the well-known high cost, time and computational, of the calculus of the DCT inverse for reconstruction of the image at each step.

# 6 Conclusions

In this paper we compare three proposed algorithms for lossy and adaptive encoding of digital images 2D based on SVD, DCT and JPEG2000. This encoding is useful for progressive transmission of images and an algorithm for reconstruction is developed.

In addition, the developed encoding allows the progressive transmission of ROIs, sequentially or several at the same time, with simple changes in transmission and reconstruction algorithms, avoiding re-encoding and redundancy

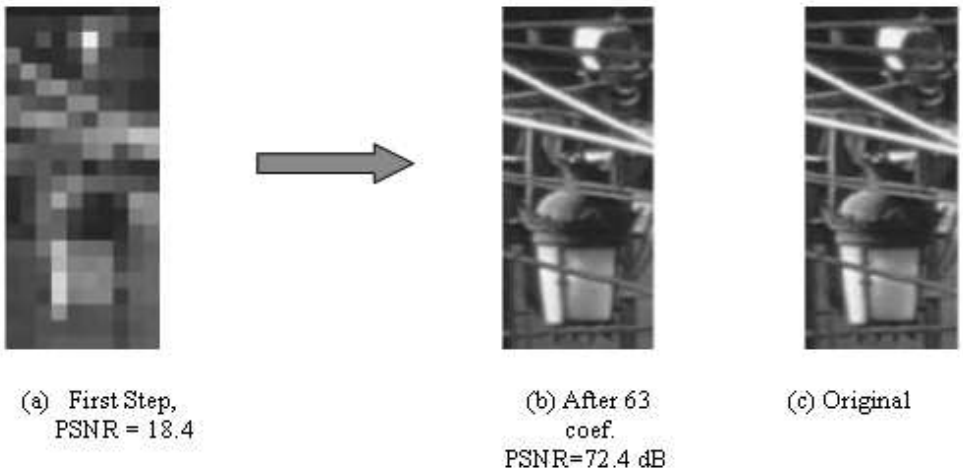



Figure 6: Magnified DCT ROI at first step (a), and final reconstruction (b) using all 64 DCT coefficients.

in data transmission. Moreover, the client can select the ROIs at any step during the transmission process.

Furthermore, a comparative with JPEG and JPEG2000 has been carried out, with different results. These results must be understood accordingly to the limitations of each one of the schemes shown here.

In the first comparative, JPEG 2000 MAXSHIFT versus SVD, there can be only one ROI, and this ROI must be defined a priori, as MAXSHIFT has no way to select ROIs on the fly. Under this circumstance, JPEG 2000 is the best method known for whole images. On the other hand, if the client selects other ROI, the whole image should be encoded again and the transmission should be re-started from the beginning. Redundancy and interlaced data are the main drawbacks of this method.

The SVD encoding increases the amount of data but wise use of data quantization can alleviate this drawback in with exchange of losing of image quality. On the other hand, SVD allows definition and handling of several ROIs at the same time on the fly, and there is no data redundancy as the encoding is done only once. This saves also computation time.

Both methods are comparable at low ratios.

In the second comparative, SVD versus JPEG, we have choose a special scheme for JPEG using just the JPEG's DCT, as this format has no native

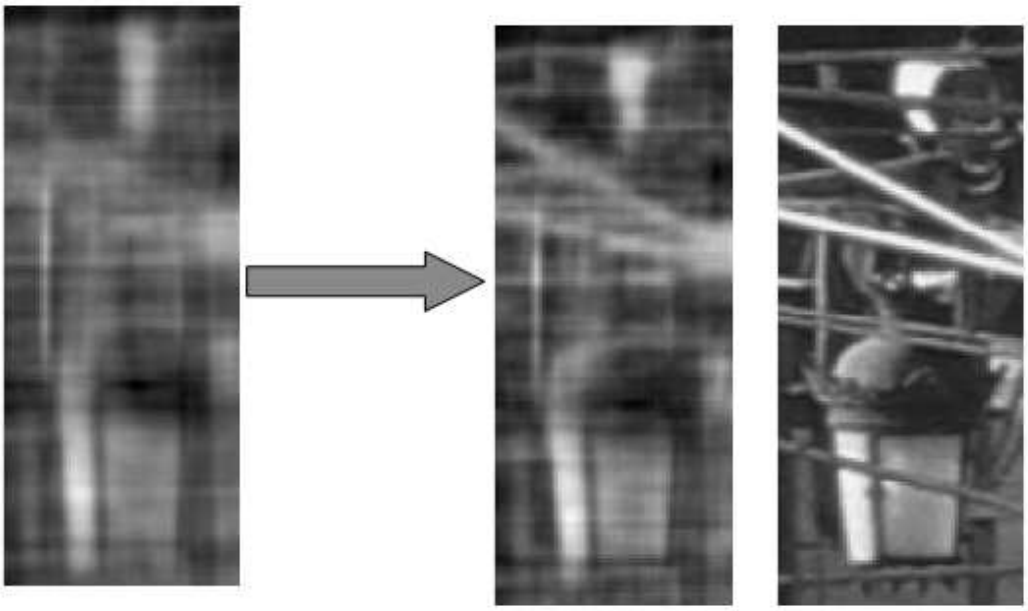



Figure 7: Magnified SVD ROI at first step (a), and final reconstruction (b) using same number of bits than in DCT after 63 coefficients.

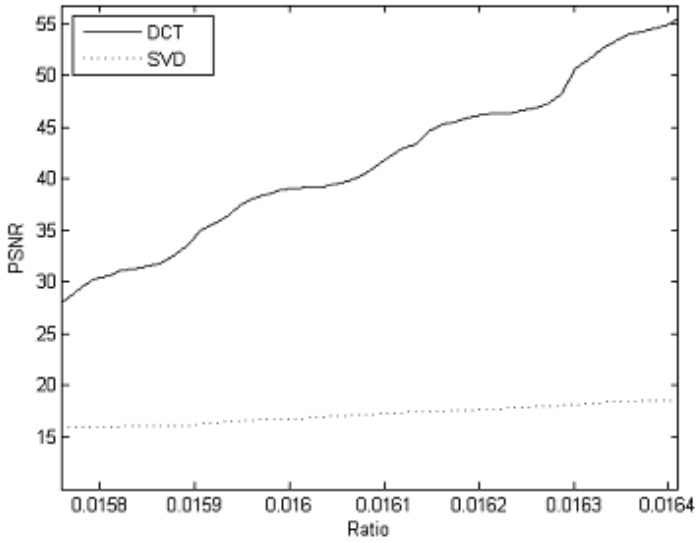



Figure 8: DCT versus SVD.

ROI handling. Using the proposed scheme, we discovered that DCT is a good method for progressive transmission of ROIs. It encodes the image increasing the resulting size, with pixelation observed at the first stage of data transmission. Reconstruction is carried out with inverse DCT at each step and consequently, leading to a high computational cost as main inconvenient. Despite of SVD ROI reconstruction being smoother than DCT at the firsts steps, DCT is better than SVD when image data increases.

# References

[1] Kim, Y.-S. , Kim, W.-Y., *Reversible correlation method for progressive transmission of 3-D medical image*, IEEE Trans. Medical Imaging, vol. 17, no. 3 (1998), pp. 383-394.

[2] Tzou, K. , *Progressive image transmission: A review and comparison of techniques*, Opt. Eng., vol. 26, pp. 581-589, July 1987.

[3] JPEG Standard (JPEG ISO/IEC 10918-1 ITU-T Recommendation T.81).

[4] G. Wallance, *The JPEG Still Picture Compression Standard*, IEEE Tansactions on Consumer Electronics, Volume 38, Issue 1, Feb 1992 Page(s):xviii - xxxiv.

[5] http://en.wikipedia.org/wiki/JPEG.

[6] T.E. Liptay, J. Barron, I. Gargantini, *A WWW interactive progressive local image transmission system* Ph.D., Department of Computer Science, The University of Western Ontario London, Canada, May 31, 2000.

[7] A.Vlaciu, S. Lungu, N. Crisan, and S.Persa. *New compression techniques for storage and transmission of 2-D and 3-D medical images*, In Advanced Image and Video Communications and Storage Technologies, volume 2451, pages 370-7, Amsterdam, Netherlands, March 1995.

[8] N. Panagiotidis, S. Kollias, *Region-Of-Interest Based Compression of Magnetic Resonance Imaging Data*, Proceedings IWISP 1996, pp. 31-35.

[9] JPEG 2000 standards, http://www.jpeg.org/jpeg2000/CDs15444.html

[10] http://en.wikipedia.org/wiki/JPEG_2000.

[11] http://en.wikipedia.org/wiki/Cohen-Daubechies-Feauveau_wavelet.

[12] Pervez Akhtar, Muhammad Iqbal Bhatti, Tariq Javid Ali1 & Muhammad Abdul Muqeet, *Significance of ROI Coding using MAXSHIFT Scaling applied on MRI Images in Teleradiology-Telemedicine*, in the J. Biomedical Science and Engineering, 2008, 110-115.

[13] George K. Anastassopoulos and Athanassios N. Skodras. *JPEG2000 ROI coding in medical imaging applications* in the Visualization, Imaging, and Image Processing (VIIP 2002), 9/9/2002 - 9/12/2002 Marbella, Spain. ISBN 0-88986-354-3.

[14] G. Menegaz, J.P. Thiran, *Lossy to lossless object-based coding of 3-D MRI data*, IEEE Transactions on Image Processing 11-9 (2002) 1053-1061.

[15] F. Pedroche, *On some capabilities of the SVD expansion to handle images*,WSEAS Transactions on Mathematics 1-2 (2002) 67-70.

[16] J.S. Walker, *A primer on wavelets and their scientific applications*, Chapman& Hall/CRC, 1999.

[17] http://www.kakadusoftware.com/

[18] D.S. Taubman, M.W. Marcellin, *JPEG2000: Image compression fundamentals, standards and practice*, Kluwer, 2002.

# Blobs-based algebraic reconstruction methods using polar grids *

C. Mora* †, M. J. Rodríguez-Álvarez‡, and I. Baeza‡

(*) Institute of Corpuscular Physics of Valencia,

Edificio Institutos de Investigación, 22085 Valencia, Spain

(‡) Instituto de Matemática Multidisciplinar,

Universidad Politécnica de Valencia,

Edificio 8G, Piso 2, 46022 Valencia, España.

December 11, 2008

## 1 Introduction

Computed tomography has had a revolutionary impact in medicine and a significative role in industrial applications. As a result of this, tomographic technology has had a quick development which has involved a continuos renewal of reconstruction algorithms.

Fourier-based methods are the most common reconstruction technique but algebraic reconstruction techniques (ART) provide more quality reconstruction than Fourier-based methods under noisy conditions or with fewer projections. Thus, a wealth of studies is aimed to computational cost reduction of these methods.

ART methods consist in a linear system of equations which relates the projections with the reconstructed images. Images features are highly determined by the system matrix model. Therefore, system matrix calculation

*This work is partially supported by the *Generalitat Valenciana* grant GVPRE2008\303.

†e-mail: cimomo@doctor.upv.es

implies a tradeoff between an accurate model and a computational efficient model. Reconstructed images quality has been improved using the modified Kaiser-Bessel window functions, also known as blobs. In one hand, system matrix calculations using blobs are highly simplified, in the other hand, blob functions lead to a system matrix less sparse than the matrix calculated using conventional pixels. A less sparse system matrix implies an increase of the reconstruction process complexity and an increase of the reconstruction time.

To place blobs on alternative grids, such hexagonal or bicubic grids, has allowed one to reduce the pixel number [1]. Nevertheless, these approaches are based on cubic arrangement of pixels. Since polar pixels grid has shown to be capable of reducing ART computational demands [2], here two alternative approaches using blobs placed on a polar grid are presented. System matrixes are used to reconstruct phantoms from simulated data for cone-beam CT scanner geometry and the reconstructed images are analyzed. Moreover, these grids are compared to conventional grid which is calculated using cubic pixels placed on regular cubic grid. All the grids are compared among them, focusing on reconstruction accuracy and computational cost.

## 2 Discretization by means of Kaiser-Bessel functions

Discrete images can be described as a sum of scaled and shifted copies of a 3D function which are placed on a grid, as follows,

$$f(x,y,z) = \sum_{i=1}^{N-1} c_i \Psi(x - x_i, y - y_i, z - z_i) \tag{1}$$

where $\Psi$ is a voxel basis function, $c_i | i = 1, 2, ..., N-1$ is the set of coefficients to represent the image and $(x_i, y_i, z_i) | i = 1, 2, ...N$ is the set of sample points which are nodes of the grid over the region represented in the reconstructed images.

Cubic voxels functions placed on Cartesian regular grid is the most widespread option but several studies has proven that this is not the most efficient one in terms of quality and computational cost [3]. Since Lewis [4] proposed

blobs as an alternative to cubic voxels several publications have explored the advantages of blobs pixels [1, 5, 6].

Blobs are spherically-symmetric functions which correspond to a Kaiser-Bessel window,

$$b^{(m,\alpha)} = \begin{cases} \frac{1}{I_m(\alpha)}\sqrt{(1-(r/a)^2)^m} I_m(\alpha\sqrt{(1-(r/a)^2)}), & 0 \le r \le a \\ 0, & other\ case \end{cases} \quad (2)$$

where $a$ is the blob radius, $I_m$ is the maximum value of the Kaiser-Bessel modified function of order $m$ which control decay smoothing in border $r = a$ and $\alpha$ which is a non-negative parameter responsible of blob shape. As equation 2 shown, the function have the maximum value in the centre $(x, y, z)$ and its value is decaying to zero with increasing distance from their centre. This shape allow us to obtain a mixture value of blobs neighbors. As a consequence of the superimposition of elements, blobs images are smoother than images using voxels. Although the reconstruction noise is less in blobs images than in voxels images, computational burden of reconstruction is increased if blobs were used.

Standard blob of $m = 2$, $a = 2.0$ and $\alpha = 10.4$ has been proposed by Matej and Lewitt [5] as the blob most suited shape for Positron Emission Tomography applications. Thus, standard blob is the one chosen in this work.

# 3 Implementation issues

The main shortcoming of blobs is due to the increase of computational cost associated to them. Polar scheme has shown to be able to reduce computational demands of system matrix calculation [7]. Furthermore, polar grid fits better to the reconstructed area using cylindrical coordinates and makes possible to take benefits of all the symmetries of the tomographic system. These symmetries allow ones to reduce the computational requirements of algebraic iterative reconstruction [8]. Actually, the polar grid had proof to be able to provide an accurate reconstruction using a circular-symmetric basis functions different to blobs [2].

Blobs are placed using two polar grid approaches: One polar grid based on a constant blob radius and another based on variable blob radius.

Polar grid based on constant radius approach is focused on reducing the number of blobs and restricting the maximum distance between samples to the scanner spatial resolution. Thus, radial samples are fixed to the spatial resolution and the angular samples are calculated to tend to the spatial resolution value in order to reduce the number of blobs.

Sampling using a polar scheme usually supposes desigual radial and angular samples. Polar grid based on constant ratio between angular and radial samples is designed to minimize differences between both samplings in order to achieve square-like grids [2]. Since a constant ratio between angular and radial samples is fixed to one, blobs have a variable radius in this case and this radius is equal to the distance between the radial samples.

Both discretization schemes using blobs placed on polar grids have been implemented for a real cone beam geometry CT scanner. A system matrix has been calculated for each discretization approach to reconstruct simulated projection using the Maximum Likelihood Expectation Maximization (MLEM) algorithm [9].

Simulated projections have been calculated taking into account geometric considerations only. A median density cylinder which is filled with high density cylinders and low density cylinders is one of the simulated phantoms. A well-known in CT studies head of Shep-Logan [10] is used too, in order to study the quality of the reconstruction. These simulated signals allow us to measure errors between the reconstructed images and the real model. These errors provide us an objective measure in order to compare both discretization approaches using blobs.

Projections in order to calculate the system matrix assumes that each ray is composed by a set of lines. This approximation simplifies the integral calculation and makes possible a generalization of the problem for all the possible intersections between rays and blobs or voxels. The sum of all lines contributions allow us to obtain the total ray weight. Naturally, the accuracy of this technique depends on the number of lines. The number of lines enough for the case of study has been established by a previous work [11].

# 4 Results

The reconstructed images calculated from blobs placed on polar grids are compared with square pixels on regular grids for a 2D particular case of CT-Simulator scanner. The number of blobs on polar grid of constant radius

is 88.85% lower than pixels number; however, the number of blobs for the constant ratio polar grid approach increased 8.26% above the pixels number. Nevertheless, iteration time is decreased in both cases respect to the reconstruction using conventional pixels. The temporary cost of reconstruction shows that it can be obtained a decrease of 12.36% in the ratio constant case and 5.6% in the constant radius case of reconstruction time per iteration than in the pixels case. Time per iteration is reduced even in the constant ratio polar grid case despite the increment in the pixels number. This is due to use the system matrix symmetries in the reconstruction algorithm. The matrix symmetries are increased 100 times using polar pixels grid respect to square pixels grid. Consequently, these symmetries allow system matrix to reduce its size to approximately the hundredth part of the square grid system matrix.

The reconstructed images are analyzed by measuring indicators of reconstruction quality which are the Root Mean Square Error (RMSE), the Coefficient of Variation (CV) and the Contrast Recovery Coefficient (CRC). Those indicators have been calculated from images reconstructed using pixels and images reconstructed using both blobs configurations for different iterations. Results for iteration number 30 of the reconstructed Shep-Logan, in which the reconstruction quality is stabilized, shows that RMSE are fixed to 0.011 for images of pixels and blobs placed on constant radius polar grid and 0.014 for images of blobs on constant ratio grid. The CRC for a low density region of the simulated phantom of cylinders shows that the similarity between the contrast phantom model and the contrast of the reconstructed phantom is 99.35% using pixels, 99.35% using blobs of constant radius and 98.2% using blobs of constant ratio. The CV shows the inhomogeneity of the regions and for a low density region is 14.9% for pixels and blobs placed on constant radius polar grid and 15.3% for blobs placed on constant ratio polar grid.

# 5 Conclusion and Discussion

CRC and CV show a higher distortion in reconstructed images using blobs with a polar grid of constant ratio than using a polar grid of constant radius. The constant ratio implies that the radius of blobs and the superimposition of blobs among their neighbors are variable. Therefore, we think that the variation in blobs smoothing may contribute to increase artifacts in reconstructed

images.

Combining blobs with polar grids of constant radius allows us to maintain the image quality equal to the one obtained using square pixels in Cartesian grids in terms of RMSE, CV and CRC. Besides, blobs placed on polar grids of constant radius make possible to reduce the computational cost increment associated to blobs because of the fact that polar grid takes advantage of all the scanner symmetries. This property allows the time per iteration to be reduced and the storage requirements to be diminished (a 11% reduction of the storage requirements than in square pixel case).

The results presented here point to the fact that the quality provided by ART methods using blobs can be maintained and the computational cost of blobs can be reduced using blobs placed on constant radius polar grid as alternative to pixels on Cartesian grid.

# References

[1] B. M Crawford and G. T Herman. Low-dose, large-angled cone-beam helical CT data reconstruction using algebraic reconstruction techniques. *Image and Vision Comp.*, 25:78–94, 2007.

[2] C. Mora, M. J. Rodríguez-Álvarez, and J. V. Romero. New pixellation scheme for CT algebraic reconstruction to exploit matrix symmetries. *Comp. and Math. with Appl.*, 53(3):715–726, Jan. 2008.

[3] S. Matej and R. M. Lewitt. Efficient 3D grids for image reconstruction using spherically-symmetric volume elements. *IEEE Trans. Nucl. Sci*, 42(4):1361–1370, Aug. 1995.

[4] R. M. Lewitt. Alternatives to voxels for image representation in iterative reconstruction algorithms. *Phys. Med. Biol*, 37(3):705–716, Dec. 1992.

[5] S. Matej and R. M. Lewitt. Practical considerations for 3D reconstruction using spherically symmetric volume elements. *IEEE Trans. Med. Imaging*, 15(1):68–79, Feb. 1996.

[6] A. Ziegler, T. Köhler, T. Nielsen, and R. Proksa. Efficient projection and backprojection scheme for spherically symmetric basis functions in divergent beam geometry. *Medical physics*, 33(12):4653–, 2006.

[7] L. Jian, L. Litaoa, C. Penga, S. Gia, and W. Zhifang. Rotating polar-coordinate ART applied in industrial CT image reconstruction. *NDT and E International*, 40(4):333–336, Dec. 2007.

[8] V. Israel-Jost, P. Choquet, S. Salmon, C. Blondet, E. Sonnendrücker, and A. Constantinesco. Pinhole SPECT imaging: Compact projection/backprojection operator for efficient algebraic reconstruction. *IEEE Trans. Med. Imaging*, 25(2):158–167, Feb. 2006.

[9] K. Lange and R. Carson. EM reconstruction algorihtms for emission and transmission tomography. *J. Comput. Assist. Tomogr.*, (8):306–316, 1984.

[10] L. A. Shepp and B. F. Logan. The fourier reconstruction of a head section. *IEEE Trans. Nucl. Sci*, 21(3):21–43, Jun. 1974.

[11] C. Mora. *Métodos de Reconstrucción Volumétrica Algebraica de Imágenes Tomográficas*. PhD thesis, Polytechnic University of Valencia, Department of Electrical Engineering, 2008.

# Age-structured mathematical modeling of Respiratory Sincytial Virus (RSV) transmission dynamics in the Spanish region of Valencia

J. Díez-Domingo[⋆], Jose-A. Moraño[†], Rafael-J. Villanueva[†], Abraham J. Arenas[‡]

(⋆) Centro de Salud Nazaret, Valencia, Spain.

Centro Superior de Investigación en Salud Pública (CSISP), Valencia, Spain.

(†,‡) Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia, 46022 Valencia, España.

(‡) Departamento de Matemáticas y Estadística, Universidad de Córdoba, Montería, Colombia.

December 11, 2008

Respiratory syncytial virus (RSV) is the most important respiratory virus of young children, and the major cause of hospitalizations specially for bronchiolitis and pneumonia in infants [1]. Its impact on the Health Systems is increasing as the incidence of hospitalization in children for bronchiolitis increases [2]. It was not until recently that the impact of RSV on adults has been studied, as up to 18% of the pneumonia hospitalizations in older than 65 are due to RSV [3].

In Spain, there are $15000 - 20000$ visits to primary care due to RSV each year. In the Spanish region of Valencia, 1500 children younger than five years old with an average of 6 days of hospitalization per case are hospitalized each year by RSV bronchiolitis [4]. The cost of pediatric hospitalization for the Valencian Health System is about 3.5 million euros per year. RSV

is the cause of annual seasonal epidemics with minor variations each year and its coincident incidence with other widespread viral infections such as influenza or rotavirus, produces a high number of hospitalizations saturating the National Health Systems. Moreover, its transmission is very easy and the nosocomial infections are frequent [5].

Therefore, the research on RSV and other virus and the developing of strategies to control epidemics are important from both the sanitary and economic point of view. Other tool is the study of vaccines to protect individuals in early ages, when the immune system is not completely developed and, what is more important, to control the immune response because the most seriously ill cases are not due to the RSV infection, they used to be due to a child anomalous immune response [6].

Mathematical models have been revealed as a powerful tool to analyze the epidemiology of the infectious illness, to understand its behavior, to predict its social impact and to find out how external factors change the impact. In the case of RSV, the building of a reliable model is a priority to predict the medical care requirements needed in next seasons.

Mathematical models for RSV have been developed previously. For instance, in [7], a SIRS (susceptible - infectious - recovered - susceptible) and a MSEIRS (maternally derived immunity - susceptible - latent - infectious - recovered - susceptible) mathematical models with four possible re-infections are studied and applied with data from Gambia, Singapore, Florida and Finland. In [8] a nested RSV model, stochastic simulations and fitting with data from several countries are presented. Also, in [9] the authors consider the two types, A and B, of the RSV and develop a SIRS model where re-infection by any of both types of RSV virus is possible, fitting the model with data of England & Wales and Finland.

The models proposed so far took into account reinfections because the first infection do not have the same virulence as further reinfections. We have studied several models with reinfections and we found some lack of that models because the age is a variable that influence on the severity of the infection. Moreover, the everyday pediatric practice shows that the incidence of the disease is different depending on age.

The available data are only for hospitalizations in the Spanish region of Valencia by bronchiolitis (RSV and non-RSV), RSV infection and RSV-pneumonia, and only for children under 5 years. These data are for the period between January 2001 and December 2004.

In order to obtain a reliable division of the population in appropiate age

groups, only data from Hospital La Fe was taken, because it is the only hospital where RSV test are carried out to diagnose the patients. Hence, we have used data from 2034 children hospitalized in La Fe by bronchiolitis, RSV-bronchiolitis, RSV-infection and RSV-pneumonia.

We carried out a non-parametrical analysis of Mann-Whitney which shows that there is a statistically significant relation between the diagnosis (RSV, No-RSV) and the age. It has also been observed that children hospitalized for RSV had a mean age below those have been hospitalized for non-RSV bronchiolitis.

Furthermore, an analysis CHAID (Chi-Square Automatic Interaction Detection) [10] has been applied to the data. This analysis also shows the influence of age on the incidence of the virus. In particular it has been detected an age between 23 and 364 days which has a higher incidence of the virus than the rest of the children.

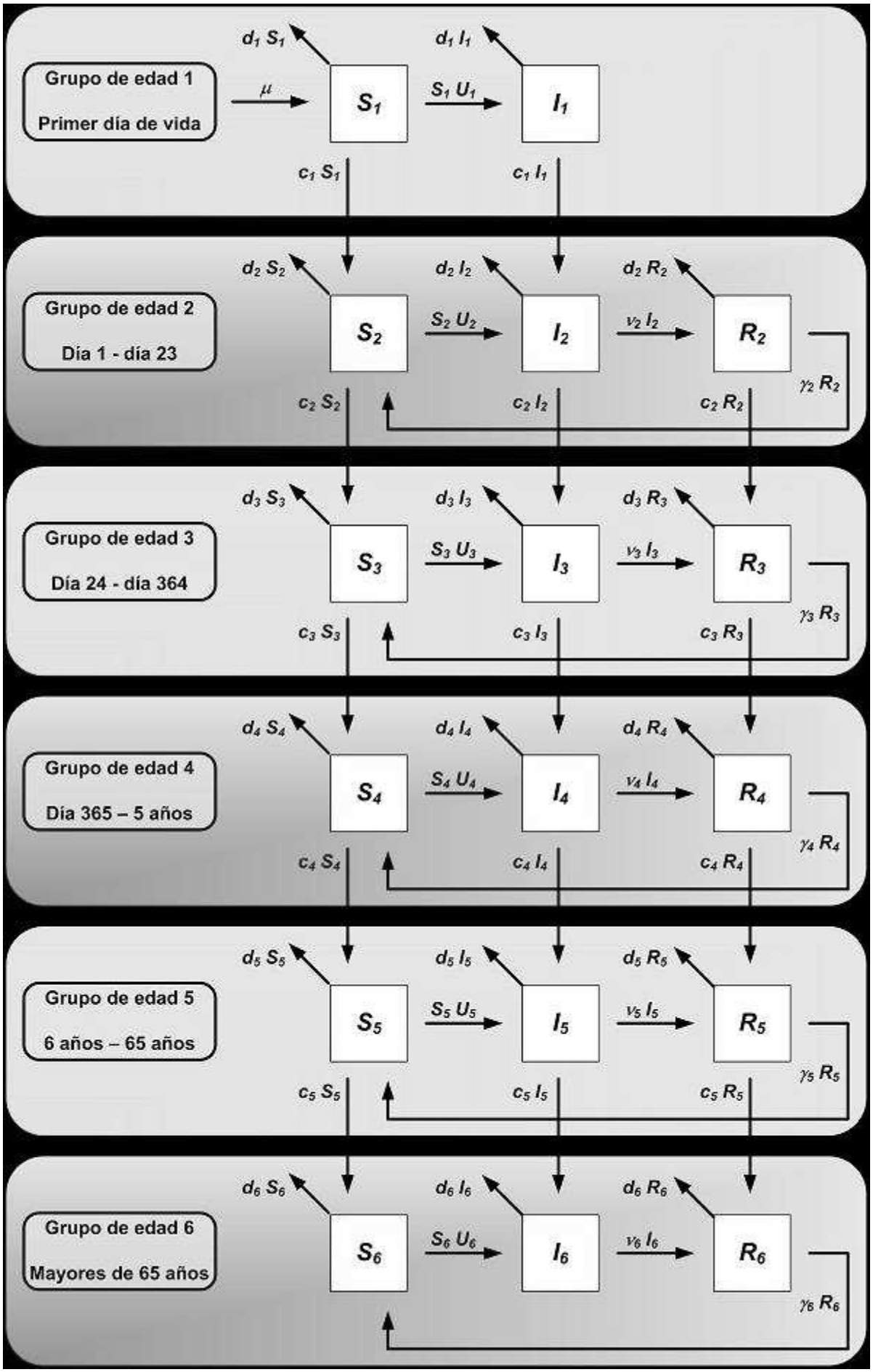
The result obtained by the CHAID analysis is acceptable because the estimated risk of error in the classification RSV/non-RSV is 0.30.

The CHAID analysis allows us to consider the following age groups:

- $G_1$, group 1, individuals with only one day of life.
- $G_2$, group 2, individuals older than 1 day of life and younger than 23 days.
- $G_3$, group 3, individuals less than 1 year old but older than 23 days. This group joint $G_2$ are the most affected by RSV and when the illness is more serious.
- $G_4$, group 4, individuals older than 1 year old and younger than 5 years old. It is an homogeneous group where the incidence of RSV is impoortant, but there is a drastic reduction of hospitalizacions.

Besides of groups obtained by CHAID analysis we have also considered two other age groups for other reasons:

- $G_5$, group 5, individuals from 6 years to 65, it is a group that we considered their defenses against the virus are developed and just pass the virus. Furthermore, in the case of infection it only produces slight symptoms.


Grupo de edad 1
Primer día de vida
Grupo de edad 2
Día 1 - día 23
Grupo de edad 3
Día 24 - día 364
Grupo de edad 4
Día 365 – 5 años
Grupo de edad 5
6 años – 65 años
Grupo de edad 6
Mayores de 65 años
$d_1 S_1$
$d_1 I_1$
$\mu$
$S_1 U_1$
$S_1$
$I_1$
$c_1 S_1$
$c_1 I_1$
$d_2 S_2$
$d_2 I_2$
$d_2 R_2$
$S_2 U_2$
$v_2 I_2$
$S_2$
$I_2$
$R_2$
$\gamma_2 R_2$
$c_2 S_2$
$c_2 I_2$
$c_2 R_2$
$d_3 S_3$
$d_3 I_3$
$d_3 R_3$
$S_3 U_3$
$v_3 I_3$
$S_3$
$I_3$
$R_3$
$\gamma_3 R_3$
$c_3 S_3$
$c_3 I_3$
$c_3 R_3$
$d_4 S_4$
$d_4 I_4$
$d_4 R_4$
$S_4 U_4$
$v_4 I_4$
$S_4$
$I_4$
$R_4$
$\gamma_4 R_4$
$c_4 S_4$
$c_4 I_4$
$c_4 R_4$
$d_5 S_5$
$d_5 I_5$
$d_5 R_5$
$S_5 U_5$
$v_5 I_5$
$S_5$
$I_5$
$R_5$
$\gamma_5 R_5$
$c_5 S_5$
$c_5 I_5$
$c_5 R_5$
$d_6 S_6$
$d_6 I_6$
$d_6 R_6$
$S_6 U_6$
$v_6 I_6$
$S_6$
$I_6$
$R_6$
$\gamma_6 R_6$

- $G_6$, group 6, individuals older than 65 years old. Elderly people which immune systems are in progressive damaging. They transmit the illness and RSV may affect them with virulence.

Once the population is divided into age groups, let us construct an age-structured SIRS model where we define $S_i, I_i, R_i;\ i = 1, ..., 6$. For the demographic model we considere that every age group has a mortality rate $d_i$. These rates as well as the number of individuals who have each age group, the birth rate, $\mu$, and the growth rates $c_i$ are obtained from the population census of Spanish Statistic Institute (INE) [11].

The spread of the disease is modeled considering the infection can happen when a susceptible individual contacts with an infectious individual of any age group. To model the transmission we considered the average number of contacts as the product of the average number of encounters by the average duration of an encounter. Different transmission effectivities have been considered depending on the individual predisposition to be infected by age group and the ability of an infectious to infect according to their age group.

On the other hand, some values of the model parameters change seasonally, taking higher values in autumn-winter in spring-summer.

# 1 References

[1] C.B. Hall, Respiratory syncytial virus and human metapneumovirus, in: R.D. Feigin, J.D. Cherry, G.J. Demmler, S.L. Kaplan (Eds.)., Textbook of Pediatric Infectious Diseases, 5th edition, Saunders, Philadelphia, PA, 2004, pp. 2315 - 2341.

[2] J.M. Langley, J.C. Leblanc, B. Smith, E.E.L. Wang, Increasing incidence of hospitalization for bronchiolitis among Canadian children 1980-2000, J. Infect. Dis. 188 (2003) 1764 - 1767.

[3] L. Han, J. Alexander, L. Anderson, Respiratory syncytial virus pneumonia among the elderly: An assessment of disease burden, J. Infect. Dis., 179 (1999) 25 - 30.

[4] J. Díez Domingo, M. Ridao López, M.I. Úbeda Sansano, A. Ballester Sanz, Incidencia y costes de la hospitalización por bronquiolitis de las infecciones por virus respiratorio sincitial en la Comunidad Valenciana. Años 2001 y 2002, An. Pediatr. 65-4 (2006) 325 - 330.

[5] C.B. Hall, Nosocomial respiratory syncytial virus infections: The cold war has not ended, Clin. Infect. Dis. 31 (2000) 590 - 596.

[6] J.E. Crowe, Respiratory syncytial virus vaccine development, Vaccine 20 (2002) 532 - 537.
[7] A.Weber, M.Weber, P. Milligan, Modeling epidemics caused by respiratory syncytial virus (RSV), Math. Biosci. 172 (2001) 95 - 113.
[8] L.J. White, J.N. Mandl, M.G.M. Gomes, A.T. Bodley-Tickell, P.A. Cane, P. Pérez-Brena, J.C. Aguilar, M.M. Siqueira, S.A. Portes, S.M. Straliotto, M.Waris, D.J. Nokes, G.F. Medley, Understanding the transmission dynamics of respiratory syncytial virus using multiple time series and nested models, Math. Biosci. (2007), doi:10.1016/j.mbs.2006.08.018.14
[9] L.J. White, M. Waris, P.A. Cane, D.J. Nokes, G.F. Medley, The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection, Epidemiol. Infect. 133 (2005) 279 - 289.
[10] Kass G. V., An Exploratory Technique for Investigating LArge Quantities of Categorical Data. Applied Statistics, 29. n$^o$ 2. 119-127 (1980).
[11] Instituto Nacional Estadística, [on-line]. Available from: http://www.ine.es.

# A difference scheme for call option pricing under transaction costs *

R. Company†, L. Jódar‡and José–Ramón Pintos §

Instituto de Matemática Multidisciplinar

Universidad Politécnica de Valencia. E–46022 Valencia, España.

December 11, 2008

## 1 Introduction

It is well-known that Black-Scholes (B-S) model is acceptable in idealized financial markets where one assumes that volatility is observable or transaction costs are not taken into account. Under the transaction costs, the continuous trading required by the hedging portfolio is prohibitively expensive, [1]. Several alternatives lead to pricing models that are equal to Black-Scholes one but with an adjusted volatility denoted by $\sigma$, see [2], [3], [4], [5] and [6],

$$V_t + \frac{1}{2}\left(\sigma(S,t,V_S,V_{SS})\right)^2 S^2 V_{SS} + rSV_S - rV = 0,\ S > 0,\ t \in [0,T[, \quad (1)$$

where $V$ is the option value that is a function of the underlying security $S$ and the time $t$. Here $r \geq 0$ denotes the riskless interest rate.

A more complex model has been proposed by Barles and Soner [1], assuming that investor's preferences are characterized by an exponential utility function. In their model the nonlinear volatility reads

$$\sigma^2 = \sigma_0^2(1 + \Psi[\exp(r(T-t)a^2S^2V_{SS})]) \ , \quad (2)$$

*This paper has been supported by the Spanish Ministery of Science and Education grant TRA2007-68006-C02-02 and the Generalitat Valenciana grant GVPRE/2008/092

†e-mail: rcompany@imm.upv.es

‡e-mail: ljodar@imm.upv.es

§e-mail: jrpt60@gmail.com

where $T$ is the maturity, and $a = \mu\sqrt{\gamma N}$ , with risk aversion factor $\gamma$ and the number $N$ of options to be sold. When $a = 0$, there is no transaction cost and classical Black-Scholes equation is recovered. The function $\Psi$ is the solution of the nonlinear initial-value problem

$$\Psi'(A) = \frac{\Psi(A) + 1}{2\sqrt{A\Psi(A)} - A} \ , \ A \neq 0 \ , \ \Psi(0) = 0 \ . \tag{3}$$

In the mathematical literature, only a few results can be found on the numerical discretization of B-S equation, mainly for linear B-S equations. The numerical approaches vary from finite element discretizations [7, 8], to finite difference approximations [9]. The numerical discretization of the B-S equations with the nonlinear volatility (3) has been performed using explicit finite-difference schemes [1]. However, explicit schemes have the disadvantage that restrictive conditions on the discretization parameters (for instance, the ratio of the time and the space step) are needed in order to obtain stable, convergent schemes [10]. [11] combines high-order compact difference schemes derived by [12] and techniques to construct numerical solutions with frozen values of the nonlinear coefficient of the non-linear B-S equation to make the formulation linear. Reasonable numerical strategies like the consideration of more discretization nodes near the maturity and the strike price, are not sufficient to guarantee reliability and accuracy of the numerical approximations. Careless numerical computations may waste a good mathematical model.

In this paper we deal with an European vanilla call option pricing equation (1) where $\sigma$ is given by (2)-(3), together with final and boundary conditions taking the form

$$\left.\begin{array}{c} V(S,T) = max(0, S-E), \ S > 0, \\ \\ V(0,t) = 0, \ \lim\limits_{s\to\infty} \dfrac{V(S,t)}{S - Ee^{-r(T-t)}} = 1. \end{array}\right\} \tag{4}$$

Using the change of variable $\tau = T - t$, $U(S,\tau) = V(S,t)$ problem (1)-(4) apart from the asymptotic condition is transformed into

$$U_\tau - \frac{S^2}{2}\sigma^2 U_{SS} - rSU_S + rU = 0 \ , \ \ 0 < S < \infty \ , \ \ 0 < \tau \leq T \ , \tag{5}$$

$$U(S,0) = \max(0, S-E) \ . \tag{6}$$

## 2 Properties of the correction of volatility function

Dealing with numerical analysis of difference schemes presented in next sections, is going to be convenient to bound the approximation of the nonlinear term involving the volatility correction function $\Psi$ appearing in (2)-(3)-(5). The next three properties of $\Psi$ are established.

(P1) From theorem 1.1 of [13] it is known that $\Psi(A)$ is an increasing function mapping the real line onto the interval $]-1, +\infty[$ and $\Psi(A)$ is implicitly defined by

$$A = \left( -\frac{\text{Arcsinh } \sqrt{\Psi}}{\sqrt{\Psi+1}} + \sqrt{\Psi} \right)^2 , \quad if \ \Psi > 0 \, , \tag{7}$$

$$A = -\left( \frac{\arcsin \sqrt{(-\Psi\,)}}{\sqrt{\Psi+1}} - \sqrt{-\Psi} \right)^2 , \quad if \ -1 < \Psi < 0 \, . \tag{8}$$

(P2) $\Psi(A)$ is a convex function for $A > 0$ and $0 < \Psi(A) \leq \Psi'(A_2)A + d_2, \ \ A > 0$ where $\Psi'(A_2) \approx 1.10, \ \ d_2 \simeq \ 2.62$.

$\Psi(A)$ is a concave function for $A < 0$.

(P3) Let $\Psi(A)$ be volatility correction function appearing in (2) verifying equation (3) and let $g(A) = \ A\Psi(A)$. Then $g(A)$ is continuously differentiable at $A = 0$ and satisfies

$$|g'(A)| \leq \max\{G, \ 2|A|\Psi'(A_2) + d_2\}, \quad A \in \mathbb{R} \tag{9}$$

where $A_2$ and $d_2$ are given by

$$A_2 = \left( \sinh \ 2 - \frac{2}{\sqrt{(\sinh \ 2)^2 + 1}} \right)^2 \simeq 9.58 \, , \ \ \Psi(A_2) = (\sinh \ 2)^2,$$

$$A_1 = -\frac{(4\pi - 3\sqrt{3})^2}{36} \, ; \quad G = \max\{|g'(A)|; \ \ A_1 \leq A \leq A_2\}. \tag{10}$$

## 3 Scheme construction.

The computation of numerical solutions of the model is necessary because an exact solution is not available. In order to compute the numerical solution

it is necessary to work in a bounded domain. We choose an interval $[E - L, E + L]$ where $E$ is the strike price and $0 \leq L \leq E$ is agree with the criteria given in [14].

By replacing in (5) the partial derivatives by finite difference (FD) approximations, [15], [16], one gets a numerical explicit scheme centered in the underlying asset variable and forward in the time variable, given by

$$0 = F(u_j^n) = \frac{u_j^{n+1} - u_j^n}{k} - \frac{S_j^2}{2}\sigma_0^2 \left(1 + \Psi_j^n(u)\right) \triangle_j^n(u) - rS_j \nabla_j^n(u) + ru_j^n, \quad (11)$$

or explicitily

$$u_j^{n+1} = u_j^n + k\left(\frac{S_j^2}{2}\sigma_0^2 \left(1 + \Psi_j^n(u)\right) \triangle_j^n(u) + rS_j \nabla_j^n(u) - ru_j^n\right), \quad (12)$$

where $u_j^n \simeq U(S_j, nk)$,

$$S_{j+1} - S_j = \triangle S = h, \quad hN = 2E, \quad 0 \leq j \leq N-1,$$
$$\tau^{n+1} - \tau^n = \triangle\tau = k, \quad lk = \tau, \quad 0 \leq n \leq l-1.$$

$$\nabla_j^n = \frac{u_{j+1}^n - u_{j-1}^n}{2h}, \quad (13)$$

$$\triangle_j^n = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}, \quad (14)$$

and a linear interpolation approximation has been taken at the boundary points $E - L$ and $E + L$.

Numerical scheme (11) can be written in vector form

$$u(\tau) = \left[\prod_{n=l-1}^{n=0} (I + kM(nk))\right] u(0), \quad (15)$$

where

$$u(\tau) = u(lk) = \left[u_1^l, ..., u_{N-1}^l\right],$$

$u(0)$ involves the initial condition at the mesh points and $M(nk)$ is a tridiagonal $(N-1) \times (N-1)$ matrix such that at every row $j$, $2 \leq j \leq N-2$ which non zero entries $(M(nk))_{jj-1}$, $(M(nk))_{jj}$ and $(M(nk))_{jj+1}$ are given by the respectively coefficients of $u_{j-1}^n$, $u_j^n$ and $u_{j+1}^n$ of the bracket in (12) taking into account (13) and (14). Entries in the first and the last rows are given by the considered linear interpolation, at the boundary mesh points.

# 4 Numerical analysis.

Once the scheme is computed numerical analysis of the solution is developed and two main results are established.

A) Stability

FD scheme (15) is conditionally time stable for appropriate fixed values of $h = \triangle S$ in the sense that there exists $C(\tau)$ such that

$$\|u(\tau)\| \leq exp\left(\frac{C(\tau)}{h^2}\tau\right)\|u(0)\|, \quad 0 \leq \tau \leq T.$$

Where $\|\cdot\|$ denotes the euclidean vector norm.

B) Consistency

Dealing with reliable numerical computations of FD schemes, the consistency of the difference-scheme with the equation is a necessary requirement because this means that the exact theoretical solution of the partial differential equation approximates well to the exact solution of the difference equation as the stepsizes tend to zero, [16].

Let us represent equation (5) by $L(U) = 0$, and let $F(u_j^n) = 0$ represent the approximating difference equation defined by (11) with exact solution $\{u_j^n\}$. In accordance with [16, p.100], the FD scheme is consistent with (5) if

$$T_j^n(U) = F(U_j^n) - L(U_j^n) \ \rightarrow \ 0, \ as \ h = \triangle S \ \rightarrow \ 0, \ k = \triangle t \rightarrow \ 0, \quad (16)$$

where $U_j^n$ denotes the theoretical solution of (5) evaluated at the point $(S_j, \ nk)$, i.e., $U_j^n = U(S_j, nk)$. If in (16) one has that $T_j^n(U) = O(h^p)+O(k^q)$, then we say that the FD scheme is consistent of order $(p, q)$.

**Theorem 1** *Let* $S = E - L + jh, \quad 0 \leq \tau = nk \leq T \quad with \quad h = \triangle S,$ $k = \triangle\tau \quad with \quad 2 \leq j \leq N-2 \quad and \quad 2L = Nh.$ *Then the FD scheme (11) at the point* $(S, \tau)$ *is consistent of order* $(2, 1)$ *with equation (5)*

# 5 Example.

Consider the vanilla call option with transaction costs and parameters

$$\sigma = 0.2, \quad r = 0.1, \quad E = 100, \quad \tau = 1 \ year, \quad h = 4, \quad k = 0.005.$$

Figure 1 shows option pricing valuation of this call option for several values of the parameter $a$ as well as the pay-off function.

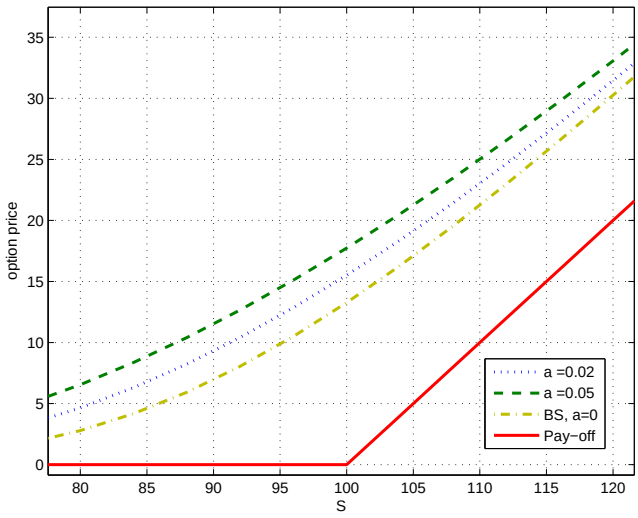



Figure 1: Valuation of a vanilla call option in both linear and nonlinear cases.

# References

[1] G. Barles, H. M. Soner, Option pricing with transaction costs and a nonlinear black–scholes equation, Finance Stochast 2 (1998) 369–397.

[2] P. Boyle, T. Vorst, Option replication in discrete time with transaction costs, J. Finance 47 (1973) 271–293.

[3] H. E. Leland, Option pricing and replication with transactions costs, J. Finance 40 (1985) 1283–1301.

[4] M. Avellaneda, A. Paras, Dynamic hedging portfolios for derivative securities in the presence of large transaction costs, Appl. Math. Finance 1 (1994) 165–193.

[5] A. E. Whalley, P. Wilmott, An asymptotic analysis of an optimal hedging model for option pricing with transaction costs, Mathematical Finance 7 (3) (1997) 307–324.

[6] T. Hoggard, A. E. Whalley, P. Wilmott, Hedging option portfolios in the presence of transaction costs, Advances in Futures and Options Research 7 (1994) 2135.

[7] P. Forsyth, K. Vetzal, R. Zvan, A finite element approach to the pricing of discrete lookbacks with stochastic volatility, Appl. Math. Finance 6 (1999) 87–106.

[8] O. Pironneau, F. Hecht, Mesh adaption for the black and scholes equations, East-West J.Numer. Math. 8 (2000) 25–35.

[9] J. Dewynne, S. Howinson, P. Wilmott, Option pricing: mathematical models and computations, Oxford Financial Press 6 (1995) 87–106.

[10] J. C. Strikwerda, Finite difference schemes and partial differential equations, Wadsworth & Brooks/Cole Mathematics Series (1989) 32–52.

[11] B. Düring, M. Fournier, A. Jungel, Convergence of a high order compact finite difference scheme for a nonlinear black–scholes equation, Esaim–Mathematical Modelling and Numerical Analysis–Modelisation Mathematique et Analyse Numerique 38 (2004) 359–369.

[12] A. Rigal, Numerical analisys of three-time-level finite difference schemes for unsteady diffusion-convection problems, J. Num. Meth. Engineering 30 (1990) 307–330.

[13] R. Company, E. Navarro, J. R. Pintos, E. Ponsoda, Numerical solution of linear and nonlinear black–scholes option pricing equations, Computers and Mathematics with Applications 56 (3) (2008) 813–821.

[14] R. Kangro, R. Nicolaides, Far field boundary conditions for black-scholes equations, SIAM Journal on Numerical Analysis 38 (4) (2000) 13571368.

[15] D. Tavella, C. Randall, Pricing financial instruments. The finite difference method, John Wiley & Sons, Inc., New york, 2000.

[16] G. D. Smith, Numerical solution of partial differential equations: finite difference methods, 3rd Edition, Clarendon Press, Oxford, 1985.