ECCOMAS Thematic Conference

3rd International Conference on

Numerical and Symbolic Computation

Developments and Applications

PROCEEDINGS

April, 06 - 07, Universidade do Minho

GUIMARÃES, Portugal

idmec IST engenharia mecânica

APMTAC

ISBN: 978-989-99410-3-8

SYMCOMP 2017 – 3rd International Conference on Numerical and Symbolic Computation: Developments and Applications

Edited by APMTAC – Associação Portuguesa de Mecânica Teórica, Aplicada e Computacional

Editors: Maria Amélia Loja (IDMEC, ISEL/GIMOSM), Joaquim Infante Barbosa (IDMEC, ISEL/GIMOSM), José Alberto Rodrigues (ISEL/GIMOSM)

April, 2017

1 – Introduction

The Organizing Committee of SYMCOMP2017 – 3rd International Conference on Numerical and Symbolic Computation: Developments and Applications, welcomes all the participants and acknowledge the contribution of the authors to the success of this event.

This Third International Conference on Numerical and Symbolic Computation, is promoted by APMTAC - Associação Portuguesa de Mecânica Teórica, Aplicada e Computacional and it was organized in the context of IDMEC - Instituto de Engenharia Mecânica, Instituto Superior Técnico, Universidade de Lisboa. With this ECCOMAS Thematic Conference it is intended to bring together academic and scientific communities that are involved with Numerical and Symbolic Computation in the most various scientific areas

SYMCOMP 2017 elects as main goals:

To establish the state of the art and point out innovative applications and guidelines on the use of Numerical and Symbolic Computation in the numerous fields of Knowledge, such as Engineering, Physics, Mathematics, Economy and Management, Architecture, ...

To promote the exchange of experiences and ideas and the dissemination of works developed within the wide scope of Numerical and Symbolic Computation.

To encourage the participation of young researchers in scientific conferences.

To facilitate the meeting of APMTAC members (Portuguese Society for Theoretical, Applied and Computational Mechanics) and other scientific organizations members dedicated to computation, and to encourage new memberships.

We invite all participants to keep a proactive attitude and dialoguing, exchanging and promoting ideas, discussing research topics presented and looking for new ways and possible partnerships to work to develop in the future.

The Executive Committee of SYMCOMP2017 wishes to express his gratitude for the cooperation of all colleagues involved in various committees, from the Scientific Committee, Organizing Committee and the Secretariat. We hope everyone has enjoyed helping to consolidate this project, which we are sure will continue in the future. Our thanks to you all.

- Amélia Loja, Chairperson (IDMEC/LAETA, ADEM/ISEL)
- Stéphane Clain, Chairperson (CMAT/UM,UM)
- Joaquim Infante Barbosa (IDMEC/LAETA, ADEM/ISEL)
- José Alberto Rodrigues (GI-MOSM, ADM/ISEL)
- António J. M. Ferreira (FEUP/INEGI)

2 – CONFERENCE BOARD

Chairperson

Maria Amélia Ramos Loja, ISEL/GIMOSM ; IDMEC/LAETA

Área Departamental de Engenharia Mecânica

Instituto Superior de Engenharia de Lisboa

Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa

Email : amelialoja@dem.isel.ipl.pt, amelialoja@tecnico.ulisboa.pt

Co-Chairperson

Stéphane Louis Clain, Co-Chairperson (CMAT/UM)

Centre of Mathematics (CMAT)

School of Sciences, Universidade do Minho

Campus de Gualtar, 4710-057, Braga, Portugal

Email : clain@math.uminho.pt

EXECUTIVE COMMITTEE

- Amélia Loja (IDMEC/LAETA, ISEL/GIMOSM)
- Joaquim Infante Barbosa (IDMEC/LAETA, ISEL/GIMOSM)
- José Alberto Rodrigues (ISEL/GIMOSM)
- Inês Carvalho Jerónimo Barbosa (ISEL/GIMOSM)

ORGANIZING COMMITTEE

Amélia Loja, Chairperson (IDMEC, GIMOSM)

Stéphane Clain, Co-Chairperson (CMAT/UM)

Joaquim Infante Barbosa (IDMEC, GIMOSM)

José Alberto Rodrigues (GI-MOSM, ADM/ISEL)

António J. M. Ferreira (FEUP/INEGI)

LOCAL ORGANIZING COMMITTEE

Stéphane Clain (CMAT/UM,UM)

Gaspar Machado (CMAT/UM,UM)

Jorge Figueiredo (CMAT/UM,UM)

Rui Pereira (CMAT/UM,UM)

SCIENTIFIC COMMITTEE

Amélia Loja (IDMEC/IST, ISEL/GIMOSM, Lisboa, Portugal)
Ana Conceição (Universidade do Algarve, Faro, Algarve, Portugal)
Ana Neves (FEUP, , Universidade do Porto , Porto, Portugal)
António J M Ferreira (FEUP, Universidade do Porto, Porto, Portugal)
Aurélio Lima Araújo (IDMEC/IST, Universidade de Lisboa, Lisboa, Portugal)
Carlos Alberto Mota Soares (IDMEC/IST, Lisboa, Portugal)
Cristóvão Manuel Mota Soares (IDMEC/IST, Lisboa, Portugal)
Dongming Wang (Beihang University, Beijing, China and CNRS, Paris, France)
Elena Vásquez-Céndon (Universidad de Santiago de Compostela, Spain)
Gaetano Giunta (Luxembourg Institute of Science and Technology, Luxembourg)
Hélder Carriço Rodrigues (IDMEC/IST, Universidade de Lisboa, Lisboa, Portugal)
Hoon Hong (North Carolina State University, United States of America)
J. N. Reddy (Texas A&M University, United States of America)
Joaquim Infante Barbosa (IDMEC/IST, ISEL/ GIMOSM, Lisboa, Portugal)
João Manuel Ferreira Calado (IDMEC/IST, ISEL/GIMOSM, Lisboa, Portugal)
José Alberto Rodrigues (ISEL/GIMOSM, DMAT, Lisboa, Portugal)
José Miranda Guedes (IDMEC/IST, Universidade de Lisboa, Portugal)
Józef Korbicz (University of Zielona-Góra, Poland)
Juan Nuñez (University of Sevilla, Spain)
Lina Vieira (ESTeSL/IPL, Lisboa, Portugal)
Lorenzo Dozio (Politecnico Milano, Italy)
Michal Bartys (Warsaw University of Technology, Poland)
Miguel Matos Neves (IDMEC/IST, Universidade de Lisboa, Lisboa, Portugal)
Pedro Areias (Universidade de Évora, Portugal)
Paulo B. Vasconcelos (FEP, University of Porto, Portugal)
Raphael Loubère (Institut de Mathématiques de Toulouse (Unversité de Toulouse, France)
Silvio Simani (Università Ferrara, Italy)
Stéphane Louis Clain (CMAT, University of Minho, Portugal)
Teresa Restivo (FEUP, Universidade do Porto, Porto, Portugal
Xesús Nogueira (Civil Engineering School, Universidad da Coruna, Spain)

SPONSORS

ECCOMAS – European Community on Computational Methods in Applied Sciences

APMTAC – Associação Portuguesa de Mecânica Teórica, Aplicada e Computacional, (Portuguese Society for Theoretical, Applied and Computational Mechanics), ECCOMAS Member Association;

IDMEC/LAETA – Instituto de Engenharia Mecânica/Laboratório Associado de Energia, Transportes e Aeronáutica (Mechanical Engineering Institute/Associated Laboratory for Energy, Transports and Aeronautics);

CMAT – Centro de Matemática (Centre of Mathematics), Universidade do Minho

ISEL/IPL – Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa

UM – Universidade do Minho

CMG - Câmara Municipal de Guimarães, Minho, Portugal

Wolfram Research

ORGANIZING INSTITUTION

IDMEC/LAETA – Instituto de Engenharia Mecânica/Laboratório Associado de Energia, Transportes e Aeronáutica.

PLACE OF THE EVENT

Auditório da Associação Fraterna

ACG - CyberCentro de Guimarães

Edificio Cybercentro, Rua de Vilaverde

Guimarães

Contents

SYMCOMP 2017
Guimarães, 6-7 April 2017

A NEW ONE-PARAMETER INVARIANT FUNCTION FOR ALGEBRAS

J.M. Escobar[1*], J. Núñez[1] and P. Pérez-Fernández[2]

1: Dpto. de Geometría y Topología
Facultad de Matemáticas
Universidad de Sevilla
Calle Tarfia s/n. 41012-Sevilla (Spain)
e-mail: {pinchamate@gmail.com, jnvaldes@us.es}

2: Dpto. de Física Aplicada III
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla
Campus de La Cartuja. Sevilla (Spain)
e-mail: pedropf@us.es

Keywords: Invariant functions for algebras; invariant function υ; contractions of algebras.

Abstract. *Invariant functions are very useful tools for the study of contractions of algebras. In 2009, Hrivnák and Novotný introduced the ψ and φ one-parameter invariant functions and by taking their procedure into consideration we introduced in 2016 the invariant two-parameter function $\bar{\psi}$. In this communication we introduce a new one-parameter invariant function for algebras, the υ function, which is related with $\bar{\psi}$. We compute the values of this new function for several types of algebras, particularly filiform Lie algebras and Malcev algebras, and for the algebra induced by the Lorentz group $SO(3)$, which allows us to prove that the n-dimensional classical-mechanical model built upon certain types of n-dimensional Lie algebras cannot be obtained as a limit process of a quantum-mechanical model based on a n-dimensional Heisenberg algebra, for certain values of n. By using the symbolic computation package SAGE as a tool, we also conjecture that $\upsilon \geq \psi$ for any algebra, which is indeed true for algebras of lower dimensions.*

1 Introduction

In this paper, we continue a research previously initiated by Hrivnák and Novotný in 2008 and later followed by ourselves in 2016, dealing with the use of invariant function as a tool for the study of contractions of algebras. Indeed, Hrivnák and Novotný introduced in [5] the invariant one-parameter functions ψ and φ with the main objective of making easier the study of such contractions, among other applications. Later, ourselves introduced the invariant two-parameter function of algebras, denoted $\bar{\psi}$, whose study was mainly focused in Malcev algebras of the type Lie, although it could also be used with any other types of algebras [2].

We now introduce a new invariant one-parameter function, that we denote by υ, which allows us to step forward in the study of contractions of algebras, since, in a certain sense, it might be considered as much as a generalization of one by Hrivnák and Novotný as a particularization of the last introduced by ourselves. Both differences might make easier the study of such contractions, although this application of the invariant function is not developed in the paper yet and it will be dealt in future work.

The main goal of the paper is to compute the invariant function υ for the case of different types of algebras, all of them of lower dimensions. The reason for dealing with algebras of lower dimensions is because there are no complete classifications of algebras of dimensions greater than 5 and, besides, few invariants for them are known. Therefore, the use of these invariant functions can be considered as a tool to get advances in the knowledge of these algebras, particularly in the study of contractions.

The structure of the paper is as follows: in Section 2 we give some preliminaries on the algebras which we will deal with, particularly, Lie and Malcev algebras, and on the invariant bi-parameter function $\bar{\psi}_{\mathfrak{g}}$. In Section 3 we introduce the invariant function υ and we dedicate Section 4 to evaluate it for different types of algebras. We show in Section 5 the main conclusions obtained from this study.

2 Preliminaries

We show in this section some preliminaries on Lie algebras and on Malcev algebras, which are the main mathematical objects used in the paper.

2.1 Preliminaries on Lie algebras.

In this subsection we show some preliminaries on Lie algebras. For a further review on this topics, the reader can consult [4].

An n-dimensional *Lie algebra* $\mathfrak{g}$ over a field K is an n-dimensional vector space over K endowed with a second inner law, named *bracket product*, which is bilinear and anti-commutative and satisfies the *Jacobi identity*

$$J(u,v,w) = [u,[v,w]] + [v,[w,u]] + [w,[u,v]] = 0, \text{ for all } u,v,w \in \mathfrak{g}.$$

The law of the n-dimensional Lie algebra $\mathfrak{g}$ is determined by the products

$$[e_i, e_j] = \sum_{k=1}^{n} c_{ij}^k e_k, \quad \text{for} \quad 1 \leq i < j \leq n,$$

where $c_{i,j}^k \in K$ are called *structure constants* of $\mathfrak{g}$. If all these constants are zero, then the Lie algebra is called *abelian*.

Two Lie algebras $\mathfrak{g}$ and $\mathfrak{h}$ are *isomorphic* if there exists a vector space isomorphism f between them such that $f([u,v]) = [f(u), f(v)]$, for all $u, v \in \mathfrak{g}$.

A mapping $d : \mathfrak{g} \longrightarrow \mathfrak{g}$ is a *derivation* of $\mathfrak{g}$ if $d([u,v]) = [d(u), v] + [u, d(v)]$, for all $u, v \in \mathfrak{g}$. The set of derivations of $\mathfrak{g}$ is denoted by $Der\mathfrak{g}$.

The *lower central series* of a Lie algebra $\mathfrak{g}$ is $\mathfrak{g}^1 = \mathfrak{g}$, $\mathfrak{g}^2 = [\mathfrak{g}^1, \mathfrak{g}]$, $\ldots$, $\mathfrak{g}^k = [\mathfrak{g}^{k-1}, \mathfrak{g}]$, $\ldots$ If there exists $m \in \mathbb{N}$ such that $\mathfrak{g}^m \equiv 0$, then $\mathfrak{g}$ is called *nilpotent* and an n-dimensional nilpotent Lie algebra $\mathfrak{g}$ is said to be *filiform* if it is verified that $\dim \mathfrak{g}^k = n-k$, for all $k \in \{2, \ldots, n\}$.

It is convenient to note that filiform Lie algebras, which were introduced by M. Vergne in the late 60's of the past century [7], are the most structured algebras within the nilpotent Lie algebras, which allows us to use and study them easier than other Lie algebras.

2.2 Preliminaries on Malcev algebras

Now we recall some preliminary concepts on Malcev algebras, taking into account that a general overview can be consulted in [6]. From here on, we are only considering finite-dimensional Malcev algebras over the complex number field $\mathbb{C}$.

A *Malcev algebra* $\mathcal{M}$ is a vector space with a second bilinear inner composition law $([\cdot,\cdot])$ called the *bracket product* or *commutator*, which satisfies: a) $[u,v] = -[v,u]$, $\forall u \in \mathcal{M}$; and b) $[[u,v],[u,w]] = [[[u,v],w],u] + [[[v,w],u],u] + [[[w,u],u],v]$, $\forall u, v, w \in \mathcal{M}$. Condition b) is named *Malcev identity* and we use the notation $M(u,v,w) = [[u,v],[u,w]] - [[[u,v],w],u] - [[[v,w],u],u] - [[[w,u],u],v]$.

Given a basis $\{e_i\}_{i=1}^n$ of a n-dimensional Malcev algebra $\mathcal{M}$, the *structure constants* $c_{i,j}^h$ are defined as $[e_i, e_j] = \sum_{h=1}^n c_{i,j}^h e_h$, for $1 \leq i, j \leq n$.

It is immediate to see that Malcev algebras and Lie algebras are not disjoint sets. Indeed, every Lie algebra is a Malcev algebra, but the converse is not true. Therefore, we can distinguish between Malcev algebras of the type Lie and Malcev algebras of the type non-Lie.

If the Jacobi identity does not hold, then the Malcev algebra is said to have a *Jacobi anomaly*. These anomalies are relevant in Physics, for instance in quantum mechanics, String Theory, or in several other physical field where algebras called *electric* and *magnetic* appear [1]. Under a mathematical point of view, it is relevant to note that both magnetic and electric algebras constitute magma algebras (see [3] for this last concept).

If $\mathfrak{g}$ is a Malcev algebra of the type Lie and $D \in Der\mathfrak{g}$ a derivation of $\mathfrak{g}$, then it immediate to see that

$$[D[x,y],[x,z]] + [[x,y],D[x,z]] = D[[[x,z],y],x] + D[[[z,x],x],y] \quad \forall x,y,z \in \mathfrak{g} \quad (1)$$

2.3 The invariant bi-parameter function $\bar{\psi}_{\mathfrak{g}}$

In this preliminary subsection we recall the definition and main properties of the invariant bi-parameter function $\bar{\psi}_{\mathfrak{g}}$, introduced in [2].

Let $\mathfrak{g} = (V, [\,,\,])$ an algebra. If $End\,\mathfrak{g}$ denotes the vector space of all linear operators of $\mathfrak{g}$ over V, the set

$$Der_{(\alpha,\beta,\gamma,\tau)}\mathfrak{g} = \{D \in End\,\mathfrak{g} : \alpha[D[x,y],[x,z]] + \beta[[x,y],D[x,z]] = \gamma D[[[x,z],y]x] + \tau D[[[z,x],x],y]\}$$

$\forall(\alpha,\beta,\gamma,\tau) \in \mathbb{C}^4$, is called *the set of the* $(\alpha,\beta,\gamma,\tau)$*-derivations* of $\mathfrak{g}$. It is denoted by $Der_{(\alpha,\beta,\gamma,\tau)}\mathfrak{g}$. It is proved that $\dim_{(1,1,1,1)}\mathfrak{g}$ is an algebraic invariant of $\mathfrak{g}$.

In [2] we proved that if $\mathfrak{g}$ and $\bar{\mathfrak{g}}$ are two Malcev algebras of the type Lie and $f : \mathfrak{g} \to \bar{\mathfrak{g}}$ is an isomorphism of algebras, then the mapping $\rho : \text{End } \mathfrak{g} \to \text{End } \bar{\mathfrak{g}}$, defined by $D \longrightarrow fDf^{-1}$ is an isomorphism between the vector spaces $Der_{(\alpha,\beta,\gamma,\tau)}\mathfrak{g}$ and $Der_{(\alpha,\beta,\gamma,\tau)}\bar{\mathfrak{g}}$, $\forall(\alpha,\beta,\gamma,\tau) \in \mathbb{C}^4$.

This result implies that the dimension of the vector space $Der_{(\alpha,\beta,\gamma,\tau)}\mathfrak{g}$ is an invariant of the algebra, for all $(\alpha,\beta,\gamma,\tau) \in \mathbb{C}^4$.

Moreover, by using some technical lemmas, it is proved in [2] that under the previous conditions and for all $(\alpha,\beta,\gamma,\tau) \in \mathbb{C}^4$, then it exists $(\lambda_1,\lambda_2) \in \mathbb{C}^2$ such that $Der_{(\alpha,\beta,\gamma,\tau)}\mathfrak{g} \subset \mathbb{C}^2$ is one of the following four sets: $Der_{(0,0,\lambda_1,\lambda_2)}\mathfrak{g}$, $Der_{(1,-1,\lambda_1,\lambda_2)}\mathfrak{g}$, $Der_{(1,0,\lambda_1,\lambda_2)}\mathfrak{g}$ or $Der_{(1,1,\lambda_1,\lambda_2)}\mathfrak{g}$. This result allows us to give the following

Definition 2.1 *The functions* $\bar{\psi}_{\mathfrak{g}}, \bar{\psi}^0_{\mathfrak{g}} : \mathbb{C}^2 \mapsto \mathbb{N}$ *defined as* $(\bar{\psi}_{\mathfrak{g}})(\alpha,\beta) = \dim Der_{(1,1,\alpha,\beta)}\mathfrak{g}$ *and* $(\bar{\psi}^0_{\mathfrak{g}})(\alpha,\beta) = \dim Der_{(1,0,\alpha,\beta)}\mathfrak{g}$, *respectively, are called* $\bar{\psi}_{\mathfrak{g}}$ *and* $\bar{\psi}^0_{\mathfrak{g}}$ invariant functions *corresponding to the* $(\alpha,\beta,\gamma,\tau)$*-derivations of* $\mathfrak{g}$.

Note that it is immediate to prove that if two Malcev algebras of the type Lie $\mathfrak{g}$ and $\mathfrak{f}$ are isomorphic, then $\bar{\psi}_{\mathfrak{g}} = \bar{\psi}_{\mathfrak{f}}$ and $\bar{\psi}^0_{\mathfrak{g}} = \bar{\psi}^0_{\mathfrak{f}}$.

3 Introducing the new invariant one-parameter function υ

In accordance with the notations indicated in Preliminaries and starting from the definition of the invariant bi-parameter function $\bar{\psi}$, we now define

Definition 3.1 *Let* $\mathfrak{g}$ *be an algebra. The function* $\upsilon_{\mathfrak{g}} : \mathbb{C} \mapsto \mathbb{N}$ *defined as* $\upsilon_{\mathfrak{g}}(\lambda) = \bar{\psi}_{\mathfrak{g}}(1,\lambda) = \dim Der_{(1,1,1,\lambda)}\mathfrak{g}$ *is called* $\upsilon_{\mathfrak{g}}$ invariant function *corresponding to the* $(1,1,1,\lambda)$*-derivations of* $\mathfrak{g}$.

Let us observe that this function is a particular case of the invariant bi-parameter function $\bar{\psi}$ when $\alpha = 1$. This particularization presents, however, some advantages, which will be commented in Section 5.

4 The invariant function υ in the case of model filiform Lie algebras of lower dimensions

We now compute the values of the invariant function υ for model filiform Lie algebras of lower dimensions. We show here the computations related with these algebras of dimension 3, 4 and 5. The results obtained allow us to give a general expression for the value of this function in this type of algebras.

4.1 The 3-dimensional filiform Lie algebra $\mathfrak{f}_3$

Let $\mathfrak{f}_3$ be the (model) filiform Lie algebra of dimension 3 defined by the bracket $[e_1, e_3] = e_2$.

Let consider $D \in Der_{(1,1,1,\lambda)}\mathfrak{f}_3$, with $\lambda \in \mathbb{C}$ and let

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

be the matrix associated with the endomorphism D.

In this case, only one triple (e_1, e_2, e_3) can be taken into consideration.

Since $[e_2, e_3] = [e_1, e_2] = 0$, we see that when equation (1) is applied all the terms are null. This implies that there are no restrictions for the elements a_{ij} of the matrix A. Consequently,

$$\upsilon_{\mathfrak{f}_3}(\lambda) = \dim\left(Der_{(1,1,1,\lambda)}\mathfrak{f}_3\right) = 9, \quad \forall \lambda \in \mathbb{C}.$$

4.2 The 4-dimensional filiform Lie algebra $\mathfrak{f}_4$

Let $\mathfrak{f}_4$ be the (model) filiform Lie algebra of dimension 4 defined by the brackets $[e_1, e_3] = e_2, [e_1, e_4] = e_3$.

Let consider $D \in Der_{(1,1,1,\lambda)}\mathfrak{f}_4$ with $\lambda \in \mathbb{C}$ and let $A = (a_{ij})$, $1 \leq i, j \leq 4$, be the matrix associated with the endomorphism D.

We observe that the only triple producing non-null results when applying equation (1) is (e_1, e_3, e_4). Indeed, by that equation for that triple we have

$$[D[e_1, e_3], [e_1, e_4]] + [[e_1, e_3], D[e_1, e_4]] = D[[[e_1, e_4], e_3], e_1] + \lambda D[[[e_4, e_1], e_1], e_3]$$

the only non-null term is $[D[e_1, e_3], [e_1, e_4]]$. The result for that term is $[D[e_1, e_3], [e_1, e_4]] = [D(e_2), e_3] = [a_{21}e_1 + a_{22}e_2 + a_{23}e_3 + a_{24}e_4, e_3] = a_{21}e_2$.

Then, $a_{21}e_2 = 0$ if and only if $a_{21} = 0$.

It implies that the equation system which allows us to obtain the elements of the endomorphism D has one equation and 16 variables. Therefore,

$$\upsilon_{\mathfrak{f}_4}(\lambda) = \dim\left(Der_{(1,1,1,\lambda)}\mathfrak{f}_4\right) = 16 - 1 = 15, \quad \forall \lambda \in \mathbb{C}.$$

4.3 The 5-dimensional filiform Lie algebra $\mathfrak{f}_5$

Let $\mathfrak{f}_5$ be the (model) filiform Lie algebra of dimension 5 defined by the brackets $[e_1, e_3] = e_2, [e_1, e_4] = e_3, [e_1, e_5] = e_4$.

Let consider $D \in Der_{(1,1,1,\lambda)}\mathfrak{f}_5$ with $\lambda \in \mathbb{C}$ and let $A = (a_{ij})$, $1 \leq i, j \leq 5$, be the matrix associated with it.

Now, applying equation (1) to all the possible triples among the generators of the algebra, we obtain expressions involving the elements of the matrix A. Indeed, from the triple (e_1, e_2, e_3) we do not obtain any constraint because all the terms in equation (1) are null, but from (e_1, e_3, e_4) we obtain that $a_{21} = 0$. And this same result is obtained starting from the triple (e_1, e_3, e_5), whereas starting from the triple (e_1, e_4, e_5), we obtain $a_{31} = 0$.

So, the equation system which allows us to obtain a basis of the vector space $Der_{(1,1,1,\lambda)}\mathfrak{f}_5$ is constituted by 2 equations and 25 variables. This implies that

$$\upsilon_{\mathfrak{f}_5}(\lambda) = \dim\left(Der_{(1,1,1,\lambda)}\mathfrak{f}_5\right) = 23, \quad \forall \lambda \in \mathbb{C}.$$

4.4 n-dimensional filiform Lie algebras $\mathfrak{f}_n$

By proceeding in the same way, it is easy to prove the following assertion

Theorem 4.1 $\upsilon_{\mathfrak{f}_n}(\lambda) = \dim\left(Der_{(1,1,1,\lambda)}\mathfrak{f}_n\right) = n^2 - (n-3)$, $\forall \lambda \in \mathbb{C}$, *for all model n-dimensional filiform Lie algebra* $\mathfrak{f}_n$.

5 The invariant function υ in the case of other algebras of lower dimensions

In this section we deal with the invariant function υ in the case of different types of algebras, all of them of lower dimensions.

5.1 The $\mathfrak{so}(3)$ algebra

Let $\mathfrak{so}(3)$ be the 3-dimensional Lie algebra defined by the brackets $[e_1, e_2] = -e_3$, $[e_2, e_3] = -e_1$ and $[e_3, e_1] = -e_2$

Let consider $D \in Der_{(1,1,1,\lambda)}\mathfrak{so}(3)$, with $\lambda \in \mathbb{C}$ and let

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

be the matrix associated with the endomorphism D.

Applying equation (1) to the only possible triple (e_1, e_2, e_3), we have

$$
\begin{aligned}
[D[e_1,e_2],[e_1,e_3]]+[[e_1,e_2],D[e_1,e_3]] &= D[[[e_1,e_3],e_2],e_1]+\lambda D[[[e_3,e_1],e_1],e_2] \quad \forall\lambda\in\mathbb{C}.\\
[D[e_1,e_2],[e_1,e_3]] &= -a_{33}e_1+a_{31}e_3,\\
[[e_1,e_2],D[e_1,e_3]] &= -a_{22}e_1+a_{21}e_2,\\
D[[[e_1,e_3],e_2],e_1] &= 0,\\
\lambda D[[[e_3,e_1],e_1],e_2] &= -\lambda a_{11}e_1-\lambda a_{12}e_2-\lambda a_{13}e_3.
\end{aligned}
$$

Therefore, $(-a_{33}-a_{22})e_1+a_{21}e_2+a_{31}e_3=-\lambda a_{11}e_1-\lambda a_{12}e_2-\lambda a_{13}e_3$, which implies $a_{33}+a_{22}=\lambda a_{11}, a_{21}=-\lambda a_{12}$ and $a_{31}=-\lambda a_{13}$. Then, we have

$$\upsilon_{so(3)}(\lambda)=\dim\left(Der_{(1,1,1,\lambda)}\mathfrak{so}(3)\right)=9-3=6,\ \forall\lambda\in\mathbb{C}.$$

5.2 The 3-dimensional Malcev algebra M_3

Let M_3 the Malcev algebra of dimension 3 defined by the law $[e_1,e_2]=e_1+e_3$, $[e_2,e_3]=e_1+e_2+e_3$ and $[e_3,e_1]=-e_1$

Let consider $D\in Der_{(1,1,1,\lambda)}M_3$, with $\lambda\in\mathbb{C}$ and let

$$A=\begin{pmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\\a_{31}&a_{32}&a_{33}\end{pmatrix}$$

be the matrix associated with the endomorphism D.

In this case, similarly to what happened in Subsection 4.1, only one triple (e_1,e_2,e_3) can be taken into consideration. By applying equation (1) we have $[D(e_1+e_3),-e_1]]+[e_1+e_3,D[(-e_1)]=D[[-e_1,e_2],e_1]$. The first member of this expression is: $[D(e_1)+D(e_3),-e_1]]+[e_1+e_3,D(-e_1)]=(-a_{33}-a_{11}+a_{32}+a_{12})e_1+a_{12}e_2+(a_{32}+a_{12})e_3$ and the second one is: $D[[-e_1,e_2],e_1]=-a_{11}e_1-a_{12}e_2-a_{13}e_3$.

So, we obtain three conditions on the elements of the endomorphism matrix given by $-a_{33}-a_{11}+a_{32}=-a_{11},\ a_{12}=0$ and $a_{32}+a_{12}=-a_{13}$.

Therefore, we have

$$\upsilon_{M_3}(\lambda)=\dim\left(Der_{(1,1,1,\lambda)}M_3\right)=9-3=6,\ \forall\lambda\in\mathbb{C}.$$

5.3 The 3-dimensional abelian algebra $\mathfrak{g}_3$ and the algebra $\mathfrak{g}_{2,1}\oplus\mathfrak{g}_1$

Let g_3 be the 3-dimensional abelian Lie algebra, given by $[e_i,e_j]=0, 1\leq i,j\leq 3$, and $\mathfrak{g}_{2,1}\oplus\mathfrak{g}_1$ be the 3-dimensional algebra defined by the only bracket $[e_1,e_2]=e_2$.

By using the previously procedure with these two algebras, of the same dimension, we observe that no restrictions can be obtained, due to all the terms of equation (1) are identically null, $\forall\lambda\in\mathbb{C}$. So,

$$\upsilon_{g_3}(\lambda)=\upsilon_{\mathfrak{g}_{2,1}\oplus\mathfrak{g}_1}(\lambda)=9.$$

6 Certain conclusions

In this paper, we have introduced a new one-parameter invariant function of algebras, $\upsilon_{\mathfrak{g}}(\lambda) = \dim\left(Der_{(1,1,1,\lambda)}\mathfrak{g}\right)$, and we have obtained its values for different types of algebras. In a certain sense, this function is similar to the previous invariant function $\psi_{\mathfrak{g}}(\lambda) = \dim\left(Der_{(\lambda,1,1)}\mathfrak{g}\right)$, by Hrivnák and Novotný [5], although we think that the new functions has some advantages over this last one.

Indeed, in the first place, the computations by using υ are easier than the ones over ψ. It implies that the economic cost in time and memory when using a computer program will be much lower using υ instead of ψ. And secondly, unlike ψ, the values obtained for υ are independent of α in all cases obtained so far, although it is true that we have only dealt with algebras of lower dimensions and we do not know if it will also happen when the dimension increases.

In the following table, we show the comparative between both functions in the studied cases

Algebra	Function ψ		Function υ	
$\mathfrak{g}_3$	$\psi_{\mathfrak{g}_1}(\alpha) = 9,$	$\forall\alpha \in \mathbb{C}.$	$\upsilon_{\mathfrak{g}_1}(\alpha) = 9,$	$\forall\alpha \in \mathbb{C}.$
$\mathfrak{g}_{2,1} \oplus \mathfrak{g}_1.$	$\psi_{\mathfrak{g}_{2,1}\oplus\mathfrak{g}_1}(\alpha) = 6,$ $\psi_{\mathfrak{g}_{2,1}\oplus\mathfrak{g}_1}(\alpha) = 4,$	for $\alpha = 0,$ otherwise.	$\upsilon_{\mathfrak{g}_{2,1}\oplus\mathfrak{g}_1}(\alpha) = 9,$	$\forall\alpha \in \mathbb{C}.$
$\mathfrak{f}_3:$	$\psi_{\mathfrak{f}_3}(\alpha) = 6,$	$\forall\alpha \in \mathbb{C}.$	$\upsilon_{\mathfrak{f}_3}(\alpha) = 9.$	$\forall\alpha \in \mathbb{C}.$
$\mathfrak{f}_4:$	$\psi_{\mathfrak{f}_4}(\alpha) = 7,$	$\forall\alpha \in \mathbb{C}.$	$\upsilon_{\mathfrak{f}_4}(\alpha) = 15,$	$\forall\alpha \in \mathbb{C}.$
$\mathfrak{f}_5:$	$\psi_{\mathfrak{f}_5}(\alpha) = 9,$	$\forall\alpha \in \mathbb{C}.$	$\upsilon_{\mathfrak{f}_5}(\alpha) = 23,$	$\forall\alpha \in \mathbb{C}.$
$so(3):$	$\psi_{so(3)}(\alpha) = 3,$ $\psi_{so(3)}(\alpha) = 1,$	if $\alpha = -1, 1,$ otherwise.	$\upsilon_{so(3)}(\alpha) = 6,$	$\forall\alpha \in \mathbb{C}.$
$M_3:$	$\psi_{M(3)}(\alpha) = 1,$ $\psi_{M(3)}(\alpha) = 0,$	if $\alpha = -1, 0,$ otherwise.	$\upsilon_{M_3}(\alpha) = 6,$	$\forall\alpha \in \mathbb{C}.$

Therefore, although the authors have only dealt with algebras of lower dimensions, the examples considered make us think that $\eta_{\mathfrak{g}} \geq \psi_{\mathfrak{g}}$, for all algebra $\mathfrak{g}$. At this respect, we would like to tackle the proof of this condition in future work. Indeed, we think that the proof must not be difficult, due to the fact that in the definition of $\upsilon_{\mathfrak{g}}$ appears the same terms as in the definition of $\psi_{\mathfrak{g}}$, although one vector is repeated. Therefore, it must imply that such terms are canceled more easily than the terms in $\psi_{\mathfrak{g}}$,which means that the endomorphism involves fewer constraints. In any case, we propose the following

Conjecture 6.1 *For all algebra* $\mathfrak{g}$, *it is verified*

$$v_{\mathfrak{g}}(\lambda) = \dim\left(Der_{(1,1,1,\lambda)}\mathfrak{g}\right) \geq \dim\left(Der_{(\lambda,1,1)}\mathfrak{g}\right) = \psi_{\mathfrak{g}}(\lambda),\ \forall\lambda \in \mathbb{C}.$$

REFERENCES

[1] Claudio Chamon, Christopher Mudry, "Magnetic translation algebra with or without magnetic field in the continuum or on arbitrary Bravais lattices in any dimension" *Phys. Rev. B* Vol. **86**:195125 (2012), 14 pages.

[2] Escobar, J.M., Núñez, J. and Pérez-Fernández, P., "The invariant two-parameter function $\bar{\psi}$". Submitted.

[3] Falcón O. J., Falcón R.M., Núñez J. "A computational algebraic geometry approach to enumerate Malcev magma algebras over finite fields" *Math. Meth. Appl. Sci.* In press, 2017.

[4] Humphreys, J.E. "Introduction to Lie algebras and representation theory". Springer-Verlag, New York, 1972.

[5] Novotný, P. and Hrivnák, J. "On (α, β, γ)-derivations of Lie algebras and corresponding invariant functions" *Journal of Geometry and Physics* Vol. **58(2)** (2008), 208–217.

[6] A.A. Sagle, "Malcev Algebras" *Trans. Amer. Math. Soc.* Vol. **101** (1961), 426–458.

[7] Vergne, M. "Cohomologie des algèbres de Lie nilpotentes, Application à l'étude de la variété des algebres de Lie nilpotentes" *B*ull. Soc. Math. France Vol. **98** (1970), 81-116.

SYMCOMP 2017
Guimarães, 6-7 April 2017

DEALING WITH FUNCTIONAL COEFFICIENTS WITHIN TAU METHOD

M. S. Trindade[1*], J. M. A. Matos[2] and P. B. Vasconcelos[3]

1: Departamento de Matemática, Faculdade de Ciências da Universidade do Porto.
Porto, Portugal
e-mail: marcelo.trindade@fc.up.pt

2: Instituto Superior de Engenharia do Porto and Centro de Matemática da Universidade do Porto
Porto, Portugal
e-mail: jma@isep.ipp.pt

3: Faculdade de Economia and Centro de Matemática da Universidade do Porto
Porto, Portugal
e-mail: pjv@fep.up.pt

Keywords: Spectral methods, Integro-differential equations, Numerical computation

Abstract. *The tau method is a spectral method originally proposed by Lanczos for the solution of linear differential problems with polynomial coefficients. In this contribution we present three approches to deal with differential equations with non-polinomial functional coefficients: (i) making use of function of matrices, (ii) applying orthogonal interpolation and (iii) solving auxiliary differential problems by tau method itself. These approaches are part of the Tau Toolbox efforts for deploying a numerical library for the solution of integro-differential problems. Numerical experiments illustrate the use of all these polynomial approximations in the context of the tau method.*

1 INTRODUCTION

The tau method, introduced by [10], is a spectral method originally developed to find a polynomial that approximates the solution of an ordinary linear differential equation with polynomial coefficients. In the tau method sense, a polynomial y_n of degree n to approximate the solution y of a differential problem is obtained imposing that y_n solves exactly a perturbed problem. This perturbation is performed adding a polynomial τ_n, which is projected on a basis of orthogonal polynomials to attain good error minimization properties. A linear system of algebraic equations is build from matrices that translate the differential problem together with its boundary conditions into an algebraic problem.

The approximation y_n is computed by solving this linear system equations truncated at the order $n+1$. This truncation is related to the perturbing polynomial τ_n.
Many studies have been developed on applications of the tau method to approximate the solution of ordinary and partial, linear and nonlinear integro-differential and fractional derivatives problems [2, 11, 13, 14, 1, 17]. However, in all these works the tau method is applied only to solve specific problems. This range of applicability of the tau method took advantage of the work of [15] with the introduction of the algebraic formulation for the method.
There is no numerical library to solve general integro-differential problems by the tau method. In order to deliver such a numerical tool, an automatic process to polynomially approach non-polynomial functional coefficients is required.
This paper focus (i) on proposing the tau method for integro-differential problems with non-polynomial functional coefficients, making use of three alternative approaches: function of matrices, orthogonal interpolation and the tau method itself and (ii) on presenting the `Tau Toolbox` – a MATLAB numerical library for the solution of integro-differential problems.

2 PRELIMINARES

Let $\mathcal{Z} = [Z_0,\ Z_1,\ \ldots] \subseteq \mathbb{P}$ be an orthogonal basis for the polynomials space $\mathbb{P}$ of any non-negative integer degree, $\mathcal{X} = [1,\ x,\ x^2,\ \ldots] \in \mathbb{P}$ the power basis for $\mathbb{P}$ and $y = \sum_{i\geq 0} a_i Z_i = \mathcal{Z}\mathsf{a},\ \mathsf{a} = [a_0,\ a_1,\ \ldots]^T$, a formal series. The following result introduces matrices M, N and O that set, respectively, polynomial multiplication, differentiation and integration into algebraic operations.

Lemma 1. *Let* V *be the matrix such that* $\mathcal{Z} = \mathcal{X}\mathsf{V}$ *and* $\mathsf{a} = \mathsf{V}^{-1}\mathsf{a}_{\mathcal{X}}$. *Then* $xy = \mathcal{Z}\mathsf{M}\mathsf{a}$, $\frac{d}{dx}y = \mathcal{Z}\mathsf{N}\mathsf{a}$ *and* $\int y dx = \mathcal{Z}\mathsf{O}\mathsf{a}$ *where*

$$\mathsf{M} = \mathsf{V}^{-1}\mathsf{M}_{\mathcal{X}}\mathsf{V}, \quad \mathsf{M}_{\mathcal{X}} = [\mu_{ij}]_{\infty\times\infty},\ \mu_{i+1,i} = 1,\ i = 1, 2, ...,$$

$$\mathsf{N} = \mathsf{V}^{-1}\mathsf{N}_{\mathcal{X}}\mathsf{V}, \quad \mathsf{N}_{\mathcal{X}} = [\eta_{ij}]_{\infty\times\infty},\ \eta_{i,i+1} = i,\ i = 1, 2, ...$$

and

$$\mathsf{O} = \mathsf{V}^{-1}\mathsf{O}_{\mathcal{X}}\mathsf{V}, \quad \mathsf{O}_{\mathcal{X}} = [\theta_{ij}]_{\infty\times\infty},\ \theta_{i+1,i} = \frac{1}{i},\ i = 1, 2, ...$$

Proof. See [15] for M and N, and [8] for O. □

Using these matrices, a linear ordinary differential and/or integral operator, with polynomial coefficients, can be translated into an algebraic representation.

Proposition 1. *Let* $\mathcal{D}y = \sum_{k=0}^{\nu} p_k \frac{d^k y}{dx^k}$ *be an ordinary linear differential operator and* $\mathcal{S}y = \sum_{\ell=0}^{\gamma} p_\ell \left(\int \cdots \int y dx^\ell\right)$ *a linear integral operator, both acting on* $\mathbb{P}$ *with polynomial*

coefficients, then

$$\mathcal{D}y = \mathcal{Z}\mathsf{Da}, \quad \mathsf{D} = \sum_{k=0}^{\nu} p_k(\mathsf{M})\mathsf{N}^k \quad \textit{and} \quad \mathcal{S}y = \mathcal{Z}\mathsf{Sa}, \quad \mathsf{S} = \sum_{\ell=0}^{\gamma} p_\ell(\mathsf{M})\mathsf{O}^\ell, \tag{1}$$

with $p_r(\mathsf{M}) = \sum_{i=0}^{n_r} p_{r,i}\mathsf{M}^i$, $r = k, \ell$ *and* $n_r \in \mathbb{N}_0$.

Proof. See [21]. □

Furthermore, using M and O, we can translate the Volterra and Fredholm linear integral operators into an algebraic representation.

Proposition 2. *The linear Volterra integral operator* $\mathcal{S}_V y = \int_{x_0}^{x} K(x,t)y(t)dt$ *and the Fredholm integral operator* $\mathcal{S}_F y = \int_a^b K(x,t)y(t)dt$, *with degenerate kernel* $K(x,t) \approx \sum_{i=0}^{n_x}\sum_{j=0}^{n_t} k_{ij}Z_i(x)Z_j(t)$, *have the following algebraic representation*

$$\mathsf{S}_V y = \sum_{i=0}^{n_x}\sum_{j=0}^{n_t} k_{ij}\left(Z_i(\mathsf{M}) - \mathsf{e}_{i+1}\mathcal{Z}\big|_{x=x_0}\right)\mathsf{O}Z_j(\mathsf{M})\mathsf{a} \tag{2}$$

and

$$\mathsf{S}_F y = \sum_{i=0}^{n_x}\sum_{j=0}^{n_t} k_{ij}\mathsf{e}_{i+1}\Big(\mathcal{Z}\big|_{x=b} - \mathcal{Z}\big|_{x=a}\Big)\mathsf{O}Z_j(\mathsf{M})\mathsf{a} \tag{3}$$

where e_i *is the ith column of the identity matrix.*

Proof. See [8] and [21]. □

Now, with the formulations (1), (2) and (3), we can find an approximate solution y_n for the linear integro-differential problem

$$\begin{cases} \mathcal{D}y + \mathcal{S}y + \mathcal{S}_V y + \mathcal{S}_F y = f \\ c_i(y) = s_i,\ i = 1, \ldots, \nu \end{cases},$$

where f is a λth degree polynomial (or a polynomial approximation of a function) and $c_i(y) = s_i$, $i = 1, \ldots, \nu$ the initial/boundary conditions, which leads, using the operational formulation, to an infinite algebraic linear system of equations

$$\begin{bmatrix} \mathsf{C} \\ \mathsf{D} + \mathsf{S} + \mathsf{S}_V + \mathsf{S}_F \end{bmatrix}\mathsf{a} = \begin{bmatrix} \mathsf{s} \\ \mathsf{f} \end{bmatrix}, \tag{4}$$

where $\mathsf{C} = [c_{ij}]_{\nu\times\infty}$, $c_{ij} = c_i(Z_{j-1})$, $i = 1, \ldots, \nu$, $j = 1, 2, \ldots$, $\mathsf{a} = [a_0, a_1, \ldots]^T$ is the vector of coefficients of y in $\mathcal{Z}$, $\mathsf{s} = [s_1, \ldots, s_\nu]^T$ and $\mathsf{f} = [f_0, \ldots, f_\lambda, 0, 0, \ldots]^T$ is the right hand side written as linear combination of $\mathcal{Z}$. Finally, the approximate solution y_n is obtained solving (4) truncated at its first $n+1$ rows and columns, with $n \geq \nu + \lambda$.

3 NON-POLYNOMIAL COEFFICIENTS APPROXIMATION

In this work, we provide alternative approaches to cope with operators $\mathcal{D}$, $\mathcal{S}$, $\mathcal{S}_V$ and $\mathcal{S}_F$, such that

$$\mathcal{D}y := \sum_{k=0}^{\nu} h_k^{(\mathcal{D})} \frac{d^k y}{dx^k} = \mathcal{Z}\mathsf{Da},$$

$$\mathcal{S}y := \sum_{\ell=0}^{\gamma} h_\ell^{(\mathcal{S})} \left(\int \cdots \int y dx^\ell \right) = \mathcal{Z}\mathsf{Sa},$$

$$\mathcal{S}_V y := h^{(\mathcal{S}_V)} \int_{x_0}^{x} K(x,t) y(t) dt = \mathcal{Z}\mathsf{S}_V\mathsf{a},$$

and

$$\mathcal{S}_F y := h^{(\mathcal{S}_F)} \int_{a}^{b} K(x,t) y(t) dt = \mathcal{Z}\mathsf{S}_F\mathsf{a},$$

where $h_k^{(\mathcal{D})}$, $h_l^{(\mathcal{S})}$, $h^{(\mathcal{S}_V)}$ and $h^{(\mathcal{S}_F)}$ are non-polynomial functional coefficients. The aim is to compute matrices D, S, S_V and S_F representing those operators with polynomial approximations for $h^{(*)} = h_k^{(\mathcal{D})}$, $h_l^{(\mathcal{S})}$, $h^{(\mathcal{S}_V)}$ and $h^{(\mathcal{S}_F)}$.
The idea is first to approximate any non-polynomial coefficient $h^{(*)}$ by an n-degree polynomial p_n. Function $h^{(*)}$ is approximated with almost machine accuracy and then the coefficient matrix $h^{(*)}(\mathsf{M}_{\mathcal{Z}})$ is approximated by $p_n(\mathsf{M}_{\mathcal{Z}})$.
In the sequel, we present and propose three different possibilities to tackle these approximations automatically.

3.1 THE USE OF INTERPOLATION

Taking for x_j, $j = 0(1)n$, the Gauss points, $h^{(*)}$ is approximated by the interpolating polynomial $h_I : h_I(x_j) = h^{(*)}(x_j)$, $j = 0(1)n$.
This can be accomplished via the Golub-Welsch algorithm [6]. It is based on the fact that if $\mathcal{Z} = [Z_0,\ Z_1,\ \ldots]$ is a sequence of orthogonal polynomials satisfying the three-term recurrence relation

$$xZ_i = \alpha_i Z_{i+1} + \beta_i Z_i + \gamma_i Z_{i-1}, \quad i \geq 0, \quad Z_{-1} = 0, \quad Z_0 = 1,$$

and if $Z_{n+1}(x_j) = 0$, then

$$\begin{bmatrix} \beta_0 & \alpha_0 & & & & \\ \gamma_1 & \beta_1 & \alpha_1 & & & \\ & \gamma_2 & \beta_2 & \alpha_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \gamma_{n-1} & \beta_{n-1} & \alpha_{n-1} \\ & & & & \gamma_n & \beta_n \end{bmatrix} \begin{bmatrix} Z_0(x_j) \\ Z_1(x_j) \\ Z_2(x_j) \\ \vdots \\ Z_{n-1}(x_j) \\ Z_n(x_j) \end{bmatrix} = x_j \begin{bmatrix} Z_0(x_j) \\ Z_1(x_j) \\ Z_2(x_j) \\ \vdots \\ Z_{n-1}(x_j) \\ Z_n(x_j) \end{bmatrix},$$

or $\mathsf{J}_{\mathcal{Z}}\mathsf{p}_{\mathcal{Z}}(x_j) = x_j\mathsf{p}_{\mathcal{Z}}(x_j)$. Thus the eigenvalues of $\mathsf{J}_{\mathcal{Z}}$ are the sought Gauss points.

Very efficient numerical methods are available to compute the eigenpairs of symmetric tridiagonal matrices. Namely the implict QR method, the bisection method with inverse iteration, the divide-and-conquer method and the recent MRRR [3]. In fact, by a diagonal similarity transformation $\mathsf{J}_{\mathcal{Z}}$ can be reduced to a symmetric tridiagonal matrix

$$\begin{bmatrix} \beta_0 & \rho_0 & & & & \\ \rho_0 & \beta_1 & \rho_1 & & & \\ & \rho_1 & \beta_2 & \rho_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \rho_{n-2} & \beta_{n-1} & \rho_{n-1} \\ & & & & \rho_{n-1} & \beta_n \end{bmatrix},$$

with $\rho_i = \sqrt{\alpha_i \gamma_{i+1}}$
Finally, the linear system of equations representing the interpolation conditions at the computed Gauss points is solved to obtain the interpolation polynomials coefficients:

$$\sum_{i=0}^{n} a_i Z_i(x_j) = h^{(*)}(x_j),\ j = 0(1)n.$$

Regarding the approximation of Fredholm and Volterra kernels $K(x,t)$, two-variable interpolation is required. Let $\mathcal{Z}_x = [Z_{x0},\ Z_{x1},\ \ldots]$ and $\mathcal{Z}_t = [Z_{t0}(t),\ Z_{t1}(t),\ \ldots]$ the two orthogonal polynomial sequences, the interpolation coefficients in $\sum_{i=0}^{n}\sum_{j=0}^{n} a_{ij} Z_{xi}(x) Z_{tj}(t)$ are the solutions of the algebraic linear equations system

$$\sum_{i=0}^{n}\sum_{j=0}^{n} a_{ij} Z_{xi}(x_r) Z_{tj}(t_s) = K(x_r, t_s),\ r, s = 0(1)n,$$

at the Gauss points x_r and t_s

3.2 THE USE OF FUNCTION OF MATRICES

Another possibility to approximate non-polynomial functions by polynomials is via the use of functions of matrices.
Given an $n \times n$ matrix A and a scalar function $F(z)$ defined on the spectrum of A then $F(\mathsf{A})$ is defined by replacing z by A. The Jordan form allows to characterize $F(\mathsf{A})$:

Theorem 1. *Let* $\mathsf{W}^{-1}\mathsf{A}\mathsf{W} = diag(\mathsf{J}_1, \cdots, \mathsf{J}_b)$ *be the Jordan canonical form of* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *with*

$$\mathsf{J}_k = \mathsf{J}_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & \cdots & 0 \\ 0 & \lambda_k & 1 & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & \lambda_k \end{bmatrix} \in \mathbb{C}^{m_k \times m_k}.$$

where W *is non-singular and* $\sum_{i=1}^{b} m_i = n$. *If* F *is analytic on the spectrum of* A,

$$F(\mathsf{A}) = \mathsf{W} diag\left(F(\mathsf{J}_1), \cdots, F(\mathsf{J}_b)\right) \mathsf{W}^{-1},$$

where

$$F(\mathsf{J}_k) = \begin{bmatrix} F(\lambda_k) & F'(\lambda_k) & & \cdots & \frac{F^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & F(\lambda_k) & \ddots & & \vdots \\ & & \ddots & \ddots & \\ \vdots & & & \ddots & F'(\lambda_k) \\ 0 & \cdots & & & F(\lambda_k) \end{bmatrix}.$$

Proof. See [5]. □

In particular if the Jordan blocks are 1×1 then $F(\mathsf{A}) = \mathsf{W} diag\left(F(\lambda_1), \cdots, F(\lambda_n)\right) \mathsf{W}^{-1}$ since $\mathsf{A} = \mathsf{W} diag\left(\lambda_1, \cdots, \lambda_n\right)) \mathsf{W}^{-1}$.
This characterization is however computationally not adequate, unless A is diagonalizable and W well conditioned. To ensure numerical stability the best possible is to apply unitary similarity transformations, where the closest diagonal matrix reachable is (upper) triangular. This leads to the Schur decomposition: $F(\mathsf{A}) = \mathsf{Q}F(\mathsf{T})\mathsf{Q}$, where T is upper triangular and Q unitary. Furthermore, in [16] Parlett provides a recurrence method to compute $F(\mathsf{T})$.
Alternatively there are approximation methods that do not require eigenvalue computations. An usual tool is through the truncation of a Taylor series.

Theorem 2. *If* $F(z)$ *has a power series representation* $F(z) = \sum_{k=0}^{\infty} c_k z^k$ *on an open disk containing the spectrum of* A *then* $F(\mathsf{A}) = \sum_{k=0}^{\infty} c_k \mathsf{A}^k$.

Proof. See [5]. □

For the moment, the `Tau Toolbox` has implemented the following function of matrices

$$\sin(\mathsf{M}_{\mathcal{Z}}) = \sum_{i=0}^{\infty} \frac{(-1)^i \mathsf{M}_{\mathcal{Z}}^{2i+1}}{(2i+1)!}, \quad \sinh(\mathsf{M}_{\mathcal{Z}}) = \sum_{i=0}^{\infty} \frac{\mathsf{M}_{\mathcal{Z}}^{2i+1}}{(2i+1)!},$$

$$\cos(\mathsf{M}_{\mathcal{Z}}) = \sum_{i=0}^{\infty} \frac{(-1)^i \mathsf{M}_{\mathcal{Z}}^{2i}}{(2i)!}, \quad \cosh(\mathsf{M}_{\mathcal{Z}}) = \sum_{i=0}^{\infty} \frac{\mathsf{M}_{\mathcal{Z}}^{2i}}{(2i)!},$$

$$\exp(\mathsf{M}_{\mathcal{Z}}) = \sum_{i=0}^{\infty} \frac{\mathsf{M}_{\mathcal{Z}}^{i}}{i!} \quad \text{and} \quad \log(\mathsf{M}_{\mathcal{Z}}) = \sum_{i=0}^{\infty} \frac{(-1)^i (\mathsf{M}_{\mathcal{Z}} - \mathbf{I})^{i+1}}{i+1}.$$

Note that $h^{(*)}$ ($h_k^{(\mathcal{D})}$, $h_l^{(\mathcal{S})}$, $h^{(\mathcal{S}_V)}$ and $h^{(\mathcal{S}_F)}$) can be a composite of functions. For error bounds on matrix functions approximations via truncated Taylor series can be found in [5, 12, 7].

3.3 THE USE OF THE TAU METHOD

Another possibility to approximate functional coefficients on an integro-differential problem is by using the tau method itself. Whenever a functional is the solution of a differential problem, then the tau method is a powerful method to compute an orthogonal polynomial approximation of this auxiliary problem. This novel approach is explored in this work. The main concern is to build a clever mechanism to reach an approximate solution for these functionals as accurate as necessary.
Table 1 shows the list of the functions implemented in `Tau Toolbox` together with the differential problem that must be tackled by the tau method. The accuracy achieved with the polynomial approximation by the tau method for Chebyshev and Legendre basis, for degrees ranging from 10 to 30 (the interval was set as $[0,\ 5]$), is presented in Table 2.

Table 1: Differential problems and their solutions

functional term	differential problem
$\sin(x)$	$\begin{cases} \frac{d^2}{dx^2}y + y = 0, & x \in]a,\ b[\\ y(a) = \sin(a), & y'(b) = \cos(b) \end{cases}$
$\sinh(x)$	$\begin{cases} \frac{d^2}{dx^2}y - y = 0, & x \in]a,\ b[\\ y(a) = \sinh(a), & y'(b) = \cosh(b) \end{cases}$
$\cos(x)$	$\begin{cases} \frac{d^2}{dx^2}y + y = 0, & x \in]a,\ b[\\ y(a) = \cos(a), & y'(b) = -\sin(b) \end{cases}$
$\cosh(x)$	$\begin{cases} \frac{d^2}{dx^2}y - y = 0, & x \in]a,\ b[\\ y(a) = \cosh(a), & y'(b) = \sinh(b) \end{cases}$
$\exp(x)$	$\begin{cases} \frac{d}{dx}y - y = 0, & x \in]a,\ b[\\ y(a) = \exp(a) \end{cases}$
$\log(x)$	$\begin{cases} x^2\frac{d}{dx}y = x, & x \in]a,\ b[\in \mathbb{R}^*_+ \\ y(a) = \ln(a) \end{cases}$

To automatically try to reach machine precision accuracy on the solution of the polynomial approximation for all functional coefficients, and not trying casuistically increasing values for the degree, an iterative scheme was developed. The idea is to build an initial LU factorization from a moderate size dimension (given degree), followed by an iterative process that increases the factors by adding a l row (or block row) and a u column (or block column). This process has the advantage of estimating the error at each iteration without additional cost.
If the function y to be approximated is a solution of the differential problem

$$\begin{cases} \mathcal{D}y = f, \quad x \in]a,\ b[\\ g_j(y) = \sigma_j, \quad j = 1, \ldots, \nu \end{cases} \tag{5}$$

then the error function $\varepsilon_n = y - y_n$, where y_n is the polynomial approximation, satisfies

$$\begin{cases} \mathcal{D}\varepsilon_n = -\tau_n, \quad x \in]a,\ b[\\ g_j(\varepsilon_n) = 0, \quad j = 1, \ldots, \nu \end{cases} \tag{6}$$

since the differential operator $\mathcal{D}$ and boundary/initial conditions functionals g_j are linear. Solving the linear system associated with (6) for a polynomial degree $m > n$, an approximation ε_m for the error is obtained. If ε_m is not sufficiently small then (5) is solved for an approximation y_m of degree m. This system solution can be obtained cheaply since (5) and (6) share the same coefficients matrix. The process continues until convergence. For details see [20].

Table 2: Errors for some functions approximations with tau method for Chebyshev ($\mathcal{T}$) and Legendre ($\mathcal{P}$).

y	$\max\lvert y - y_{10}\rvert$	$\max\lvert y - y_{20}\rvert$	$\max\lvert y - y_{30}\rvert$
$\sin(x)$	$6,5 \cdot 10^{-7}(\mathcal{T})$	$1,6 \cdot 10^{-15}(\mathcal{T})$	$1,8 \cdot 10^{-15}(\mathcal{T})$
	$3,1 \cdot 10^{-7}(\mathcal{P})$	$1,9 \cdot 10^{-15}(\mathcal{P})$	$1,3 \cdot 10^{-15}(\mathcal{P})$
$\sinh(x)$	$3,7 \cdot 10^{-5}(\mathcal{T})$	$4,2 \cdot 10^{-14}(\mathcal{T})$	$2,8 \cdot 10^{-14}(\mathcal{T})$
	$2,1 \cdot 10^{-5}(\mathcal{P})$	$2,8 \cdot 10^{-14}(\mathcal{P})$	$2,8 \cdot 10^{-14}(\mathcal{P})$
$\cos(x)$	$5,0 \cdot 10^{-6}(\mathcal{T})$	$9,9 \cdot 10^{-16}(\mathcal{T})$	$7,7 \cdot 10^{-16}(\mathcal{T})$
	$1,8 \cdot 10^{-6}(\mathcal{P})$	$1,4 \cdot 10^{-15}(\mathcal{P})$	$1,2 \cdot 10^{-15}(\mathcal{P})$
$\cosh(x)$	$3,7 \cdot 10^{-5}(\mathcal{T})$	$2,8 \cdot 10^{-14}(\mathcal{T})$	$2,8 \cdot 10^{-14}(\mathcal{T})$
	$2,1 \cdot 10^{-5}(\mathcal{P})$	$5,6 \cdot 10^{-14}(\mathcal{P})$	$5,6 \cdot 10^{-14}(\mathcal{P})$
$\exp(x)$	$1,8 \cdot 10^{-4}(\mathcal{T})$	$5,6 \cdot 10^{-14}(\mathcal{T})$	$5,6 \cdot 10^{-14}(\mathcal{T})$
	$1,7 \cdot 10^{-5}(\mathcal{P})$	$8,5 \cdot 10^{-14}(\mathcal{P})$	$8,5 \cdot 10^{-14}(\mathcal{P})$
$\ln(x)$	$5,7 \cdot 10^{-5}(\mathcal{T})$	$1,3 \cdot 10^{-9}(\mathcal{T})$	$5,7 \cdot 10^{-14}(\mathcal{T})$
	$3,4 \cdot 10^{-5}(\mathcal{P})$	$1,0 \cdot 10^{-9}(\mathcal{P})$	$4,6 \cdot 10^{-14}(\mathcal{P})$

4 TAU METHOD ON `Tau Toolbox`

The `Tau Toolbox` is a MATLAB library developed to solve, by the tau method, linear and non-linear systems of integro-differential problems (available for download at http://www.fc.up.pt/tautoolbox). It is based on object-oriented programming, consisting of a set of classes and functions.
The use of the `Tau Toolbox` is simple and all starts by defining the tau objects, specifying an orthogonal polynomial basis, domain of interest and basis dimension:

```
>> [x, y] = tau('LegendreP', [-1 1], 5)

x =                          y =

itau with properties:        dtau with properties:

     basis: 'LegendreP'           basis: 'LegendreP'
    domain: [-1 1]               domain: [-1 1]
         n: 5                         n: 5
       mat: [5x5 double]            mat: [5x5 double]
```

This code creates one independent tau variable x (itau) and one dependent tau variable y (dtau). All polynomial approximations described at section 3 are available at the `Tau Toolbox` :

Table 3: Polynomial approximations available at `Tau Toolbox` .

y	Orthogonal interpolation	Function of matrices	tau approximation
$\sin(x)$	$\mathrm{sini}(x)$	$\mathrm{sinm}(x)$	$\mathrm{sint}(x)$
$\cos(x)$	$\mathrm{cosi}(x)$	$\mathrm{cosm}(x)$	$\mathrm{cost}(x)$
$\sinh(x)$	$\mathrm{sinhi}(x)$	$\mathrm{sinhm}(x)$	$\mathrm{sinht}(x)$
$\cosh(x)$	$\mathrm{coshi}(x)$	$\mathrm{coshm}(x)$	$\mathrm{cosht}(x)$
$\exp(x)$	$\mathrm{expi}(x)$	$\mathrm{expm}(x)$	$\mathrm{expt}(x)$
$\log(x)$	$\mathrm{logi}(x)$	$\mathrm{logm}(x)$	$\mathrm{logt}(x)$
$\mathrm{sqrt}(x)$	$\mathrm{sqrti}(x)$	$\mathrm{sqrtm}(x)$	$\mathrm{sqrtt}(x)$

Remark 1. *All approximations can be applied to a composite function at the independent variable x, $u(x)$: `fun(u(x))` for any `fun` specified at Table 3.*

Remark 2. *By default if a function is specified without suffix, then the tau approximation is used: `fun=funt`. This means that the usual MATLAB functions are overlaped when applied to tau variable x.*

To compute $\cos(x)$ just perform the following command:

```
>> A = cost(x); A.mat

ans =

   0.8415   -0.0000   -0.0620    0.0000    0.0010
  -0.0000    0.7174   -0.0000   -0.0780    0.0000
  -0.3102   -0.0000    0.7555   -0.0000   -0.0892
   0.0000   -0.1820   -0.0000    0.7579   -0.0000
   0.0092    0.0000   -0.1606   -0.0000    0.8783
```

This operation is approaching

$$\cos(x) \approx \sum_{i=0}^{\kappa} f_i P_i(\mathsf{M}_{\mathcal{P}}),$$

where f_i are the coefficients of the approximation provided by the tau method. The number of terms κ is automatically set by the method in order to produce an approximation close to machine precision.
A differential operator with functional coefficients can easily be translated into an algebraic problem. For instance the operator

$$\mathcal{D}y = \cos(x)\frac{d^3}{dx^3}y(x) + \sin(x^2+1)\frac{d^2}{dx^2}y(x) + \sinh(\exp(x))\frac{d}{dx}y(x)$$

can be tackled by

```
>> A = cosm(x)*diff(y, 3)+ sini(x^2+1)*diff(y, 2)+ sinht(expt(x))*diff(y);
>> A.mat

ans =

0    1.8721    5.3509   16.1377   13.3312
0    2.5645    7.5886   23.2472   86.3176
0    1.6435    6.4770    9.1748   48.2374
0    0.8331    3.4020    9.7280    0.2318
0    0.3328    1.0924    4.4984   10.4692
```

In this case, the operator is represented by

$$\sum_{i=0}^{\lambda_1} \frac{(-1)^i \mathsf{M}_{\mathcal{Z}}^{2i}}{(2i)!}\mathsf{N}_{\mathcal{P}}^3 + \sum_{i=0}^{\lambda_2} a_i P_i(\mathsf{M}_{\mathcal{P}}^2 + \mathsf{I})\mathsf{N}_{\mathcal{P}}^2 + \sum_{i=0}^{\lambda_3} b_i P_i\left(\sum_{j=0}^{\lambda_4} c_j P_j(\mathsf{M}_{\mathcal{P}})\right)\mathsf{N}_{\mathcal{P}},$$

where $\mathsf{a} = [a_0, \ldots, a_{\lambda_2}]$ is the vector containing the coefficients of the orthogonal interpolation for $\cos(x)$, and $\mathsf{b} = [b_0, \ldots, b_{\lambda_3}]$ and $\mathsf{c} = [c_0, \ldots, c_{\lambda_4}]$ are the vectors containing, respectively, the coefficients by tau approximations for $\sinh(x)$ and $\exp(x)$ functions.

5 NUMERICAL RESULTS

The following examples are based on integro-differential problems with non-polynomial coefficients.

Example 1. *Consider the system of differential equations*

$$\begin{cases} \cos(x)\dfrac{d}{dx}y_1 - (x^2 - 3x^4)y_2 + 3\dfrac{d}{dx}y_3 = \cos(x)^2 - \cos(\exp(x))(-3x^4 + x^2) - 6\sin(2x) \\ \sin(x)\dfrac{d}{dx}y_1 + \dfrac{d}{dx}y_2 - y_3 = \sin(x)\cos(x) - \cos(2x) - \sin(\exp(x))\exp(x) \\ -y_1 + \exp(x)y_2 + x^3\dfrac{d}{dx}y_3 = \cos(\exp(x))\exp(x) - \sin(x) - 2x^3\sin(2x) \\ y_1(-\pi) = 0,\ y_2(0) = \cos(1),\ y_3(\pi) = 1 \end{cases},$$

with exact solution $y_1 = \sin(x)$, $y_2 = \cos(\exp(x))$ *and* $y_3 = \cos(2x)$, $x \in]-\pi,\ \pi[$.

Code 1. *Example 1 on* `Tau Toolbox` .

```
% Create tau objects.
[x, y] = tau('ChebyshevT', [-pi pi], 101);

% Specify problem, conditions and exact solution (if it is known).
sys = {['cosm(x)*diff(y1)-(x^2-3*x^4)*y2+3*diff(y3) =', ...
        ' cos(x)^2-cos(exp(x))*(- 3*x^4+x^2)-6*sin(2*x)'];
       ['sinm(x)*diff(y1)+diff(y2)-y3 =', ...
        ' sin(x)*cos(x)-cos(2*x)-sin(exp(x))*exp(x)'];
       ['-y1+expm(x)*y2+x^3*diff(y3) =', ...
        ' cos(exp(x))*exp(x) - sin(x) - 2*x^3*sin(2*x)']};
cond = {'y1(-pi)=0';'y2(0)=cos(pi)';'y3(pi)=1'};
es = {'sin(x)';'cos(exp(x))';'cos(2*x)'};

% Solve the problem.
a = tausolver(x, y, sys, cond, 'exact_solution', es, 'spy', 1);
```

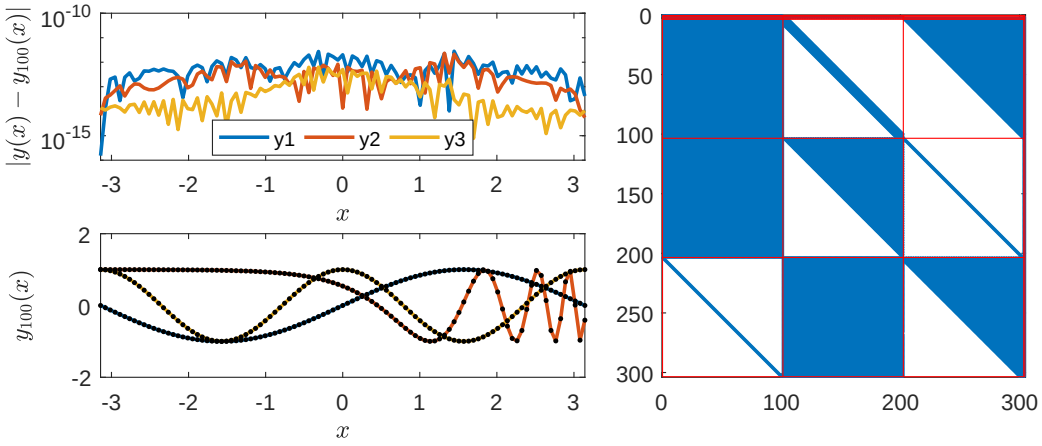


Figure 1: Error and approximate solution (left side) and a spy of the T matrix (right side) for Example 1, using Chebyshev polynomials of the first kind

.

Figure 1 (left) shows the error (bellow 10^{-11}) together with the approximate solution. At the right part of Figure 1, a spy of the coefficient matrix involved on the tau system solver is shown. This 3×3 block matrix has the blocks (1,1), (2,1) and (3,2) completely filled due to the approximation required for the non-polynomial coefficients at the respective terms.

Example 2. *Consider the differential problem*

$$\begin{cases} \lambda \cosh(x)\dfrac{d^2}{dx^2}y + 2x\exp(x)\dfrac{d}{dx}y = \dfrac{2x(\exp(2x)-1)}{\exp\left(\frac{x(\lambda+x)}{\lambda}\right)\sqrt{\lambda\pi}}, \ x \in]-1, \ 1[, \ \lambda \in \mathbb{R}^+ \\ y(-1) = erf\left(-\dfrac{1}{\sqrt{\lambda}}\right), \ y(1) = erf\left(\dfrac{1}{\sqrt{\lambda}}\right) \end{cases},$$

with exact solution

$$y = erf\left(\frac{x}{\sqrt{\lambda}}\right), \ erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp(-w^2)dw.$$

Note that y is a good signal function approximation for λ values sufficiently close to zero:

$$\lim_{\lambda \to 0^+} erf\left(\frac{x}{\sqrt{\lambda}}\right) = sign(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}.$$

This example allows to check the tau approximation behavior at the neighborhood of $x = 0$, for different polynomial degrees and values for λ. Code 2 implements Example 2 on `Tau Toolbox` for $n = 500$ and $\lambda = 10^{-8}$.

Code 2. *Example 2 on `Tau Toolbox` .*

```
% Create tau objects.
[x, y] = tau('ChebyshevT', [-1 1], 501);

% Specify parameter, problem, conditions and exact solution.
lambda = 10^-8;
ode = [num2str(lambda),'*cosh(x)*diff(y, 2)+2*x*exp(x)*diff(y)=', ...
                '(2*x*(exp(2*x)-1))/(exp((x*(',num2str(lambda),'+x))/', ...
                  '(',num2str(lambda),'))*sqrt(',num2str(lambda),'*pi))'];
conditions = {['y(-1)=erf(-1/sqrt(',num2str(lambda),'))']; ...
                                  ['y(1)=erf(1/sqrt(',num2str(lambda),'))']};
solution = ['erf(x/sqrt(',num2str(lambda),'))'];

% Solve the problem.
a = tausolver(x, y, ode, conditions, ...
                                  'exact_solution', solution, 'numstep', 16);
```

Figure 2 shows the approximate solutions and errors for Example 2 with several values for λ and n. The solutions are obtained as p piecewise polynomials.

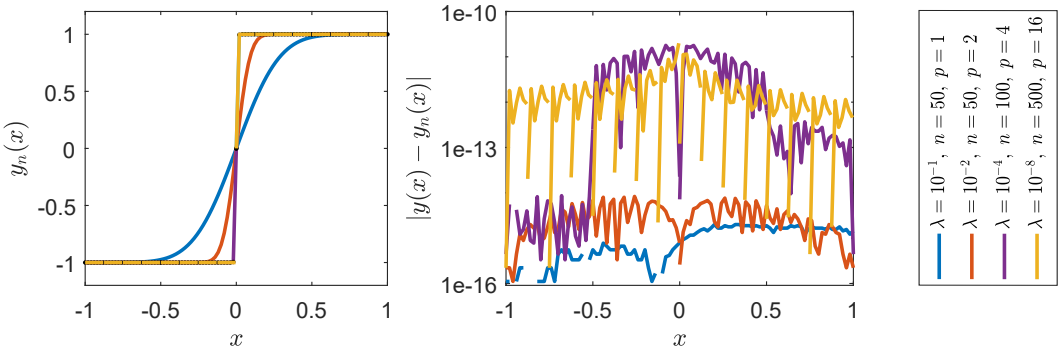


Figure 2: Approximate solutions and errors for Example 2, using the Legendre basis

.

Accurate solutions even for very small values for λ. Note that for decreasing values of λ the domain of integration in $\operatorname{erf}\left(x/\sqrt{\lambda}\right)$ increases and the problem becomes harder to solve.

Example 3. *Consider the Fredholm-Volterra integro-differential problem*

$$\begin{cases} \cos(x)\dfrac{d^2}{dx^2}y - \sin(x^3-x)\dfrac{d}{d}y + \exp(\dfrac{x^2}{2})y - \cosh(x)\displaystyle\int_0^x \sin(x-4t)y(t)dt \\ \qquad\qquad + \sinh(x)\displaystyle\int_0^1 \cos(t^2-x)y(t)dt = \exp(\sin(x^3-x)+x^2) \\ y(0) = 1,\ y(1) = 0 \end{cases}$$

Code 3. *Example 3 on* `Tau Toolbox` .

```
% Create tau objects.
[x, y] = tau('ChebyshevU', [0 1], 50);

% Solve the problem.
a = tausolver(x, y, ['cos(x)*diff(y, 2)','-sin(x^3-x)*diff(y)', ...
                '+exp(0.5*x^2)*y-cosh(x)*volt(y,''sin(x-4*t)'')', ...
                'sinh(x)*fred(y,''cos(t^2-x)'')=exp(sin(x^3-x)+x^2)'], ...
                {'y(0)=1';'y(1)=0'});
```

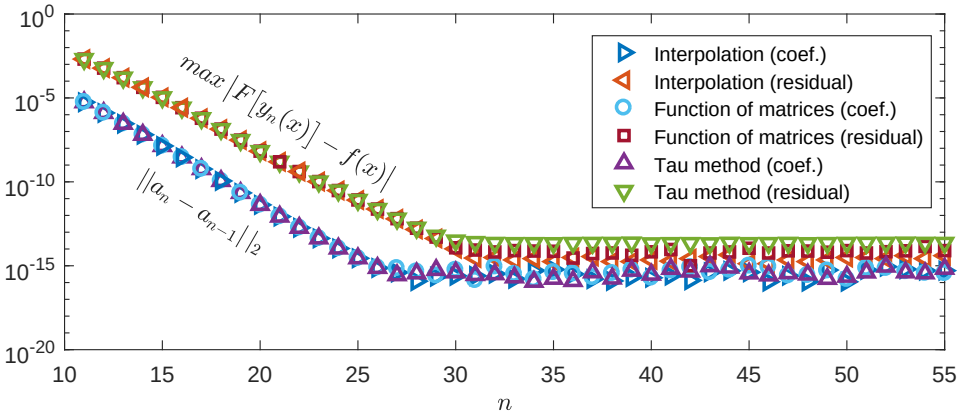


Figure 3: Maximum of the residuals and norms of differences between consecutive coefficients approximations for the Example 3, using the Chebyshev basis of second kind.

Figure 3 illustrates that for all three approaches, the accuracy increases with the degree n, reaching machine precision for n below 30.

6 CONCLUSIONS

Three polynomial approximations were proposed for non-polynomial functional coefficients in the context of the Lanczos tau method. Along with the use of polynomial

interpolation and functions of matrices, the tau method itself was explored to this purpose. The accuracy attained by all these approximations is able to introduce error at the machine precision. This feature extends the applicatibility of the tau method and it is a crucial enhancement to propose a numerical tool to solve integro-differential problems. Furthermore, an automatic mechanism to provide those approximations within the tau method, as an auxiliary problem, prescribing the accuracy required is presented.
The `Tau Toolbox` is a MATLAB numerical library for the solution of integro-differential problems using the Lanczos tau method. Numerical experiments illustrate the use of all these polynomial approximations.

REFERENCES

[1] S. Abbasbandy and A. Taati. Numerical solution of the system of nonlinear volterra integro-differential equations with nonlinear differential part by the operational tau method and error estimation. *Journal of Computational and Applied Mathematics*, 231(1):106 – 113, 2009.

[2] M. R. Crisci and E. Russo. An extension of ortiz' recursive formulation of the tau method to certain linear systems of ordinary differential equations. *Mathematics of Computation*, 41(163):27–42, 1983.

[3] J.W. Demmel, *Applied Numerical Linear Algebra*, SIAM, 1997

[4] A. Gil, J. Segura and N. M. Temme, *Numerical methods for special functions*, SIAM, 2007

[5] G. Golub and C. Van Loan, *Matrix computations*, JHU Press, 2012

[6] G. H. Golub and J. H. Welsch, "Calculation of Gauss quadrature rules", *Math. Comput.*, **23(106)** pp. 221-230, 1969

[7] N. J. Higham: *Functions of matrices: theory and computation*, SIAM, 2008

[8] S. M. Hosseini, and S. Shahmorad, "Numerical solution of a class of integro-differential equations by the Tau method with an error estimation", *Appl. Math. Comput.*, **136**, pp. 559-570, 2003

[9] K. Ito, B. Jin and T. Takeuchi, On a Legendre Tau method for fractional boundary value problems with a Caputo derivative. *Fractional Calculus and Applied Analysis*, 19(2):357–378, 2016.

[10] C. Lanczos, "Trigonometric interpolation of empirical and analytical functions", *J. Math. Phys.*, **17(3)**, pp. 123-199, 1938

[11] K. Liu and C. Pan. The automatic solution to systems of ordinary differential equations by the tau method. *Computers & Mathematics with Applications*, 38(9–10):197 – 210, 1999.

[12] R. Mathias, "Approximation of matrix-valued functions", *J. Matrix Anal. Appl.*, SIAM **14(4)**, pp. 1061-1063, 1993

[13] J. Matos, M. J. Rodrigues, and P. B. Vasconcelos. New implementation of the tau method for {PDEs}. *Journal of Computational and Applied Mathematics*, 164–165:555 – 567, 2004.

[14] E. L. Ortiz and A. P. N. Dinh. Linear recursive schemes associated with some nonlinear partial differential equations in one dimension and the tau method. *SIAM Journal on Mathematical Analysis*, 18(2):452–464, 1987.

[15] E. L. Ortiz and H. Samara, "An operational approach to the tau method for the numerical solution of nonlinear differential equations", *Comput.*, Springer, **1(27)**, pp. 15-25, 1981

[16] B. Parlett, *Computation of Functions of Triangular Matrices*, Memorandum ERL-M481, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, 1974

[17] J. Pour-Mahmoud, M. Y. Rahimi-Ardabili, and S. Shahmorad. Numerical solution of the system of fredholm integro-differential equations by the tau method. *Applied Mathematics and Computation*, 168(1):465–478, Sept. 2005.

[18] L. L. Saeedi, A. Tari, and S. H. M. Masuleh, "Numerical solution of some nonlinear Volterra integral equations of the first kind", *Appl. Appl. Math. - Int. J. (AAM)*, **8(1)**, pp. 214-226, 2013

[19] L. N. Trefethen, *Spectral methods in MATLAB*, SIAM, 2000

[20] M. S. Trindade, J. Matos and P. B. Vasconcelos, "Towards a Lanczos' τ-Method Toolkit for Differential Problems", *Math. Comput. Sci.*, Springer, **10(3)**, pp. 313-329, 2016

[21] P. B. Vasconcelos, J. Matos and M. S. Trindade, "Spectral Lanczos' tau method for systems of nonlinear integro-differential equations", *Integral Methods on Science and Engineering*, Springer, forthcoming

STRENGTH PREDICTION OF ADHESIVELY-BONDED JOINTS WITH CZM LAWS ESTIMATED BY DIGITAL IMAGE CORRELATION

Ulisses Carvalho [1], Raul Campilho [1*]

1: Instituto Superior de Engenharia do Porto
Porto, Portugal
e-mail: ulissescarvalho88@gmail.com , raulcampilho@gmail.com

Keywords: Finite Element Method, Cohesive Zone Models, J-integral, Direct method, Structural adhesive

Abstract *Modern and competitive structures are sought to be strong, reliable and lightweight, which increased the industrial and research interest in adhesive bonding, namely improving the materials strength and fracture properties. With this joining technique, design can be oriented towards lighter structures, not only regarding the direct weight saving advantages of the joint over fastened or welded joints, but also because of flexibility to joint different materials. In any field of industry, the large-scale application of a given joint technique supposes that reliable tools for design and failure prediction are available. Within this scope, Cohesive Zone Models (CZM) integrated with the Finite Element technique are a powerful tool, combining continuum mechanics concepts to promote damage initiation in structures, together with fracture mechanics-based criteria to propagate damage. For damage modelling under generic mixed-mode conditions, the CZM laws of the adhesive bond in tension and shear are required as input in the models. This work experimentally evaluates by the J-integral/direct method the tensile and shear CZM laws of three adhesives with distinct ductility. The experimental work consists of the tensile and shear fracture characterization of the bond by the J-integral technique, providing the estimations of the tensile fracture toughness (GIC) and shear fracture toughness (GIIC). Additionally, by the direct method, the precise shape of the cohesive law in tension and shear of the adhesives is defined. The Double-Cantilever Beam (DCB) specimen was considered to obtain the tensile CZM law of the adhesives. With this purpose, a digital image correlation method was used for the evaluation of the adhesive layer's tensile displacement at the crack tip (delta_n) and adherends rotation at the crack tip (theta_o) during the test, coupled to a Matlab® sub-routine for extraction of this parameter automatically. On the other hand, the End-Notched Flexure (ENF) test was considered the CZM laws in shear. The same method and Matlab® sub-routine were used to measure the adhesive layer's shear displacement at the crack tip (delta_s). After obtaining the tensile and shear CZM laws, triangular, exponential and trapezoidal CZM laws were built to reproduce their behaviour. Validation of these CZM laws was undertaken with a mixed-mode geometry (single-lap joint) considering the same three adhesives and varying overlap lengths. The strength prediction by this technique revealed accurate predictions for CZM law shapes depending of the adhesive ductility, showing that this technique can be valuable in predicting the behaviour of bonded joints.*

EVALUATION OF THE EXTENDED FINITE ELEMENT METHOD TO PREDICT THE MECHANICAL BEHAVIOUR OF ADHESIVELY-BONDED JOINTS

João Xará [1], Raul Campilho [1*]

1: Instituto Superior de Engenharia do Porto
Porto, Portugal
e-mail: 1110236@isep.ipp.pt , raulcampilho@gmail.com

Keywords: Carbon-epoxy, Single-L Joint, eXtended Finite Element Method, Structural adhesive

Abstract *Adhesive bonding is a permanent joining process between the components of a structure that uses an adhesive to bond the components after solidification/curing. This bonding process is used to manufacture complex-shaped structures, which could not be manufactured in a single piece, in order to provide a structural joint that theoretically should be at least as resistant as the base material. Concurrently to the evolution of adhesive joints' technology, synthetic fibres-reinforced composite materials are becoming increasingly popular in many applications, as a result of some competitive advantages over conventional materials. Composite materials are typically used in structures that require high specific strength and stiffness, thereby reducing the weight of the components, while keeping the necessary strength and stiffness to support the imposed loads. With regard to the manufacture of composite structures, although the manufacturing methods allow reducing to the maximum the connections by means of advanced manufacturing techniques, it is still necessary to use joints due to the typical size of the components and design, technological and logistical limitations. On the other hand, it is known that in many high performance structures, it is necessary join components in composite materials with other light metals such as aluminium, for the purpose of structure optimization. This work is an experimental and numerical analysis of single-L adhesive joints between aluminium components and carbon-epoxy composites subjected to peeling loads, considering different adhesives and varying geometrical configurations This study aims to a detailed understanding of these joints' behaviour in which regards the strength and failure modes, by proposing the optimal joint geometry and adhesive type. Numerically, the Finite Element Method is employed to perform a detailed stress analysis that justifies the different failure and strength behaviour between joint configurations. For strength prediction, the Extended Finite Element Method (XFEM) is used, considering different criteria for damage initiation and softening laws. This analysis enables assessing the predictive capabilities of this numerical technique as a design tool for composite bonded joints. The experimental tests validated the numerical results and provided design guidelines for single-L joints. It was shown that the L-part geometry and adhesive type highly influence the joints strength.*

SYMCOMP 2017
Guimarães, 6-7 April 2017

NEURAL NETWORKS BASED ON RADIAL BASIS FUNCTION IN HIGH DIMENSION DOMAINS

Antonio J. Tallón-Ballesteros[1]*, María Rodríguez-Romero[2] and Luís Correia[3]

1: Department of Languages and Computer Systems
Higher Technical School of Computer Science Engineering
University of Seville (Spain)
Reina Mercedes Av. 41012-Seville (Spain)
e-mail: atallon@us.es, web: http://www.lsi.us.es

2: Higher Technical School of Computer Science Engineering
University of Seville (Spain)
Reina Mercedes Av. 41012-Seville (Spain)
e-mail: marrodrom8@alum.us.es, web: http://www.eii.us.es

3: Department of Computer Science
University of Lisbon.
Campo Grande. Lisbon. 1749-016 Lisbon (Portugal).
e-mail: Luis.Correia@ciencias.ulisboa.pt, web: http://ciencias.ulisboa.pt/pt/di

Keywords: Classification, Machine Learning, Data preparation, High dimensionality

Abstract *This paper focuses on radial basis function neural networks to cope with the matter of effectively predicting the right class of unknown samples in the context of multi-class supervised machine learning problems. One of the main drawbacks of this kind of networks is the poor performance that is achieved as the number of dimensions grows. Basically, the idea emerges from the situation that is produced by the fact that the samples are far from the centre of the cluster. As an eventual breakthrough, some guidelines are proposed to be applied in the scope of high dimension domains. The analysis is supported by an empirical study in problems with more features than samples.*

1. INTRODUCTION

Supervised machine learning goals to do predictions within a group of categories with unseen data. Roughly speaking, there is a great deal of approaches such as eager or lazy methods. The taxonomy of algorithms could be divided into several types: neural networks, decision trees, nearest-neighbour and rule-based classifiers [1]. Concerning neural networks, there are two well-known topologies: feed-forward and recurrent neural networks. This paper copes with feed-forward neural networks (F2N2) and more concretely with radial basis function neural networks (RBFN2) in the context of classification problems [2].

This paper aims at shedding light on supervised machine learning problems modelled with radial basis function neural networks that are applied to high dimensionality data sets where the number of features is greater than the number of samples. It is a very empirical paper because neural networks are not widespread with big problems and the computational cost is also very high in terms of memory and processor requirements.

This paper is organized as follows: Section 2 reviews some concepts on artificial neural networks; Section 3 describes briefly the proposal; Section 4 collects the experimental results and the settings in order to make easier the reproducibility of the results; finally, Section 5 summarizes the conclusions that achieve this research.

2. BACKGROUND

It has been theoretically shown that the RBFN2 performance decreases as the dimensionality grow up. Data are thus sparsely distributed in the data space, which actually looks nearly empty, and all distances between two data elements seem roughly equal. Another misleading intuition is the fact that prediction should be easier with more information than with little, and hence prediction should be easier with high-dimensional data than with low-dimensional data. Unfortunately, what is really observed is different: the performances of the data analysis tools decrease as dimensionality grows large. R. Bellman introduced the terms “curse of dimensionality” in [3], to refer to the fact that, in order to optimise a function of d binary variables by exhaustive search over the input domain, one needs to perform 2^d evaluations of the function. This of course becomes quickly computationally intractable.

In RBFN2 the activation of hidden neurons are determined by the closeness of inputs to weights. In other words, the closeness is determined by the squared Euclidean of the inputs and the weights (centres) which are divided by the width of the spherical Gaussian. Basically, the determination of the centres and widths could be done by several approaches although the k-means is the most common one.

3. PROPOSAL

Our proposal analyses deeply the setting parameters and tries to give some future directions for the usage of RBFN2 in high dimensionality problems with more than twelve thousands of inputs and a number of classes greater or equal than five. The data preparation step has been the application of a correlation-based filter selection (CFS) method to the training data and finally the reduced number of inputs is above two hundreds.

4. EXPERIMENTAL RESULTS

This section is devoted to the explanation of the experiments that were carried out along with their configurations.

The experimental design follows the hold-out cross validation technique that consists of dividing the data into two sets: a training and a test set. The former is employed to train the neural network and the latter is used to test the training process and to measure neural network generalization capability. In our case, the size of the training set is 3N/4 and that of the test set is approximately N/4, where N is the number of samples in the problem. We have used a stratified holdout where the two sets are stratified so that the class distribution of the samples in each set is approximately the same as in the original data set. The tables with the results report the accuracy and the Cohen's kappa (K) of the test set that is the performance with the unseen data that have been averaged with 30 runs. We recall that we have initially pre-processed the data set with CFS method which has been successfully used in problems with fewer features [4] in the context of neural networks and also in general purpose classifiers [5].

Data set	#Samples	#Attributes	#CFS(Attributes)	Classes
Lung	203	12600	474	5
SALL	327	12558	268	7

Table 1. Summary of the test-bed

The different settings of the models with RBFN2 are rooted on the number of clusters (B) for the K-means algorithm and also the standard deviation of the clusters (W) [6]. Typically, on most of the implementations these parameters are configured with values 2 and 0.10, respectively.

4.1 Lung

Table 2 depicts the test results with different settings for Lung. According to the results the best parameter values are (B,W)=(10,0.1).The number of clusters is set to 10 which means the double of the number of classes of the problem. It should not be taken as a rule of thumb but sometimes more clusters than classes may be required. K is very likely one of the most important measures in the context of multi-class problems. More details about it could be read in [7] .

Table 3 reports the test results obtained by keeping the number of cluster to 2. There are some variations as the distance increases and the behaviour is better. However, the standard deviation is increasing which means that the W could not be increased unlimitedly [8].

Data set	Setting Value		Performance measures			
	B	W	Acc (%)	SD	K	SD
Lung	2	0.1	94.05	0.80	0.871	0.019
	3	0.1	92.03	1.60	0.820	0.039
	4	0.1	90.98	2.12	0.796	0.054
	5	0.1	89.48	3.60	0.755	0.038
	6	0.1	91.11	3.61	0.796	0.099
	7	0.1	92.09	2.61	0.824	0.068
	8	0.1	93.46	2.39	0.856	0.060
	9	0.1	93.86	2.67	0.865	0.066
	10	0.1	95.10	1.81	0.895	0.041
Avg			92.46		0.831	
SD			1.79		0.045	

Table 2. Test results with different values of B for Lung

Data set	Setting Value		Performance measures			
	B	W	Acc (%)	SD	K	SD
Lung	2	0.05	93.73	1.18	0.864	0.028
	2	0.10	92.03	1.60	0.820	0.039
	2	0.15	94.31	0.93	0.877	0.021
	2	0.20	94.84	0.94	0.889	0.021
	2	0.25	94.84	1.07	0.890	0.024
	2	0.30	95.75	1.44	0.911	0.031
Avg			94.25		0.875	
SD			1.28		0.031	

Table 3. Test results with different values of W for Lung

Table 4 summarises the best results for Lung problem. With the information reported in the previous tables is a bit unclear to draw some conclusions so we have added the recall, that is defined as TP/(TP+FN), to untie the situation. The best pair of values for the parameters are (B,W)=(10,0.1) particularly for the higher performance measures such as Accuracy and K. The configuration of the second row is not best because in two labels the recall is around sixty per cent. The first row compared with the third row is more or less similar, however in the latter there are mistakes in more classes.

Data set	Setting Value		Performance measures				Remarks Recall (%)
	B	W	Acc (%)	SD	K	SD	
Lung	10	0.1	95.10	1.81	0.895	0.041	SQ(54.67) SMCL(80.00) AD(99.90)
	2	0.3	95.75	1.44	0.911	0.031	SQ(60.00) SMCL(66.67)
	10	0.3	94.84	1.56	0.891	0.034	SQ(53.33) SMCL(100) NL(93.33) AD(99.24)
Avg			95.23		0.899		
SD			0.47		0.011		

Table 4. Comparison of the best approaches for Lung

4.2 SALL

Table 5 shows the results for SALL problem which mean that the number of clusters does not affect outstandingly this 7-class problem.

Data set	Setting Value		Performance measures			
	B	W	Acc (%)	SD	K	SD
SALL	2	0.1	92.07	2.37	0.902	0.030
	3	0.1	91.10	2.48	0.890	0.031
	4	0.1	90.00	1.85	0.876	0.023
	5	0.1	89.63	1.99	0.871	0.025
	6	0.1	90.04	1.87	0.876	0.023
	7	0.1	90.04	2.32	0.876	0.029
	8	0.1	90.45	2.14	0.882	0.027
	9	0.1	90.61	2.09	0.884	0.026
	10	0.1	91.14	1.91	0.891	0.024
Avg			90.56		0.883	
SD			0.76		0.010	

Table 5. Test results with different values of B for SALL

Table 6 represents the behaviour with a great deal of values for W keeping the number of clusters in 2. Contrary to the previous scenario, the mean performance is slightly better. An increase in the W values contributes to improve the assessment. More measures are not included in the table but we have carried a further analysis and have reached the conclusion that the best configuration is obtained with the value 2 and 0.20, for B and W, respectively.

Data set	Setting Value		Performance measures			
	B	W	Acc (%)	SD	K	SD
SALL	2	0.05	91.87	2.45	0.899	0.03
	2	0.10	92.07	2.37	0.902	0.03
	2	0.15	92.28	2.28	0.905	0.028
	2	0.20	93.37	2.17	0.918	0.027
	2	0.25	93.78	1.79	0.923	0.022
	2	0.30	93.58	1.83	0.921	0.023
Avg			92.82		0.911	
SD			0.84		0.010	

Table 6. Test results with different values of B for SALL

Table 7 outlines the best results for SALL data set. We have picked up the best value of Tables 5 and 6. The best approach is get with (B,W)=(2,0.2). A high number of clusters is not required and at the same not a very wide distance for this problem with seven classes.

Data set	Setting Value		Performance measures				Remarks Recall (%)
	B	W	Acc (%)	SD	K	SD	
SALL	2	0.2	93.37	2.17	0.918	0.027	BCR-ABL(65.00) MLL(97.33) OTH(86.83)
	2	0.1	92.07	2.37	0.902	0.030	BCR-ABL(57.50) MLL(80.00) OTH(87.17)
Avg			92.72		0.910		
SD			0.92		0.011		

Table 7. Comparison of the best approaches for SALL

5. CONCLUSIONS

This paper went in depth radial basis function neural networks in the context of real-world high dimensionality problems. According to the results we cannot generalised that an increase in the number of clusters is convenient in general terms. It is true that for some problems the number of clusters should be higher than the number of classes, but roughly speaking the distance may be around 0.15 but it is not clear that values greater than 0.30 could be convenient especially due to the increase in the standard deviation. For the future, we plan to experiment with more problems and also to provide some plots to try to understand the drawbacks of the curse of dimensionality.

REFERENCES

[1] Duda, R.O., Hart, P.E., Stork, D. *Pattern Classification*, second ed., Wiley, 2001

[2] Lippmann, R.P. Pattern classification using neural networks, *IEEE Communications Magazine* Vol. 27, pp. 47-64, 1989.

[3] Bellman, R. E. *Adaptive control processes: a guided tour*. Princeton university press, 1961.

[4] Tallón-Ballesteros, A.J., Hervás-Martínez, C., Riquelme, J.C., Ruiz, R. "Improving the accuracy of a two-stage algorithm in evolutionary product unit neural networks for classification by means of feature selection". *IWINAC 2011, Proceedings of the 4th International work-conference on the Interplay Between Natural and Artificial Computation*. Lecture Notes in Computer Science (LNCS) 6687 (vol. II). Springer, La Palma - Spain, pp. 381-390, 2011.

[5] Tallón-Ballesteros, A.J., Riquelme, J.C. "Tackling ant colony optimization meta-heuristic as search method in feature subset selection based on correlation or consistency measures". *IDEAL 2014*, *International Conference on Intelligent Data Engineering and Automated Learning*. Lecture Notes in Computer Science (LNCS) 8669. Springer International Publishing, pp. 386-393, 2014.

[6] Howlett, R.J., Jain L.C. *Radial Basis Function Networks 1: Recent Developments in Theory and Applications*, Springer, Heidelberg, Germany, 2001.

[7] Tallón-Ballesteros, A.J., Riquelme, J.C. "Data Mining Methods Applied to a Digital Forensics Task for Supervised Machine Learning". *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications*. Studies in Computational Intelligence (SCI) 555. Springer International Publishing, 413-428, 2014.

[8] Bishop, C.M. *Neural networks for pattern recognition*, Oxford University Press, New York, 1995.

IMPLEMENTING MATHEMATICAL MODELS FOR SINGULAR INTEGRALS

Paula Ventura Martins[1]* and Ana C. Conceição[2]

1: Research Centre of Spatial and Organizational Dynamics (CIEO)
Faculdade de Ciências e Tecnologia
Universidade do Algarve
8005-139 Faro
e-mail: pventura@ualg.pt, web: http://w3.ualg.pt/~pventura/

2: Center for Functional Analysis, Linear Structures and Applications (CEAFEL)
Faculdade de Ciências e Tecnologia
Universidade do Algarve
8005-139 Faro
e-mail: aicdoisg@gmail.com web: http://w3.ualg.pt//~aconcei/

Keywords: Domain-Specific Language, Singular Integrals, Symbolic Computation

Abstract *In recent years, several software applications with extensive capabilities of symbolic computation had been available to the general public. These applications, known as computer algebra systems (CAS), allow to delegate to a computer all, or a significant part, of the symbolic calculations present in many mathematical algorithms. A mathematical model is a description of a system using mathematical concepts and language. Mathematical models are used in natural sciences and engineering, as well as in social sciences. The main goal of this work is to provide a simple and efficient textual language to formalize mathematical models in the domain of singular integrals. In other words, a mathematician may wish to formulate problems using a computer language specialized to this particular application domain. This new language to compute singular integrals (also known as a Domain-Specific Language - DSL) will not have the complexity that is normally found in general-purpose languages like C, Java and Python. With Xtext and Eclipse, the authors will create a syntax highlighting, error checking and auto-completion editor for singular integrals. To show the effectiveness of the proposed language, a test case is presented with the model of the [SInt] algorithm that computes Cauchy type singular integrals.*

1. INTRODUCTION

In recent years, several software applications with extensive capabilities of symbolic computation had been available to the general public. These applications, known as computer algebra systems (CAS), allow to delegate to a computer all, or a significant part, of the symbolic calculations present in many mathematical algorithms. A mathematical model is a description of a system using mathematical concepts and language [1]. Mathematical models are used in natural sciences and engineering, as well as in social sciences.

Modelling languages provide the best approach to represent complex problems for non-programmers since no sophisticated programming skills are required [10]. Most of mathematicians are not programming experts, the lack of knowledge of how to program is a barrier. So, it is important to provide simple and attractive languages to implement their systems without having to program with general and complex programming languages.

There are different kinds of mathematical modelling languages [8] for describing and solving high complexity problems in different mathematics domains. As an example, we mention operational research that originally presented a modelling language to solve combinatorial optimization problems. A Mathematical Programming Language (AMPL) [6] is an example of a modelling language applied in this domain. In order to describe different classes of problems related to singular integrals and considering the lack of modelling languages and related tools, we decided to develop a Domain-Specific Language (DSL) for this particular application domain. In other areas, such as software development and business processes management, is usual to develop suitable languages considering the specificities of a problem. The main purpose of this work is to help specific stakeholders to solve different classes of problems related to the computation of singular integrals.

This paper is organized in the following sections. Section 2 discusses literature on Domain-Specific Modelling. Section 3 presents the proposed DSL to support model definition related to singular integrals problems. Furthermore, section 4 briefly sketches an example showing how to apply the DSL. Finally, section 5 concludes and gives and outlook on future work.

2. RELATED WORK

A Domain-Specific Language (DSL) is a programming language focused on representing problems and the solutions of a particular domain [7]. The most important characteristics are:

- having a well-defined domain;
- having clear notation (constructs must correspond to important domain concepts);
- empowering experts to easily comprehend and specify logic of their applications;
- improving users' productivity and communication among domain experts.

DSLs offer a restricted suite of notations and abstractions. Languages such as C, Java or Python are regarded as general-purpose languages, because their expressive power is not restricted to an application domain. DSLs also offers the opportunity for interdisciplinary activity and can assist in reaching a shared understanding of intuitive or vague notions above general-purpose languages such as UML [13]. Although many DSLs have been developed over the last decades, the systematic study of DSLs is recent. The most important subjects of

current research are: terminology, risks and opportunities, example DSLs, DSL design methodology and DSL implementation strategies. In this section, we will present same example DSLs in the area of mathematics.

AMPL is an algebraic modelling language implemented around 1985 by Fourer, Gay and Kernighan [6]. This language resembles the symbolic algebraic notation used by modellers to describe mathematical programs in a regular and formal language. The initial focus is dedicated to formulate the underlying model and to generate the computational data structures. The main goal is to make the process easier and less error-prone.

Zinc [9] is a high-level modelling language designed to support experimentation with different solving techniques, namely constraint programming; mixed integer programming; and incomplete search using local search methods. It allows specification of models using a natural mathematical-like notation, resulting in a relatively simple and restricted language. It can be extended to different application areas and provides type instantiation checking which allows early detection of errors in models.

General Algebraic Modelling System (GAMS) [2] is a high-level modelling system for mathematical programming and optimization. The GAMS language [2] is formally similar to common programming languages but allows the user to formulate mathematical models in a way that is similar to its mathematical description. So, it can be used not only by programmers but also by domain experts, in this case, mathematicians.

3. A DSL IN THE DOMAIN OF SINGULAR INTEGRALS

Considering that there is no proper modelling approach to represent operator theory problems, in this section we intend to describe a new language related with the problematic of the computation of singular integrals defined in the unit circle.

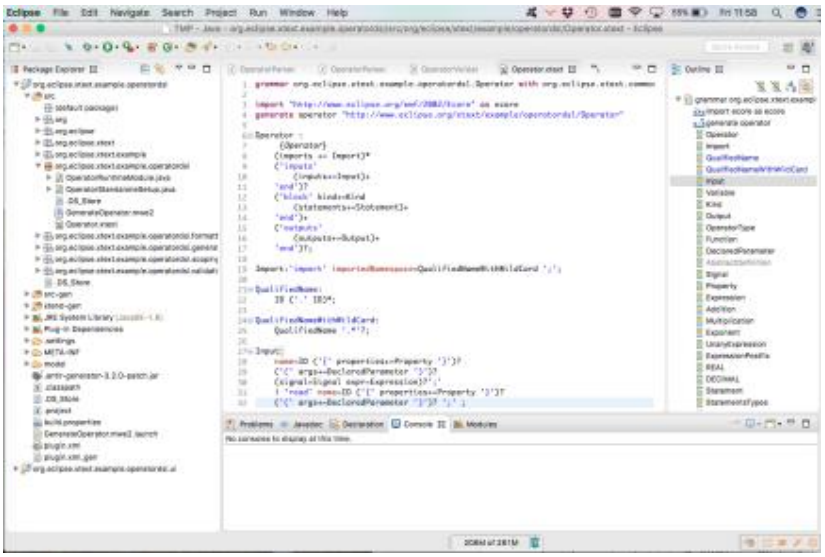


Figure 1. Eclipse framework.

We developed a *Singular Integrals Language* (SIL), the syntax of this DSL includes the concepts: to create lists of equations solutions; to decompose rational fractions; to apply operators to obtained expressions (involving elements of the created lists and elements resulting from the decomposition), among others.
The authors will create, with Eclipse [11] and Xtext [12], a syntax highlighting, error checking and auto-completion editor for singular integrals. In our conceptual model, it is possible to identify different concepts in the domain of singular integrals. Figure 1 illustrates the development environment.

- Rule to represent the different main components of the model

  Figure 2 concerns to the Operator Model that contains some imports and three main components: *(i) Input; (ii) Block of Statements* and *(iii) Output*. Furthermore, the block can be classified in three different types: initial, decompose and calculate. We use the *import* to include another library with grammar definitions of other elements.

```
6  Operator :
7        {Operator}
8       (imports += Import)*
9       ('inputs'
10          (inputs+=Input)+
11      'end')?
12      ('block' kind+=Kind
13          (statements+=Statement)+
14      'end')+
15      ('outputs'
16          (outputs+=Output)+
17      'end')?;
18
```

Figure 2. Main structure of the Operator Model.

- Rule to describe input data elements used in the mathematical model

```
27 Input:
28      name=ID ('{' properties+=Property '}')?
29      ('(' args+=DeclaredParameter ')')?
30      (signal=Signal expr=Expression)?';'
31      | 'read' name=ID ('{' properties+=Property '}')?
32      ('(' args+=DeclaredParameter ')')? ';' ;
```

Figure 3. Partial code of the input rule.

Each *Input* element has an identifier which is initialised with a value or an

expression. This component also includes the request of user inputs.

- Rule to describe a type of block that includes several types of statements

```
99  Statement:
100     '{'
101      (state+=StatementsTypes ';')+
102      '}';
103
104 StatementsTypes:
105     Assigment | Expression | StatementIf | StatementWhile;
```

Figure 4. Rule of the different types of statements.

Each block component will include several types of statements: *assignment*, *expression*, *if* statement and *while* statement. Figure 5 is an example of an expression rule.

```
124 Comparison returns Expression:
125         Equals
126         (({Comparison.left=current} op=("<") ) right=Equals)*;
```

Figure 5. Example of an expression rule.

- Rule to describe the excepted final result

```
40  Output:
41      op=OperatorType fun=Function ';';
42
43  OperatorType:
44      name=ID '{' properties+=Property '}';
45
46  Function:
47      name=ID '(' args+=DeclaredParameter ')';
```

Figure 6. Partial code of the output rule.

The *Output* element format includes an operator and a function. It represents the singular integral that results from the operator applied to the function.

4. EXAMPLE

In this section, we present an example on the subclass of Cauchy type singular integrals that illustrates the application of our DSL-SIL. The [SInt] algorithm [4] computes Cauchy type singular integrals (1), defined on the unit circle Π.

$$S_{\Pi}\varphi(t) = \frac{1}{\pi i}\int_{\Pi} \frac{\varphi(\tau)}{\tau - t} d\tau, \quad \varphi(t) = \mathrm{r}(t)\big(x_{+}(t) + y_{-}(t)\big) \tag{1}$$

Figure 7 illustrates the design of this algorithm developed by a mathematician without applying a specific notation. These elements are not sufficient to describe this class of problems, and at the moment no other satisfactory language has been suggested in the domain of the computation of singular integrals.

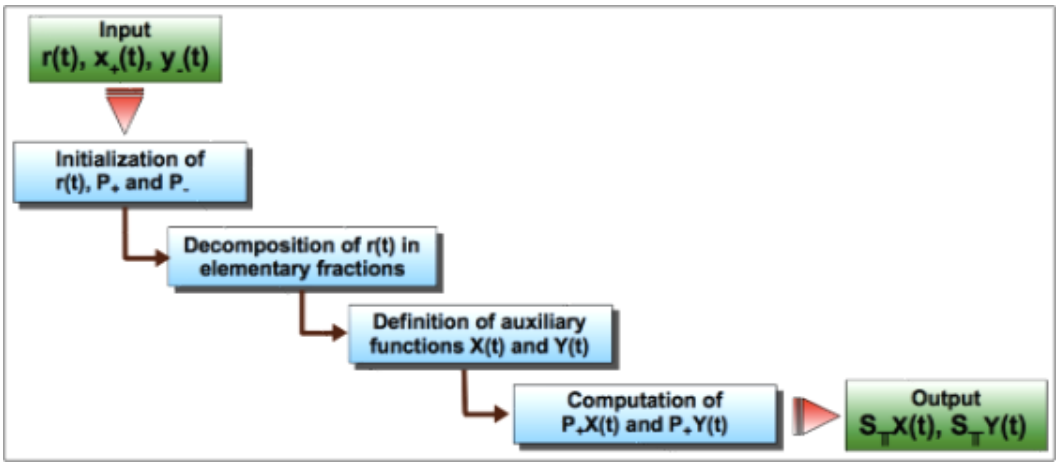


Figure 7. Flowchart of the [SInt] algorithm [4].

As observed in Figure 7, the algorithm presents three different components: *(i) Input; (ii) Block* and *(iii) Output*. This structure represents the basis of our grammar as represented in the previous section. In the following we will describe how to build some components of the [SInt] algorithm using the SIL language. The following segments of the model will be created:

i. The **Input** component of the model allows the user to introduce a particular rational function r and general, or particular, functions x_+ e y_-.

```
inputs
    t;
    read r(t);
    read x{+}(t);
    read y{-}(t);
end
```

Figure 8. Example of an input.

ii. This algorithm includes three different kinds of blocks: initial, decompose and

calculate.

```
block calculate
{
    X(t)=r(t)*x{+}(t);
    Y(t)=r(t)*y{-}(t);
}
end
```

Figure 9. Example of a calculate block.

Changing *r(t)* that had been declared in the **input block** to *v(t)* that wasn't declared. We will receive the notification of an error.

```
block calculate
{
    X(t)=v(t)*x{+}(t);
    Y(t)=r(t)*y{-}(t);
}
end
```

Couldn't resolve reference to AbstractDefinition 'v'.

Figure 10. Example of an error notification.

iii. The **Output** component of the model gives the expected singular integrals.

```
outputs
    S{π}X(t);S{π}Y(t);
end
```

Figure 11. Example of an output.

5. CONCLUSIONS AND FUTURE WORK

The SIL language present several innovative aspects, since it is the first textual language tailored to the computation of singular integrals without implementation details. The possibility to express the algorithm at a higher-level of abstraction is one of the most important features of this DSL. The use of the proposed language allows to implement the model for this class of problems in a user-friendly environment.
Our work will continue in two directions:

- There are other aspects within operator theory related to other algorithms [3][5] that should be considered in the future;
- We intend to work on text-to-text transformations to automatically generate code for the computer algebra system *Mathematica*. We hope that our proposal will allow to translate a mathematical model into a computational structure.

REFERENCES

[1] Bhargava, H.K., ”Computer-aided model construction”, *Decision Support Systems*, Vol. 9(1), pp. 91-111, 1993.

[2] Bussieck, M.R., Meeraus, A., ”General Algebraic Modeling System (GAMS)”, Springer US Vol. 88, pp. 137-157, 2004.

[3] Conceição, A.C., Kravchenko, V.G., ”About explicit factorization of some classes of non-rational matrix functions”, *Mathematisch Nachrichten* , Wiley-Vch Verlag Vol. 280(9-10), pp. 1022-1034, 2007.

[4] Conceição, A.C., Kravchenko, V.G., Pereira, J.C., ”Computing some classes of Cauchy type singular integrals with Mathematica software”, *Advances in Computational Mathematics*, Springer US Vol. 39(2), pp. 273-288, 2013.

[5] Conceição, A.C., Pereira, J.C., ”Exploring the Spectra of Some Classes of Singular Integral Operators with Symbolic Computation”, *Mathematics in Computer Science*, Springer International Publishing Vol. 10(2), pp. 291-309, 2016.

[6] Fourer, R., Gay, D.M., Kernighan, B.W., *AMPL: A Modeling Language for Mathematical Programming*, Cengage Learning, 2002.

[7] Fowler, M., *Domain-Specific Languages*, Addison-Wesley Professional, 2010.

[8] Kallrath, J., *Modeling Languages in Mathematical Optimization*, Springer US, 2004.

[9] Marriott, K., Nethercote, N., Rafeh, R., Stuckey, P.J., Garcia de la Banda, M., Wallace, M., ”The Design of the Zinc Modelling Language”, *Constraints*, vol. 13(3), pp. 229-267, 2008.

[10] Rafeh, R., Marriott, K., Wallace, M., Garcia de la Banda, M., ”Towards the new modelling language Zinc. ”, OSDC2005 *Proceedings of the Open Source Developers' Conference*, Ed. B. Balbo, Melbourne-Australia, pp. 138-142, 2005.

[11] The Eclipse Foudation, ”Eclipse DSL Tools”, Retrieved January 6, 2017, from https://eclipse.org.

[12] The Eclipse Foundation, ”Xtext - The Grammar Language”, Retrieved February 1, 2017, from https://eclipse.org/Xtext/.

[13] Tolvanen, J.-P., ”Industrial Experiences on Using DSLs in Embedded Software Development”, *Proceedings of the Embedded Software Engineering Kongress*, Sindelfingen-Germany, 2011.

SYMBOLIC APPROACH TO THE GENERAL QUADRATIC POLYNOMIAL DECOMPOSITION: ORTHOGONAL AND SYMMETRIC CASES

Ângela Macedo[1], Teresa A. Mesquita[2], Zélia da Rocha[3]

1: Departamento de Matemática & CMAT, Polo CMAT-UTAD, http://www.utad.pt
2: Instituto Politécnico de Viana do Castelo & Centro de Matemática da Universidade do Porto, http://www.estg.ipvc.pt
3: Departamento de Matemática & Centro de Matemática da Universidade do Porto, FCUP , http://www.cmup.pt
e-mails: amacedo@utad.pt ; teresa.mesquita@fc.up.pt ; mrdioh@fc.up.pt

Keywords: Quadratic decomposition, orthogonal polynomials, symmetric polynomials, symbolic computations

Abstract

The quadratic decomposition (QD) of a symmetric orthogonal polynomial sequence into two components was already discussed by L. Carlitz and T. S. Chihara in the second half of the twentieth century. Since then, the type of QD has been enlarged in order to decompose a nonsymmetric sequence and to admit a full quadratic mapping leading to the so called general quadratic decomposition (GQD) previously studied in [1, 3]. In this work we deal with the symbolic approach to the GQD analogously with what was done in [2].

We focus our efforts in expliciting relevant properties of the four component sequences of the GQD like orthogonality and symmetry.

REFERENCES

[1] Â. Macedo and P. Maroni, General quadratic decomposition, J. Differ. Equ. Appl. 16 (11) (2010), p. 1309-1329.

[2] T.A. Mesquita, Z. da Rocha, Symbolic approach to the general cubic decomposition of polynomial sequences. Results for several orthogonal and symmetric cases, Opuscula Math. 32 (4) (2012), 675-687.

[3] T.A. Mesquita, Â. Macedo, Chebyshev polynomials via quadratic and cubic decompositions of the canonical sequence, Integr. Transf. Spec. F., 26 (12) (2015), 956-970.

SYMCOMP 2017
Guimarães, 6-7 April 2017

PAGERANK COMPUTATION WITH MAAOR AND LUMPING METHODS

I.R. Mendes[1*] and P.B. Vasconcelos[2]

1: Instituto Superior Engenharia, Instituto Politécnico do Porto
Porto, Portugal
e-mail: irm@isep.ipp.pt

2: Faculdade de Economia and Centro de Matemática da Universidade do Porto
Porto, Portugal
e-mail: pjv@fep.up.pt

Keywords: Web matrix, linear system computation, Lumping, AOR, GAOR and MAAOR methods

Abstract. *How does the search engine Google determine the order in which to display web pages? The major ingredient in determining this order is the PageRank vector, which assigns a score to every web page. The PageRank vector is the left principal eigenvector of a web matrix that is related to the hyperlink structure of the web, the Google matrix. PageRank is one of the numerical methods Google uses to compute a page's importance and it is at the base of the success of the search engine. This numerical method can be mathematically explored either as an eigenvalue problem or as the solution of a homogeneous linear system. In both cases the Google matrix involved is large and sparse, so tuned algorithms must be developed to tackle it. One of such tunings is the Lumping method approach [19, 13, 18]. Furthermore, the accuracy of the ranking vector needs not to be very precise, so inexpensive iterative methods are preferred. In this work the recent Matrix Analogue of the AOR (MAAOR) iterative method [10], which contains as particular cases the Accelerated Overrelaxation (AOR) [11] and the Generalized AOR (GAOR) [14] stationary family of methods, is explored for the PageRank computation. Additionally Lumping methods have been applied to the eigenproblem formulation and we propose a novel approach combining the Lumping and MAAOR methods for the solution of the linear system. Numerical experiments illustrating the MAAOR method and the MAAOR method combined with the Lumping methods applied to PageRank computations are presented.*

1 INTRODUCTION

PageRank is a crucial numerical algorithm to order the relative importance of web documents based on a web graph (within the World Wide Web). The weight of a page,

PageRank, is computed recursively and depends on the number and PageRank of all incoming links, and therefore, *a web page is important if it is pointed to by other important pages.*
A large number of publications have been produced on this topic. Most of them concerned with numerical approaches to solve the problem as fast as possible. Indeed, the success of Google's lies at PageRank. Nowadays, the PageRank algorithm, originally proposed in [4], by Google founders Larry Page and Sergey Brin, is much more complex, involving other topics such as spammers' control, author rank, among others.
Two basic approaches can be followed to solve the PageRank problem:

(i) as an *eigenvector problem*: $\pi^T = \pi^T G \quad \pi^T e = 1$, or

(ii) as a *linear homogeneous system*: $\pi^T (I - G) = 0^T \quad \pi^T e = 1$

where $\pi^T = (\pi(1), \cdots, \pi(n))^T$ is the *PageRank vector*, $\|\pi\|_1 = 1$, $\pi(i)$ is the PageRank of page i, n is the number of pages in Google's index of the Web and G is the *Google matrix* of order n. In both approaches, the normalization equation $\pi^T e = 1$ ensures that π^T is a probability vector.
The model is founded on matrix

$$S = H + dw^T \tag{1}$$

composed by the matrix H reflecting the link structure of the Web, $h_{i,j} = \frac{1}{n_i}$ if page i links to page j and 0 otherwise, being n_i the number of outlinks of page i, and dw^T to eliminate the zero rows produced by pages with no outlinks, $d_i = 1$ if $n_i = 0$ and 0 otherwise. Vector $w \geq 0$ is known as the *dangling node vector* and it is a probability vector. A dangling node represents a page with no links to other pages (such as image files or protected pages); if a link exists, then the node is referred as *nondangling*.
The uniqueness of π is guaranteed for G irreducible and aperiodic, which leads to the *Google matrix*

$$G = \alpha S + (1 - \alpha) E. \tag{2}$$

with $E = ev^T$ (rank-1 matrix), where e is the vector of all ones, $v \geq 0$ is a probability vector known as *personalization* or *teleportation vector*. It is usual to assume $v = w = \frac{1}{n} e$. Finally, $\alpha \in [0, 1]$ is the *damping factor*, the fraction of time that the random walk follows a link.
Large dimension matrices are involved within PageRank computations, so special care and tuned algorithms must be developed to cope with this problem. One of such numerical techniques are the *Lumping methods* [19, 13, 18] which proceed with a matrix reordering according to dangling and nondangling nodes. These techniques were developed and applied to the eigenvalue formulation and in this paper we will extend them to the linear system formulation of the PageRank problem. Furthermore, a broader class of iterative

stationary numerical methods for the solution of the linear system will be explored in the context of PageRank computations, which is also a novelty.
The rest of the paper is structured as follows. Section 2 presents a stationary method for solving linear systems, the recent MAAOR method. The Lumping methods are briefly described in Section 3. A new hybrid approach that combines lumping techniques with the MAAOR method is presented in section 4. Numerical results illustrating the MAAOR method and the MAAOR method combined with the lumping methods applied to PageRank computations are given in section 5. Finally section 6 presents a few concluding remarks.

2 MAAOR METHODS FOR LINEAR SYSTEMS

In this section we will unveil the importance of sparse computation for the PageRank computations and briefly summarize the MAAOR (Matrix Analogue of the Accelerated Overrelaxation) family of stationary iterative methods for the solution of linear systems of equations.

2.1 Sparse computations

Matrix G is large but not necessarily sparse and yet it is built from the sparse matrix H. Theorem 1 shows that a sparse linear system formulation version exists.

Theorem 1 (Sparse linear system for the PageRank problem)**.** *Solving the sparse linear system*

$$x^T (I - \alpha H) = v^T \tag{3}$$

and letting $\pi^T = \frac{x^T}{x^T e}$ *produces the PageRank vector.*

Proof. If $\pi^T G = \pi^T$ and $\pi^T e = 1$ then π^T is the PageRank vector. Knowing that $\pi^T G = \pi^T \Leftrightarrow \pi^T (I - G) = 0^T \Leftrightarrow x^T (I - G) = 0^T$, we will prove that $x^T (I - G) = 0^T$:

$$\begin{aligned} x^T (I - G) &= x^T \left[I - \alpha H - \alpha d v^T - (1 - \alpha) e v^T \right] \\ &= x^T (I - \alpha H) - x^T \left[\alpha d + (1 - \alpha) e \right] v^T \end{aligned}$$

We have $x^T (\alpha d - (1 - \alpha) e) = 1$ due to the fact that v^T is a probability vector and

$$\begin{aligned} 1 &= v^T e \\ &= x^T (I - \alpha H) e \\ &= x^T e - \alpha x^T H e \quad \text{and} \quad He = e - d \\ &= x^T e - \alpha x^T (e - d) \\ &= x^T e - \alpha x^T e + \alpha x^T d \\ &= (1 - \alpha) x^T e + \alpha x^T d \\ &= x^T \left[(1 - \alpha) e + \alpha d \right]. \end{aligned}$$

So, it results that

$$\begin{aligned} x^T (I - G) &= x^T (I - \alpha H) - x^T [\alpha d + (1 - \alpha) e] v^T \\ &= v^T - v^T \\ &= 0^T. \end{aligned}$$

□

Some properties of the coefficient matrix $(I - \alpha H)$ at Theorem 1 are relevant to highlight.

Proposition 2. *Matrix $I - \alpha H$ has the following properties:*

- *$I - \alpha H$ is nonsingular*
- *$I - \alpha H$ is an M-matrix*
- *$\|I - \alpha H\|_\infty = 1 + \alpha$, provided H is nonzero*
- *The row sums of $I - \alpha H$ are either $1 - \alpha$ for nondangling or 1 for dangling nodes.*
- *$I - \alpha H$ is an M-matrix so $(I - \alpha H)^{-1} \geq 0$*
- *The row sums of $(I - \alpha H)^{-1}$ are equal to 1 for dangling nodes and less than or equal to $\frac{1}{(1-\alpha)}$ for nondangling nodes*
- *The condition number $k_\infty (I - \alpha H) \leq \frac{1+\alpha}{1-\alpha}$*
- *The row of $(I - \alpha H)^{-1}$ corresponding to dangling node i is e_i^T where e_i is the i^{th} column of the identity matrix.*

Proof. See [17, 9, 20]. □

2.2 MAAOR family of methods

For long iterative methods have been investigated, particularly stationary, Jacobi, Gauss-Seidel and SOR [1, 7], nonstationary, such as BiCGSTAB and GMRES, along with the use of preconditioners and reordering [8, 16, 6].
The recent Matrix Analogue of the AOR (MAAOR) iterative method [10], which includes the Accelerated Overrelaxation (AOR) [11] and the Generalized AOR (GAOR) [14] stationary family of methods, is explored for the first time in the context of PageRank computations.
Consider the system

$$Ax = b \tag{4}$$

and the splitting of matrix A

$$A = D - L - U, \tag{5}$$

where $D = diag(A)$, $-L$ and $-U$ represent the diagonal, strictly lower triangular and strictly upper triangular part of A, respectively. Assume $det(D) \neq 0$.
The MAAOR iterative method can be written as

$$x^{(k+1)} = J_{R,W}x^{(k)} + d_{R,W} \quad k = 0, 1, 2, \ldots \tag{6}$$

with

$$J_{R,W} = \left(I - R\tilde{L}\right)^{-1} \left[(I - W) + (W - R)\,\tilde{L} + W\tilde{U}\right]$$

and

$$d_{R,W} = \left(I - R\tilde{L}\right)^{-1} W\tilde{b}$$

where $\tilde{L} = D^{-1}L$, $\tilde{U} = D^{-1}U$, $\tilde{b} = D^{-1}b$, $R \in \mathbb{R}^{n\times n}$ is any diagonal matrix and $W \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $det(W) \neq 0$. It can be shown, with ease, that iteration (6) solves (4).
Considering r and $\omega \neq 0$ two scalars, $R = r\Omega_1$, $W = \omega\Omega_2$, $\Omega_1, \Omega_2 \in \mathbb{R}^{n\times n}$ diagonal matrices with $det(diag(\Omega_2)) \neq 0$, I and O, respectively, the identity and the null matrix well-known iterative methods [12] can be recovered by parameter specification, see Table 1.

Table 1: Iterative methods for specific values of the parameters R and W: SOR = Successive OverRelaxation; EJ = Extrapolated Jacobi; JOR = Jacobi Overrelaxation; EGS = Extrapolated Gauss-Seidel; AOR = Accelerated Overrelaxation; GSOR = Generalized SOR; GAOR = Generalized AOR; MAAOR = Matrix Analogue of the AOR

r	ω	Ω_1	Ω_2	$R = r\Omega_1$	$W = \omega\Omega_2$	iterative method
0	1	I	I	O	I	Jacobi
1	1	I	I	I	I	Gauss-Seidel
ω	ω	I	I	ωI	ωI	SOR
0	ω	I	I	O	ωI	EJ or JOR
1	ω	I	I	I	ωI	EGS
r	ω	I	I	rI	ωI	AOR
1	1	Ω_2	Ω_2	Ω_2	Ω_2	GSOR
r	1	Ω_2	Ω_2	$r\Omega_2$	Ω_2	GAOR
1	1	Ω_1	Ω_2	Ω_1	Ω_2	MAAOR

3 LUMPING METHODS FOR PAGERANK

In [19] lumping methods combined with extrapolation to the eigensystem formulation of PageRank computations were proposed. This allows to accelerate the convergence of the power method. In this work we propose a novel approach incorporating lumping methods into MAAOR methods for the linear system formulation of PageRank.
A brief presentation of lumping methods is now presented.

The basic idea of the Lumping methods is to perform the PageRank computation for the nondangling nodes separately.

Lumping 1 method The dangling nodes can be lumped into a single node to obtain a stochastic reduced matrix with the same eigenvalues as the full matrix. The H matrix can be partitioned into

$$H = \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix}$$

where $H_{11} \geq 0$, $k \times k$, represents the links among the nondangling nodes (ND), and $H_{12} \geq 0$, $k \times (n-k)$, represents the links from nondangling nodes to dangling nodes (D). The $(n-k)$ zero rows in H represent the dangling nodes and the k first rows sums one, see Figure 1.

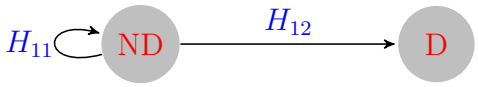


Figure 1: A simple model of the link structure of the Web divided in D and ND nodes.

Then matrix (1) can be casted as $S = \begin{bmatrix} H_{11} & H_{12} \\ ew_1^T & ew_2^T \end{bmatrix}$ where $d = \begin{bmatrix} 0 \\ e \end{bmatrix}$, $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$, w_1 $k \times 1$ and w_2 $(n-k) \times 1$. With this partition we rewrite matrix (2) as $G = \begin{bmatrix} G_{11} & G_{12} \\ eu_1^T & eu_2^T \end{bmatrix}$, with $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$,

$$u_i = \alpha w_i + (1-\alpha)\, v_i, \quad \textit{and} \quad G_{1i} = \alpha H_{1i} + (1-\alpha)\, e v_i^T, \quad i = 1, 2.$$

Lumping can be viewed as a similarity transformation of the Google matrix.

Theorem 3. *Let* $X = \begin{bmatrix} I_k & 0 \\ 0 & L \end{bmatrix}$, *with* $L = I_{n-k} - \frac{1}{n-k}\hat{e}e^T$, $\hat{e} = e - e_1 = [0, 1, 1, \ldots, 1]^T$ *the first canonical basis vector, and* $I_n = [e_1 \cdots e_n]$ *the identity matrix of order* n.
Then, $XGX^{-1} = \begin{bmatrix} G^{(1)} & * \\ 0 & 0 \end{bmatrix}$, *where* $G^{(1)} = \begin{bmatrix} G_{11} & G_{12}e \\ u_1^T & u_2^T e \end{bmatrix}$.
The matrix $G^{(1)}$ *is stochastic of order* $k+1$ *with the same nonzero eigenvalues as* G. *Furthermore, let* $\sigma^T G^{(1)} = \sigma^T$, $\sigma^T \geq 0$, $\|\sigma\| = 1$, *with partition* $\sigma^T = \begin{bmatrix} \sigma_{1:k}^T & \sigma_{k+1} \end{bmatrix}$, *where* σ_{k+1} *is a scalar. Then the PageRank vector* π^T *equals* $\pi^T = \left[\sigma_{1:k}^T \quad \sigma^T \begin{pmatrix} G_{12} \\ u_2^T \end{pmatrix} \right]$.

Proof. See [19] based on [13]. □

Theorem 3 illustrates how the PageRank π^T can be given in terms of σ of the small matrix $G^{(1)}$: Lumping 1 method (it comprises two theorems from [13]).

Lumping 2 method An alternative more complex to this approach considers a refined division of the ND nodes in strongly nondangling nodes (SND), pages with links to pages that are not dangling nodes, and weakly nondangling nodes (WND), pages that are not dangling but that point to only dangling nodes (D). This division leads to a matrix H of the form

$$H = \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} H_{11}^{11} & H_{11}^{12} & H_{12}^{1} \\ 0 & 0 & H_{12}^{2} \\ 0 & 0 & 0 \end{bmatrix}$$

where H_{11}^{11}, $k_1 \times k_1$ represents the links among SND, H_{11}^{12}, $k_1 \times k_2$, the links from SND to WND, H_{12}^{1}, $k_1 \times (n-k)$, the links from SND to D, H_{12}^{2}, $k_2 \times (n-k)$, the links from WND to D, see Figure 2.

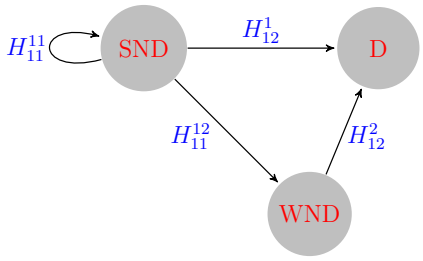


Figure 2: A simple model of the link structure of the Web divided in SND, WND and D nodes.

With the Lumping 1 method the dangling nodes are lumped into a single node resulting in a stochastic reduced matrix with the same eigenvalues as the full matrix. The PageRank computation for the nondangling nodes are performed separately. Following this idea, in [18] it was shown that weakly nondangling nodes can be also lumped into a single node and the PageRank of the strongly nondangling nodes can be computed separately. The further reduced matrix is also stochastic with the same nonzero eigenvalues as the Google matrix G. Moreover, the full PageRank vector π can be easily recovered from the stationary distribution of the smaller matrix.

Theorem 4. *Using the notation above and defining* $G^{(2)}$ *by*

$$G^{(2)} = \begin{bmatrix} G_{11}^{11} & G_{12}^{1} e & G_{11}^{12} e \\ {u_1^{(1)}}^T & u_2^T e & {u_1^{(2)}}^T e \\ (1-\alpha)\, {v_1^{(1)}}^T & \alpha + (1-\alpha)\, v_2^T e & (1-\alpha)\, {v_1^{(2)}}^T e \end{bmatrix},$$

then $G^{(2)}$ *is a stochastic matrix of order* $k_1 + 2$ *with the same nonzero eigenvalues as the full Google matrix* G. *Furthermore, let* $\hat{\sigma}^T = \hat{\sigma}^T G^{(2)}$, $\hat{\sigma} \geq 0$, $\|\hat{\sigma}\| = 1$, *partitioned by* $\hat{\sigma}^T = \begin{bmatrix} \hat{\sigma}_{1:k_1}^T & \hat{\sigma}_{k_1+1} & \hat{\sigma}_{k_1+2} \end{bmatrix}$, *where* $\hat{\sigma}_{k_1+1}$ *and* $\hat{\sigma}_{k_1+2}$ *are two scalars, then the PageRank vector* π *is given by*

$$\pi^T = \begin{bmatrix} \sigma_{1:k}^T & \sigma^T \begin{pmatrix} G_{12} \\ u_2^T \end{pmatrix} \end{bmatrix} \tag{7}$$

where the vector σ is

$$\sigma^T = \begin{bmatrix} \hat{\sigma}^T_{1:k_1} & \hat{\sigma}^T \begin{pmatrix} G^{12}_{11} \\ {u_1^{(2)}}^T \\ (1-\alpha)\, {v_1^{(2)}}^T \end{pmatrix} & \hat{\sigma}_{k_1+1} \end{bmatrix}. \tag{8}$$

Proof. See [19]. □

Theorem 4 shows that to compute the PageRank vector π, we can compute the stationary distribution $\hat{\sigma}$ of the stochastic matrix $G^{(2)}$ and then recover the PageRank vector π according to (7) and (8).

4 LUMPING-MAAOR METHODS FOR PAGERANK: AN HYBRID APPROACH

Based on the work developed by Ipsen and Selee [13] and Lin et al. [18], in [19] we developed new algorithms where the Lumping methods and extrapolation were applied to calculate the PageRank vector as an eigenvector problem. Yu et al. [22] developed two algorithms analogous to the ones by Ipsen and Lin applied to the linear system formulation for the PageRank: Lumping with Jacobi iterations. In this work we propose a new hybrid approach that combines Lumping techniques with the MAAOR method, which generalizes the approach in [22].

4.1 Lumping 1 with MAAOR

Next we will combine the MAAOR method with the Lumping 1 method. We know how to exclude the dangling nodes (D) with their artificial links from the computations, by lumping all the dangling nodes into a single node to obtain a smaller matrix. The PageRank of the nondangling nodes (ND) can be computed separately from that of the dangling nodes using the MAAOR method.
Let us permute the rows and columns of matrix H so that the rows corresponding to dangling nodes are at the bottom of the hyperlink matrix, i.e.,

$$\hat{H} = XHX^T = \begin{array}{c} \\ ND \\ D \end{array} \begin{array}{c} \begin{array}{cc} ND & D \end{array} \\ \begin{bmatrix} H_{11} & H_{12} \\ 0 & O \end{bmatrix} \end{array} \tag{9}$$

where X is a permutation matrix where each row and column has exactly one 1 and all other entries are 0. The H_{11} submatrix represents the links among nondangling nodes (ND), the H_{12} submatrix represents the links from nondangling nodes (ND) to dangling nodes (D) and the zero rows of $\hat{H}$ are associated with the dangling nodes.

It follows from $\pi^T (I - \alpha H) = v^T$ and (9) that

$$\begin{aligned} \pi^T (I - \alpha H) &= v^T \\ \pi^T (I - \alpha H) X^T &= v^T X^T \\ \pi^T X^T X (I - \alpha H) X^T &= v^T X^T \\ \pi^T X^T \left(X I X^T - \alpha X H X^T \right) &= v^T X^T \end{aligned}$$

that is,

$$\hat{\pi}^T \left(I - \alpha \hat{H} \right) = \hat{v}^T \tag{10}$$

where $\hat{\pi}^T = \pi^T X^T$ and $\hat{v}^T = v^T X^T$. Multiplying $\hat{\pi}^T = \pi^T X^T$ by X on the right we obtain the PageRank vector

$$\pi^T = \hat{\pi}^T X. \tag{11}$$

According to (9), $\hat{\pi} = \left[\hat{\pi}_1^T, \hat{\pi}_2^T\right]^T$, $\hat{v} = \left[\hat{v}_1^T, \hat{v}_2^T\right]^T$ and

$$\begin{aligned} \hat{\pi}^T \left(I - \alpha \hat{H} \right) &= \hat{v}^T \\ \left[\hat{\pi}_1^T, \hat{\pi}_2^T\right] \left[\begin{bmatrix} I & O \\ O & I \end{bmatrix} - \alpha \begin{bmatrix} H_{11} & H_{12} \\ O & O \end{bmatrix} \right] &= \left[\hat{v}_1^T, \hat{v}_2^T\right] \\ \left[\hat{\pi}_1^T (I - \alpha H_{11}), -\alpha \hat{\pi}_1^T H_{12} + \hat{\pi}_2^T I\right] &= \left[\hat{v}_1^T, \hat{v}_2^T\right]. \end{aligned}$$

Then,

$$\hat{\pi}_1^T (I - \alpha H_{11}) = \hat{v}_1^T \Leftrightarrow \hat{\pi}_1^T = \hat{v}_1^T (I - \alpha H_{11})^{-1} \tag{12}$$

and

$$-\alpha \hat{\pi}_1^T H_{12} + \hat{\pi}_2^T I = \hat{v}_2^T \Leftrightarrow \hat{\pi}_2^T = \alpha \hat{\pi}_1^T H_{12} + \hat{v}_2^T. \tag{13}$$

Therefore, with the division of the nodes into nondangling nodes and dangling nodes, it is sufficient to apply the MAAOR algorithm on a smaller matrix H_{11} (of the nondangling nodes) to compute the PageRank vector. A new algorithm that combines Lumping 1 with MAAOR method, Lumping1-MAAOR method, is presented in Algorithm 1.

Algorithm 1 Lumping1-MAAOR method

Reorder the hyperlink matrix and vector v to get (9)
function π = LUMPING1-MAAOR(H_{11}, H_{12}, $\hat{v}_1$, $\hat{v}_2$, α, *tol*)
  Solve $\hat{\pi}_1^T (I - \alpha H_{11}) = \hat{v}_1^T$ with the MAAOR method
  Compute $\hat{\pi}_2^T = \alpha \hat{\pi}_1^T H_{12} + \hat{v}_2^T$
  Set $\hat{\pi} = \left[\hat{\pi}_1^T, \hat{\pi}_2^T\right]^T$
  Compute the PageRank vector $\pi^T = \hat{\pi}^T X$
  Normalize $\pi^T = \frac{\pi^T}{\|\pi^T\|_1}$
end function

4.2 Lumping 2 with MAAOR method

Next we combine the MAAOR method with the Lumping 2 method. As mentioned before in the Lumping 2 method all the dangling nodes (D) are lumped into a single node and the same is done with the weakly nondangling nodes (WND). The PageRank of the strongly nondangling nodes (SND) is computed separately.
Again a permutation of the rows and columns of H is required to obtain a new matrix

$$\tilde{H} = YHY^T = \begin{array}{c} \\ SND \\ WND \\ D \end{array} \begin{array}{c} \begin{array}{ccc} SND & WND & D \end{array} \\ \begin{bmatrix} H_{11}^{11} & H_{11}^{12} & H_{12}^{1} \\ O & O & H_{12}^{2} \\ O & O & O \end{bmatrix} \end{array} \tag{14}$$

where submatrix H_{11}^{11} represents the links among SND, H_{11}^{12} the links from SND to WND, H_{12}^{1} the links from SND to D, H_{12}^{2} the links from WND to D and Y is a permutation matrix. According to the partition in (14), $\tilde{\pi} = \left[\tilde{\pi}_1^T, \tilde{\pi}_2^T\right]^T = \left[\tilde{\pi}_1^{(1)^T}, \tilde{\pi}_1^{(2)^T}, \tilde{\pi}_2^T\right]^T$ and $\tilde{v} = \left[\tilde{v}_1^T, \tilde{v}_2^T\right]^T = \left[\tilde{v}_1^{(1)^T}, \tilde{v}_1^{(2)^T}, \tilde{v}_2^T\right]^T$.
From (10) and (14) we obtain

$$\tilde{\pi}^T\left(I - \alpha\tilde{H}\right) = \tilde{v}^T$$

$$\left[\tilde{\pi}_1^{(1)^T}, \tilde{\pi}_1^{(2)^T}, \tilde{\pi}_2^T\right]\left[\begin{bmatrix} I & O & O \\ O & I & O \\ O & O & I \end{bmatrix} - \alpha \begin{bmatrix} H_{11}^{11} & H_{11}^{12} & H_{12}^{1} \\ O & O & H_{12}^{2} \\ O & O & O \end{bmatrix}\right] = \left[\tilde{v}_1^{(1)^T}, \tilde{v}_1^{(2)^T}, \tilde{v}_2^T\right]$$

$$\left[\tilde{\pi}_1^{(1)^T}, \tilde{\pi}_1^{(2)^T}, \tilde{\pi}_2^T\right]\begin{bmatrix} I - \alpha H_{11}^{11} & -\alpha H_{11}^{12} & -\alpha H_{12}^{1} \\ O & I & -\alpha H_{12}^{2} \\ O & O & I \end{bmatrix} = \left[\tilde{v}_1^{(1)^T}, \tilde{v}_1^{(2)^T}, \tilde{v}_2^T\right]$$

$$\left[\tilde{\pi}_1^{(1)^T}\left(I - \alpha H_{11}^{11}\right), -\alpha\tilde{\pi}_1^{(1)^T}H_{11}^{12} + \tilde{\pi}_1^{(2)^T}, -\alpha\tilde{\pi}_1^{(1)^T}H_{12}^{1} - \alpha\tilde{\pi}_1^{(2)^T}H_{12}^{2} + \tilde{\pi}_2^T\right] = \left[\tilde{v}_1^{(1)^T}, \tilde{v}_1^{(2)^T}, \tilde{v}_2^T\right]$$

Then,

$$\tilde{\pi}_1^{(1)^T}\left(I - \alpha H_{11}^{11}\right) = \tilde{v}_1^{(1)^T} \Leftrightarrow \tilde{\pi}_1^{(1)^T} = \tilde{v}_1^{(1)^T}\left(I - \alpha H_{11}^{11}\right)^{-1}, \tag{15}$$

$$\tilde{\pi}_1^{(2)^T} = \alpha\tilde{\pi}_1^{(1)^T}H_{11}^{12} + \tilde{v}_1^{(2)^T}, \tag{16}$$

$$\tilde{\pi}_2^T = \alpha\tilde{\pi}_1^{(1)^T}H_{12}^{1} + \alpha\tilde{\pi}_1^{(2)^T}H_{12}^{2} + \tilde{v}_2^T. \tag{17}$$

With this in mind we will present the following Lumping algorithm, Lumping2-MAAOR algorithm (Algorithm 2) with respect to three type of nodes for computing the PageRank. The MAAOR method will be applied on a even smaller matrix H_{11}^{11} (of the strongly nondangling nodes).

Algorithm 2 Lumping2-MAAOR method

Reorder the hyperlink matrix and vector v to get (14)
function π = LUMPING2-MAAOR(H_{11}^{11}, H_{11}^{12}, H_{12}^{1}, H_{12}^{2}, $\tilde{v}_1^{(1)}$, $\tilde{v}_1^{(2)}$, $\tilde{v}_2$, α, tol)
Solve $\tilde{\pi}_1^{(1)^T}(I - \alpha H_{11}^{11}) = \tilde{v}_1^{(1)^T}$ with the MAAOR method
Compute $\tilde{\pi}_1^{(2)^T} = \alpha\tilde{\pi}_1^{(1)^T} H_{11}^{12} + \tilde{v}_1^{(2)^T}$ and $\tilde{\pi}_2^T = \alpha\tilde{\pi}_1^{(1)^T} H_{12}^{1} + \alpha\tilde{\pi}_1^{(2)^T} H_{12}^{2} + \tilde{v}_2^T$
Set $\tilde{\pi} = \left[\tilde{\pi}_1^{(1)^T}, \tilde{\pi}_1^{(2)^T}, \tilde{\pi}_2^T\right]^T$
Compute the PageRank vector $\pi^T = \tilde{\pi}^T Y$
Normalize $\pi^T = \frac{\pi^T}{\|\pi^T\|_1}$
end function

Table 2: EPA matrix: Number of iterations and convergence time in seconds for MAAOR method, Lumping1-MAAOR and Lumping2-MAAOR methods for $tol < 10^{-8}$ and $\alpha = 0.85$. The symbol – appears whenever the maximum number of iterations was exceeded.

method	MAAOR		Lumping1--MAAOR		Lumping2 -MAAOR	
	it.	time	it.	time	it.	time
Jacobi	698	9.68E-02	723	2.95E-03	1188	9.86E-03
Gauss-Seidel	345	5.21E-02	308	4.98E-03	598	1.71E-02
SOR ($\omega = 0.5$)	685	1.03E-01	665	1.05E-02	1390	3.57E-02
SOR ($\omega = 1.5$)	249	3.91E-02	212	3.58E-03	377	1.16E-02
Extrapolated Jacobi ($\omega = 0.5$)	1120	1.53E-01	1167	1.72E-02	–	–
Extrapolated Gauss-Seidel ($\omega = 1.5$)	294	4.75E-02	266	4.37E-03	478	1.33E-02
Extrapolated Gauss-Seidel ($\omega = 0.5$)	478	7.55E-02	416	6.61E-03	932	2.34E-02
AOR ($\omega = 1.5;\ r = 0.5$)	382	6.08E-02	373	5.96E-03	648	1.70E-02
AOR ($\omega = 0.5;\ r = 1.5$)	380	6.07E-02	305	5.09E-03	673	1.77E-02
AOR ($\omega = 0.5;\ r = 2$)	325	5.24E-02	249	4.09E-03	530	1.45E-02
AOR ($\omega = 0.5;\ r = 5$)	201	3.31E-02	143	2.51E-03	201	7.01E-03
GSOR	345	5.17E-02	308	4.85E-03	598	1.69E-02
GAOR ($r = 0.5$)	464	7.43E-02	452	7.34E-03	839	2.49E-02
GAOR ($r = 1.5$)	287	4.75E-02	240	4.05E-03	457	1.49E-02
GAOR ($r = 0$)	698	9.59E-02	723	1.06E-02	1188	3.06E-02
GAOR ($r = 1$)	345	5.18E-02	308	4.90E-03	598	1.70E-02
GAOR ($r = 3$)	213	3.51E-02	161	2.89E-03	256	9.16E-03
MAAOR ($\omega = 1.5;\ r = 0.5$)	382	6.09E-02	373	5.93E-03	648	1.69E-02
MAAOR ($\omega = 0.5;\ r = 1.5$)	380	6.10E-02	305	4.97E-03	673	1.79E-02
MAAOR ($\omega = 0.8;\ r = 3$)	227	3.70E-02	170	2.89E-03	277	8.63E-03

5 NUMERICAL EXPERIMENTS

To complement the description, this section gives an indication of the computing time (seconds) in typical uses. All the tests have been run on an Intel Core i7-3770 CPU at 3.40 GHz with 4 cores.
Numerical experiments were executed on three matrices:

- EPA matrix – Kleinberg: Pajek network, pages linking to www.epa.gov. A 4772 × 4772 matrix with 8965 nonzeros, 70.18% dangling nodes, 18.76% strong nondangling nodes and 11.06% weakly nondangling nodes.
- wikipedia-20070206 matrix – Gleich: Wikipedia pages at Feb 6, 2007. A 3566907 × 3566907 matrix with 45030389 nonzeros, 2.84% dangling nodes, 88.07% strong nondangling nodes and 9.09% weakly nondangling nodes.
- test matrix – authors: proposed to deliver large sets of pages with equal PageRank. A $10^6 \times 10^6$ matrix with 1410000 nonzeros, 65% dangling nodes, 22% strong nondangling nodes and 13% weakly nondangling nodes.

Table 3: wikipedia matrix: Normalized convergence time for MAAOR method, Lumping1-MAAOR and Lumping2-MAAOR methods for $tol < 10^{-8}$ and $\alpha = 0.85$.

method	MAAOR	Lumping1 -MAAOR	Lumping2 -MAAOR
Jacobi	1.24E+02	1.42E+01	1.36E+01
Gauss-Seidel	2.03E+01	1.29E+01	1.33E+01
SOR ($\omega = 0.5$)	2.27E+01	1.43E+01	1.50E+01
SOR ($\omega = 1.5$)	1.97E+01	1.25E+01	1.27E+01
Extrapolated Jacobi ($\omega = 0.5$)	1.41E+02	8.37E+01	9.69E+01
Extrapolated Gauss-Seidel ($\omega = 1.5$)	2.33E+01	1.52E+01	1.48E+01
Extrapolated Gauss-Seidel ($\omega = 0.5$)	2.35E+01	1.52E+01	1.48E+01
AOR ($\omega = 1.5;\ r = 0.5$)	2.65E+01	1.69E+01	1.67E+01
AOR ($\omega = 0.5;\ r = 1.5$)	2.23E+01	1.46E+01	1.42E+01
AOR ($\omega = 0.5;\ r = 2$)	2.13E+01	1.41E+01	1.35E+01
AOR ($\omega = 0.5;\ r = 5$)	2.15E+01	1.35E+01	1.29E+01
GSOR	2.08E+01	1.32E+01	1.35E+01
GAOR ($r = 0.5$)	2.80E+01	1.86E+01	1.88E+01
GAOR ($r = 1.5$)	2.29E+01	1.46E+01	1.40E+01
GAOR ($r = 0$)	1.25E+02	7.39E+01	8.47E+01
GAOR ($r = 1$)	2.07E+01	1.35E+01	1.36E+01
GAOR ($r = 3$)	2.13E+01	1.42E+01	1.35E+01
MAAOR ($\omega = 1.5;\ r = 0.5$)	2.78E+01	1.87E+01	1.91E+01
MAAOR ($\omega = 0.5;\ r = 1.5$)	2.21E+01	1.48E+01	1.41E+01
MAAOR ($\omega = 0.8;\ r = 3$)	2.11E+01	1.42E+01	1.35E+01

Table 2 reports the number of iterations and time until convergence ($||r|| < 10^{-8}||b||$, r the residual and b the right hand side of the linear system) for the EPA matrix obtained with MAAOR, Lumping1-MAAOR and Lumping2-MAAOR methods, for several MAAOR parameters, for the standard $\alpha = 0.85$. The best results were obtained with AOR with $\omega = 0.5$ and $r = 5$, GAOR with $r = 3$ and MAAOR with $\omega = 0.8$ and $r = 3$. The overall speed of convergence is high due to the tuned sparse matrix operations used. In general, best results were obtained for the hybrid Lumping1-MAAOR both in computation time

(can be up to 10× faster with respect to MAAOR) as well as in the number of iterations to reach convergence. Lumping2-MAAOR for some variants was faster than MAAOR but always requiring more iterations. There are still cases where Lumping2-MAAOR is the worse of the three approaches. This EPA model is small, and the results are certainly very influenced by the distribution of dangling and nondangling nodes (weak and strong). In order to understand if the computation has acceptable costs, we present in Table 3 results with the much larger matrix wikipedia-20070206. Now lumping is always to be preferred to nonlumping, and Lumping2-MAAOR is sometimes faster than Lumping1-MAAOR. This problem is particularly interesting since the percentage of dangling nodes is extremely reduced. The effort in solving the reduced system is thus smaller and the cost to update the solution is higher. Even though, lumping showed it is a crucial strategy to gain performance. As for the EPA matrix, the best results are provided by the same versions of MAAOR, along with Gauss-Seidel and SOR with $\omega = 1.5$.

Table 4: test matrix: Number of iterations and convergence time for MAAOR method, Lumping1-MAAOR and Lumping2-MAAOR methods for $tol < 10^{-8}$ and $\alpha = 0.85$.

method	MAAOR		Lumping1 -MAAOR		Lumping2 -MAAOR	
	it.	time	it.	time	it.	time
Jacobi	31	2.04E+00	28	7.03E-01	31	7.27E-01
Gauss-Seidel	17	1.41E+00	15	4.01E-01	17	3.28E-01
SOR ($\omega = 0.5$)	59	3.74E+00	53	8.45E-01	59	8.79E-01
SOR ($\omega = 1.5$)	37	2.55E+00	39	6.19E-01	38	6.16E-01
Extrapolated Jacobi ($\omega = 0.5$)	70	4.02E+00	62	8.68E-01	71	9.64E-01
Extrapolated Gauss-Seidel ($\omega = 1.5$)	33	2.51E+00	32	5.59E-01	33	5.92E-01
Extrapolated Gauss-Seidel ($\omega = 0.5$)	44	3.11E+00	42	6.92E-01	45	7.56E-01
AOR ($\omega = 1.5;\ r = 0.5$)	608	3.50E+01	667	8.74E+00	687	9.36E+00
AOR ($\omega = 0.5;\ r = 1.5$)	38	2.77E+00	38	6.35E-01	40	6.90E-01
AOR ($\omega = 0.5;\ r = 2$)	48	3.35E+00	45	7.19E-01	46	7.72E-01
AOR ($\omega = 0.5;\ r = 5$)	1400	8.01E+01	302	4.04E+00	1405	1.92E+01
GSOR	40	2.60E+00	35	5.26E-01	38	5.89E-01
GAOR ($r = 0.5$)	46	3.15E+00	40	6.21E-01	45	7.34E-01
GAOR ($r = 1.5$)	32	2.37E+00	29	4.93E-01	31	5.50E-01
GAOR ($r = 0$)	51	2.99E+00	45	6.15E-01	50	7.02E-01
GAOR ($r = 1$)	40	2.61E+00	35	5.31E-01	38	5.88E-01
GAOR ($r = 3$)	36	2.60E+00	31	5.23E-01	33	5.82E-01
MAAOR ($\omega = 1.5;\ r = 0.5$)	41	2.96E+00	42	7.00E-01	46	7.65E-01
MAAOR ($\omega = 0.5;\ r = 1.5$)	73	4.77E+00	67	1.06E+00	71	1.11E+00
MAAOR ($\omega = 0.8;\ r = 3$)	44	3.12E+00	38	6.90E-01	42	7.10E-01

Another feature important to access is to understand how these methods deal well with a web with large sets of pages with equal PageRank. Also, it is known that for increasing values of α the computation of the PageRank computations brings additional difficulties to convergence. For that purpose we use a very large generated matrix to mimic this

behavior. Table 4 shows the number of iterations and time in seconds for the three approaches for $\alpha = 0.85$ and Table 5 for $\alpha = 0.95$.

Table 5: test matrix: Number of iterations and normalized convergence time for MAAOR method, Lumping1-MAAOR and Lumping2-MAAOR methods for $tol < 10^{-8}$ and $\alpha = 0.95$. The symbol – appears whenever the maximum number of iterations was exceeded.

method	MAAOR		Lumping1 -MAAOR		Lumping2 -MAAOR	
	it.	time	it.	time	it.	time
Jacobi	45	2,66E+00	40	7,90E-01	39	9,55E-01
Gauss-Seidel	24	1,76E+00	19	3,41E-01	22	3,79E-01
SOR ($\omega = 0.5$)	80	4,82E+00	61	9,02E-01	65	9,47E-01
SOR ($\omega = 1.5$)	39	2,65E+00	39	6,26E-01	39	6,21E-01
Extrapolated Jacobi ($\omega = 0.5$)	98	5,38E+00	78	1,05E+00	77	1,02E+00
Extrapolated Gauss-Seidel ($\omega = 1.5$)	34	2,53E+00	32	5,59E-01	33	5,90E-01
Extrapolated Gauss-Seidel ($\omega = 0.5$)	57	3,85E+00	47	7,58E-01	52	8,48E-01
AOR ($\omega = 1.5;\ r = 0.5$)	–	–	–	–	–	–
AOR ($\omega = 0.5;\ r = 1.5$)	49	3,38E+00	44	–	48	7,86E-01
AOR ($\omega = 0.5;\ r = 2$)	67	4,41E+00	61	9,33E-01	67	1,04E+00
AOR ($\omega = 0.5;\ r = 5$)	–	–	–	–	–	–
GSOR	60	3,71E+00	53	7,13E-01	57	8,12E-01
GAOR ($r = 0.5$)	69	4,49E+00	61	8,54E-01	65	1,00E+00
GAOR ($r = 1.5$)	49	3,35E+00	44	6,56E-01	47	7,54E-01
GAOR ($r = 0$)	78	4,31E+00	68	8,44E-01	73	9,39E-01
GAOR ($r = 1$)	60	3,70E+00	53	7,15E-01	57	8,06E-01
GAOR ($r = 3$)	47	3,26E+00	40	5,99E-01	47	7,55E-01
MAAOR ($\omega = 1.5;\ r = 0.5$)	45	3,23E+00	44	6,93E-01	50	8,15E-01
MAAOR ($\omega = 0.5;\ r = 1.5$)	107	6,80E+00	96	1,35E+00	103	1,52E+00
MAAOR ($\omega = 0.8;\ r = 3$)	59	4,03E+00	52	7,79E-01	60	9,38E-01

Comparing tables 4 and 5 it can be seen that higher α makes the problem more difficult to solve, but nonetheless, only two of the several variants tested for the three methods failed to convergence (for the maximum number of iterations fixed). For this matrix, since it is sparser than the others, the number of floating-point operations is reduced, in spite of the large dimension. Thus computation times are similiar to the lumping versions and smaller than the non-lumping one.

6 CONCLUSIONS

We have reviewed the MAAOR algorithm, which encloses the most well-known stationary iterative methods, as well as the Lumping methods to segregate dangling and nondangling nodes within the PageRank computations. Hybrid methods combining both were proposed: Lumping1-MAAOR and Lumping2-MAAOR. Numerical experiments illustrate the effectiveness of the two proposed methods. While the Lumping part allows for a clever computation on a smaller portion of the coefficients matrix, MAAOR provides a plethora

of non-expensive numerical iterative linear solvers. This combination has reveled to be well adapted to the problem at hands as well as requiring low computational costs. Quite large problems were solved fast. Future work should explore even more the huge number of possible combinations of the MAAOR parameters, delivering different iterative approaches. Additionally, profiting from all the necessary simulations, get acquainted with additional characteristics of the iteration matrix that influence the selection of the (possibly) best MAAOR parametrization.

REFERENCES

[1] Arasu, A., Novak, J., Tomkins, A.,Tomlin, J. "PageRank computation and the structure of the web: Experiments and algorithms", *Proceedings of the Eleventh International World Wide Web Conference, Poster Track* , pp. 107-117, 2002

[2] Bianchini, M., Gori, M., Scarselli, F. "Inside pagerank", *ACM Transactions on Internet Technology (TOIT)* , ACM Vol. **5(1)** , pp. 92-128, 2005

[3] Brezinski, C., Redivo-Zaglia, M. "The PageRank vector: Properties, computation, approximation, and acceleration", *SIAM Journal on Matrix Analysis and Applications* , SIAM Vol. **28(2)** , pp. 551-575, 2006

[4] Brin, S., Page, L. "Reprint of: The anatomy of a large-scale hypertextual web search engine", *Computer networks* , Elsevier Vol. **56(18)** , pp. 3825-3833, 2012

[5] Cleve, M. "Numerical computing with MATLAB", *SIAM Philadelphia* , 2004

[6] Del Corso, G., Gulli, A., Romani, F. "Exploiting Web matrix permutations to speedup PageRank computation", *Informe técnico* , 2004

[7] Del Corso, G., Gulli, A., Romani, F. "Fast PageRank computation via a sparse linear system", *Internet Mathematics* , Taylor & Francis Vol. **2(3)** , pp. 251-273, 2005

[8] Gleich, D., Zhukov, L., Berkhin, P. "Fast parallel PageRank: A linear system approach", *Yahoo! Research Technical Report YRL-2004-038, available via http://research. yahoo. com/publication/YRL-2004-038. pdf* , Vol. **13** , pp. 22, 2004

[9] Golub, G.,Van Loan, C. "Matrix computations", JHU Press, Vol. **3** , 2012

[10] Hadjidimos, A. "The matrix analogue of the scalar AOR iterative method", *Journal of Computational and Applied Mathematics*, Elsevier Vol. **288** , pp. 366-378, 2015

[11] Hadjidimos, A. "Accelerated overrelaxation method", *Mathematics of Computation* , Elsevier Vol. **32(141)** , pp. 149-157, 1978

[12] Hadjidimos, A., Tzoumas, M. "On the solution of the linear complementarity problem by the generalized accelerated overrelaxation iterative method", *Journal of Optimization Theory and Applications*, Springer Vol. **165(2)** , pp. 545-562, 2015

[13] Ipsen, I.,Selee, T. "PageRank computation, with special attention to dangling nodes", *SIAM Journal on Matrix Analysis and Applications*, SIAM Vol. **29(4)**, pp. 1281-1296, 2007

[14] James, K. "Convergence of matrix iterations subject to diagonal dominance", *SIAM Journal on Numerical Analysis* , SIAM Vol. **10(2)** , pp. 478-484, 1973

[15] Kamvar, S., Haveliwala, T., Manning, C., Golub, G. "Extrapolation methods for accelerating PageRank computations", *Proceedings of the 12th international conference on World Wide Web* , Vol. **28(2)** , pp. 261-270, 2003

[16] Langville, A., Meyer, C. "A reordering for the PageRank problem", *SIAM Journal on Scientific Computing* , SIAM Vol. **27(6)** , pp. 2112-2120, 2006

[17] Langville, A.,Meyer, C. "Google's PageRank and beyond: The science of search engine rankings", Princeton University Press, 2011

[18] Lin, Y.,Shee, X. ,Wei, Y. "On computing PageRank via lumping the Google matrix", *Journal of Computational and Applied Mathematics*, Elsevier Vol. **224(2)**, pp. 702-708, 2009

[19] Mendes, I.,Vasconcelos, P. "Lumping Method with Acceleration for the PageRank Computation", *14th International Conference on Computational Science and Its Applications, Guimares* , pp. 221-224, 2014

[20] Meyer, C. "Matrix analysis and applied linear algebra", Siam, Vol. **2** , 2000

[21] Song, Y. "On the convergence of the generalized AOR method", *Linear algebra and its applications* , Elsevier Vol. **256** , pp. 199-218, 1997

[22] Yu, O., Miao, Z., Wu, G., Wei, Y. "Lumping algorithms for computing Google's PageRank and its derivative, with attention to unreferenced nodes", *Information retrieval* , Springer Vol. **15(6)** , pp. 503-526, 2012

SYMCOMP 2017
Guimarães, 6-7 April 2017

USER-FRIENDLY MATLAB TOOL FOR SIMPLIFIED LIFE CYCLE ASSESSMENT METHOD FOR DOMESTIC BUILDINGS

Carlos Campos[1]*, Fiona Bradley[2] and Graeme Hannah[3]

1: University of Strathclyde, UK, http://www.apluscrc.com
2: University of Glasgow, UK
3: BRE Group, UK

e-mails: carlos.campos@strath.ac.uk ; fiona.bradley@glasgow.ac.uk ; graeme.hannah@bre.co.uk

Keywords: Life Cycle Assessment, LCA, LCA of buildings, LCA tool, Matlab tool

Abstract *When undertaking Life Cycle Assessment (LCA) of Buildings in practice, it can become a major challenge to accurately assess given the range of scenarios, data sources, numerical data formats and calculations that the process requires. Results are usually shown with impact indicators such as Global warming Potential (KgCO2e) or Energy consumption (MJ), as well as other alternatives. LCA of buildings is understood as a complete environmental evaluation of the building process from cradle to grave. This includes for the raw materials that are extracted, the transportation loads, the construction process and the operational use until the building is no longer appropriate be used as for its primary intention, and is therefore demolished and final disposal of construction and demolition waste takes place or materials reach the end-of-waste line where they can be utilized for other applications. Those LCA processes are commonly divided in four stages named from A to D:*

• The Manufacturing (A) stage involves raw material collection, materials manufacturing, construction and transport related impacts.

• The Operational (B) stage is includes for the usage of the building during its life, including sub-stages such as water consumption, building controlled energy demand, replacements and maintenance works.

• The End-of-life (C) stage takes into consideration the impact after the building is no longer fit for its primary purpose, including demolition, waste processing, final disposal and the transportation loads.

• The fourth and last of the stages is Beyond the End-of-life (D) which is the most uncertain phase due to the time difference and industry potential evolution in a period of time which varies from 60 to 100 years in some cases since the assessment is done. This stage includes recycling, re-use, and energy recovery potential.

Current practices, within the LCA of the built environment, often involve the use of complex and heavy software tools, for evaluating (all stages or some of them) the complete process or part of it, capable of taking into account all material quantities, indicators and variables occurring within a building project life span. In addition, the life cycle of a building might be a period over the 100 years and lots of different variable should be considered. It happens that the earlier the engagement with LCA occurs, within the commonly accepted RIBA project timeline, the more possible scenarios should be considered when attempting for the best possible cradle-to-grave building solution.These actual facts make the assessment process difficult to understand for users

not very familiarized with the complete methodology and also difficult to report to the final user that should take into consideration diverse LCA indicators for the optimal performance of the building. The assessment process is understood as complex and time consuming task for which a sustainable consultant is often recommended. However, current software programs available for this matter require a high expertise and experience with that specific tool and its environmental impact method.

This is resulting, within the construction sector, on a poor usage of these tools due to their complexity and a general lack of understanding indicators, methods and reporting information systems. When trying the best sustainable practice in small projects, the best case scenario is that only some stages of the assessment are completed, which is widely called simplified process. This simplification usually consist in taking into account some of the Environmental Product Declarations of materials used within the project or just consider the manufacturing process with the transportation of materials to site included. The other stages, such as construction, operational stage, end-of-life and beyond the-end-of life are commonly avoided. This entails to incomplete life cycle assessment or evaluations with lack of analysis for important stages within the complete life cycle of a building.

Furthermore, comparisons between LCA studies from different software are almost impossible due to the diversity of database and hidden calculations and assumptions behind every stage of calculation developed for the built environment. It makes even difficult to trust, from a final user point of view, if results are given without the subjective opinion from the expert about what the building should achieve instead of based in objective and measurable facts. In order to bridge the gap between the reality and the software tool model a complete and more accurate transparent LCA is needed, letting the user know all assumptions and considerations that the software is about to use when evaluating the performance of that particular project. That needed methodology has to be according to the current LCA and sustainable standards in order to produce and be able to validate the results.

Many authors argue that a simplified life cycle methodology is needed, in order to get reliable assessment results and comparable with others from different software and tools. This simplification would be also beneficial due to the actual complexity of the LCA process and the hard task that experts approach when gathering all data involved in the already complex scenario such as a building project implies. This simplification is needed for a more transparent and bigger impact of decision-making experience when addressing this environmental evaluation at a very early stage of the building process.

This paper presents a new simplification methodology, when undertaking the environmental assessments of domestic projects, related to the building's fabric and operational embodied energy and carbon, having into account all main stages that the LCA involves. This simple methodology is implemented with a user-friendly tool using Matlab software. The proposed impact method is in line with the European Standard EN 15978 under the framework described in EN 15643-2 regarding the Assessment of environmental performance of buildings. The aim of this proposed Matlab tool is to bridge the gap between environmental assessment and reality. It also aims to improve, within that process, the transparency and accessibility to assumptions and information used. All stages of LCA, manufacturing, operational and End-of-life, are included and calculated based on the environmental database and transportation calculations for all materials too. The last and optional stage Beyond the End-of-life is not included due to the difficulty to make a close-to-real assumption. The impact method presented uses an environmental

database produced from the manufacturing embodied energy (MJ) and carbon (KgCO2e) data of typical materials included in domestic building project which is measured by floor area unit (m2). This internal environmental database, from which environmental results are based, is easily updatable and changeable for other specific values with which the user may want to test the project. The final indicators assessed in the tool are not normalized nor characterized in order to avoid the subjective part of these assessments. Hence, the method aims to get only the measurable and objective part of the environmental evaluation.

The tool, in order to produce the environmental assessment, the building measurements and the main construction elements descriptions, including finishes, need to be entered into the tool by the user, avoiding facilities and furniture installations. These are introduced by simplified selection from pre-defined possible solutions already within the tool. Transportation loads are also considered within the tool, from different types of vehicles. Distances are are fixed by default, of all materials, from the manufacturing gate to the construction site and from site to final disposal sites after demolition works take place. After gathering all building model data, the tool matches that data with the corresponding environmental impacts. After that all impacts are added and a sub-total final figure is produced. The results are then expressed by LCA stages and by floor. These results are given in simple graphics in order to get easier for final users to understand.

The tool is also able to compare different options of the same building using different materials for certain construction elements such us different type of finishes or variations with insulation thicknesses in order to compare how those changes affect within the global environmental assessment. The proposed tool will be able to calculate within minutes a complete simplified LCA for new or retrofit domestic projects. It will also could help the important decision-making task when a domestic building is about to be either be demolish or deeply refurbished.

Further research would involve the comparison of the building assessed against a LCA benchmark for domestic buildings. This would help the incorporation of the tool to current certification and quality schemes such as Home Quality Mark, within the simplified route of the assessment future work could include the upgrade of the tool in order to be able to evaluate all types of buildings and use it within BREEAM, BRE Environmental Assessment Methodology. It could also include the interface development and the compilation of the stand-alone and executable format.

SYMCOMP 2017
Guimarães, 6-7 April 2017

SELECTION OF MODELLING PARAMETERS FOR STOCHASTIC MODEL UPDATING

Tiago A. N. Silva[1,2], John E. Mottershead[3]

[1] NOVA UNIDEMI, Faculdade de Cincia e Tecnologia, Universidade Nova de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal.
[2] GI-MOSM, ADEM, ISEL - Grupo de Investigao em Modelao e Optimizao de Sistemas Multifuncionais, Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emdio Navarro, 1, 1959-007 Lisboa, Portugal.
[3] Liverpool Institute for Risk and Uncertainty, University of Liverpool, Liverpool L69 3GH, United Kingdom.
e-mail: tan.silva@fct.unl.pt

Keywords: Sensitivity analysis; Covariance matrix; Parameter selection.

Abstract. *In structural dynamics, the adjustment of a set of modelling parameters based on the minimization of the discrepancy between experimental and model responses is known as model updating. In the context of stochastic model updating, the selection of a set of updating parameters from the modelling ones is very important, both in terms of computational efficiency and of the accuracy of the solution of this stochastic inverse problem. One can find in the literature several approaches to model updating. A simple expression was developed for covariance matrix correction in stochastic model updating and by its use one may observe the relevance of choosing the correct set of updating parameters. One may conclude that if the updating parameters are correctly chosen, then the covariance matrix of the outputs is correctly reconstructed, but when the updating parameters are wrongly chosen is found that the responses covariance matrix is generally not reconstructed accurately, although the reconstructing of the responses mean values is accurate. Hence, the selection of updating parameters is developed by assessing the contribution of each candidate parameter to the responses covariance matrix, thereby enabling the selection of updating parameters to ensure that both the responses mean values and covariance matrix are reconstructed by the updated model. It is shown that the scaled output covariance matrix may be decomposed to allow the contributions of each candidate parameter to be assessed. Numerical examples are given to illustrate this theory.*

1 INTRODUCTION

Assuming that modal testing and analysis have been carefully carried out, measured data is the most reliable information on the dynamic behaviour of a structure and usually it is considered to be correct. Therefore, it is taken as reference or target data. On the other hand, structural modelling based on FEM became a standard on most industrial and research applications. As it is widely known, there are several sources of error at the modelling stage, roughly ranging from modelling assumptions to uncertainty in the modelling parameters [1]. As the initial estimates on the analytical or numerical model are often incorrect, there is a difference between the measured responses and the ones predicted using the model. Hence, it is necessary to reconcile or to ensure the agreement of the two data sets. This task is made possible by correcting the model, also referred as model updating.

Over the years, a huge range of model updating methods has been proposed and implemented in the most diverse fields. The focus of this thesis is concerned with the updating of linear models and this subject has been reviewed in detail by Mottershead and Friswell [2, 3]. In [4], one can find a detailed overview of several implementations of model updating strategies, from direct to iterative methods, including methods based on robust optimization and meta-modelling. A comprehensive review on the use of computational intelligence techniques applied to model updating can be found in [5]. A historical overview of model updating (1960–1990) can be found in [6].

This work presents the fundamentals of the sensitivity-based model updating methods to enhance the discussion on the issue of selecting a consistent parameter set from the considered modelling parameters, as it is crucial to update the covariance matrix.

2 SENSITIVITY-BASED MODEL UPDATING

The sensitivity-based model updating method was reviewed by [7]. Generally, this kind of methods consider that the experimental or reference model can be cast as a perturbation about a theoretical model approximation. By this, the error between the experimental and predicted responses is

$$\boldsymbol{\varepsilon}_{\boldsymbol{z}_j} = \boldsymbol{z}_m - \boldsymbol{z}(\boldsymbol{\theta}_j) \tag{1}$$

where $\boldsymbol{z}_m$ is the vector of experimental responses and $\boldsymbol{z}(\boldsymbol{\theta}_j)$ is the vector of the corresponding counterparts predicted by the model at the j^{th} iteration step, and therefore function of the modelling parameters vector, to be more precise the updating parameters vector $\boldsymbol{\theta}_j$.

The error given in eq. (1) can be approximated by an expansion in Taylor series of $\boldsymbol{z}(\boldsymbol{\theta}_j)$, as

$$\boldsymbol{\varepsilon}_{\boldsymbol{z}_j} = \boldsymbol{z}_m - \left(\boldsymbol{z}_j + \mathbf{S}_j\left(\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j\right) + O\left(\boldsymbol{\theta}_j^2\right)\right) \approx \left(\boldsymbol{z}_m - \boldsymbol{z}_j\right) - \mathbf{S}_j \Delta\boldsymbol{\theta}_j \tag{2}$$

where $\mathbf{S}_j$ is the sensitivity matrix assembled as

$$\mathbf{S}_j = \left.\frac{\partial z_i}{\partial \theta_l}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j} \tag{3}$$

for $i = 1, \ldots, n$ experimentally measured outputs and $l = 1, \ldots, k$ modelling parameters.

As $\boldsymbol{\varepsilon}_{\boldsymbol{z}_j} \to \mathbf{0}$, eq. (2) can be recast as

$$(\boldsymbol{z}_m - \boldsymbol{z}_j) = \mathbf{S}_j \left(\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j\right) \tag{4}$$

The exact analytical solution of eq. (3), in terms of the sensitivity of the eigensolutions was given by Fox and Kapoor [8]. The sensitivity of the predicted eigenvalues ω_i^2 is found by the partial derivative of the eigenproblem for undamped systems w.r.t. the updating parameters, as

$$\frac{\partial \omega_i^2}{\partial \theta_l} = \boldsymbol{\phi}_i^T \left[\frac{\partial \mathbf{K}}{\partial \theta_l} - \omega_i^2 \frac{\partial \mathbf{M}}{\partial \theta_l}\right] \boldsymbol{\phi}_i \tag{5}$$

As the eigenvectors $\boldsymbol{\phi}_i$ are linearly independent, the sensitivity of the predicted eigenvectors is found by a weighted linear combination of the $H \leq N$ eigenvectors of interest, defined as

$$\frac{\partial \boldsymbol{\phi}_i}{\partial \theta_l} = \sum_{h=1}^{H} a_{ilh} \boldsymbol{\phi}_i \tag{6}$$

which can be proved to result in the following expression [1],

$$\frac{\partial \boldsymbol{\phi}_i}{\partial \theta_l} = \sum_{h=1, h\neq i}^{H} \left(\frac{\boldsymbol{\phi}_h \boldsymbol{\phi}_h^T}{\omega_i^2 - \omega_h^2} \left[\frac{\partial \mathbf{K}}{\partial \theta_l} - \omega_i^2 \frac{\partial \mathbf{M}}{\partial \theta_l}\right] \boldsymbol{\phi}_i\right) - \frac{1}{2} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T \frac{\partial \mathbf{M}}{\partial \theta_l} \boldsymbol{\phi}_i \tag{7}$$

Knowing the sensitivity matrix, the updated vector of parameters is then given by

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j + \mathbf{T}_j \left(\boldsymbol{z}_m - \boldsymbol{z}_j\right) \tag{8}$$

where the transformation matrix $\mathbf{T}_j$ is generally the weighted pseudo-inverse of the sensitivity matrix $\mathbf{S}_j$,

$$\mathbf{T}_j = \left(\mathbf{S}_j^T \mathbf{W}_\varepsilon \mathbf{S}_j + \mathbf{W}_\vartheta\right)^{-1} \mathbf{S}_j^T \mathbf{W}_\varepsilon \tag{9}$$

and $\mathbf{W}_\varepsilon$ and $\mathbf{W}_\vartheta$ are scaling and regularization matrices, respectively, to allow for the regularisation of ill-posed sensitivity equations [9, 10]. Regarding regularization, a detailed discussion is given in [7].

The sensitivity method has been extended to lead with stochastic model updating as given in [11–13] and reviewed in [14]. Later, the two previous stochastic model updating techniques [11, 12] are shown to be equivalent and to reduce to the same formula within the theory of small perturbations about the mean [15].

3 SENSITIVITY-BASED STOCHASTIC MODEL UPDATING

Equivalently to the updated vector of parameters in the deterministic context (eq. (8)), the stochastic model updating equation may be written as

$$\tilde{\boldsymbol{\theta}}_{j+1} + \Delta\boldsymbol{\theta}_{j+1} = \tilde{\boldsymbol{\theta}}_j + \Delta\boldsymbol{\theta}_j + \left(\tilde{\mathbf{T}}_j + \Delta\mathbf{T}_j\right)\left(\tilde{\boldsymbol{z}}_m + \Delta\boldsymbol{z}_m - \tilde{\boldsymbol{z}}_j - \Delta\boldsymbol{z}_j\right) \tag{10}$$

where $\tilde{\bullet}$ and Δ denote the mean and perturbation on the mean, respectively. Then, by separating the zeroth-order and first-order terms, Khodaparast et al. [11] obtained the following two expressions:

$$O(\Delta^0): \quad \tilde{\boldsymbol{\theta}}_{j+1} = \tilde{\boldsymbol{\theta}}_j + \tilde{\mathbf{T}}_j\left(\tilde{\boldsymbol{z}}_m - \tilde{\boldsymbol{z}}_j\right) \tag{11}$$

$$O(\Delta^1): \quad \Delta\boldsymbol{\theta}_{j+1} = \Delta\boldsymbol{\theta}_j + \tilde{\mathbf{T}}_j\left(\Delta\boldsymbol{z}_m - \Delta\boldsymbol{z}_j\right) + \left(\sum_{i=1}^{n}\frac{\partial\tilde{\mathbf{T}}_j}{\partial z_m^{(i)}}\Delta z_m^{(i)}\right)\left(\tilde{\boldsymbol{z}}_m - \tilde{\boldsymbol{z}}_j\right) \tag{12}$$

Note that the third term on the right-hand-side of eq. (12) will be neglected on the grounds that $(\tilde{\boldsymbol{z}}_m - \tilde{\boldsymbol{z}}_j)$ is itself a small quantity of $O(\Delta^1)$.

Eq. (12) can be used to form the covariance matrix [4, 11],

$$\begin{aligned}\mathrm{cov}\left(\Delta\boldsymbol{\theta}_{j+1}, \Delta\boldsymbol{\theta}_{j+1}\right) = \mathrm{cov}\left(\Delta\boldsymbol{\theta}_j, \Delta\boldsymbol{\theta}_j\right) - \mathrm{cov}\left(\Delta\boldsymbol{\theta}_j, \Delta\boldsymbol{z}_j\right)\tilde{\mathbf{T}}_j^T - \tilde{\mathbf{T}}_j\mathrm{cov}\left(\Delta\boldsymbol{z}_j, \Delta\boldsymbol{\theta}_j\right) + \\ + \tilde{\mathbf{T}}_j\mathrm{cov}\left(\Delta\boldsymbol{z}_j, \Delta\boldsymbol{z}_j\right)\tilde{\mathbf{T}}_j^T + \tilde{\mathbf{T}}_j\mathrm{cov}\left(\Delta\boldsymbol{z}_m, \Delta\boldsymbol{z}_m\right)\tilde{\mathbf{T}}_j^T\end{aligned} \tag{13}$$

where the updated parameters $\Delta\boldsymbol{\theta}_j$ and hence the predictions $\Delta\boldsymbol{z}_j$ are assumed to be statistically independent of the measurements $\Delta\boldsymbol{z}_m$.

In the case of small parameter variability, then $\boldsymbol{\theta}_j$ may be approximated by the perturbation on the mean, as

$$\boldsymbol{\theta}_j = \tilde{\boldsymbol{\theta}}_j + \Delta\boldsymbol{\theta}_j = \tilde{\boldsymbol{\theta}}_j + \mathbf{T}(\tilde{\boldsymbol{\theta}}_j)\left(\boldsymbol{z}_j - \boldsymbol{z}(\tilde{\boldsymbol{\theta}}_j)\right) \tag{14}$$

assuming that

$$\tilde{\mathbf{T}}_j = \mathbf{T}(\tilde{\boldsymbol{\theta}}_j) \quad \text{and} \quad \tilde{\boldsymbol{z}}_j = \boldsymbol{z}(\tilde{\boldsymbol{\theta}}_j) \tag{15}$$

then,

$$\Delta\boldsymbol{\theta}_j = \tilde{\mathbf{T}}_j\left(\boldsymbol{z}_j - \tilde{\boldsymbol{z}}_j\right) = \tilde{\mathbf{T}}_j\Delta\boldsymbol{z}_j \tag{16}$$

and thus the terms of eq. (13) can by recast, leading to the very simple expression proposed in [15],

$$\mathrm{cov}\left(\Delta\boldsymbol{\theta}_{j+1}, \Delta\boldsymbol{\theta}_{j+1}\right) = \tilde{\mathbf{T}}_j\mathrm{cov}\left(\Delta\boldsymbol{z}_m, \Delta\boldsymbol{z}_m\right)\tilde{\mathbf{T}}_j^T \tag{17}$$

Eq. (17) allows the computation of the outputs covariances using the transformation matrix, $\tilde{\mathbf{T}}_j$, obtained at the final step of deterministic updating of the means using eq. (11), without the need for forward propagation.

In [4], several numerical examples are given to illustrate the potential of this updating technique and it was shown that the issue of selecting a consistent parameter set from the considered modelling parameters is crucial to update the covariance matrix, even if the parameter mean values are correctly updated.

4 SELECTION OF PARAMETERS FOR STOCHASTIC UPDATING

Recalling the perturbation method, the stochastic model updating problem defined in eq. (10) may be expressed as

$$(\boldsymbol{z}_m - \tilde{\boldsymbol{z}}_m) = \tilde{\mathbf{S}}_j \left(\boldsymbol{\theta}_{j+1} - \tilde{\boldsymbol{\theta}}_{j+1}\right) + \boldsymbol{\varepsilon}_{j+1} \tag{18}$$

by the assumption of small perturbation about the mean, as described before, $\Delta\boldsymbol{z}_m = \boldsymbol{z}_m - \tilde{\boldsymbol{z}}_m$ and $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}$. Hence, if one assumes an initial parameter estimate, for parameter selection purposes, eq. (18) can be recast as

$$(\boldsymbol{z}_m - \tilde{\boldsymbol{z}}_m) = \tilde{\mathbf{S}}_0 \left(\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}_0\right) + \boldsymbol{\varepsilon} \tag{19}$$

Considering n experimentally measured outputs and k modelling parameters, eq. (19) may be cast in a standardized form, such that

$$\begin{Bmatrix} \dfrac{z_1 - \tilde{z}_1}{\sigma_{z_1}} \\ \dfrac{z_2 - \tilde{z}_2}{\sigma_{z_2}} \\ \vdots \\ \dfrac{z_n - \tilde{z}_n}{\sigma_{z_n}} \end{Bmatrix} = \begin{bmatrix} \dfrac{\sigma_{\theta_1}}{\sigma_{z_1}}\dfrac{\partial \tilde{z}_1}{\partial \tilde{\theta}_1} & \dfrac{\sigma_{\theta_2}}{\sigma_{z_1}}\dfrac{\partial \tilde{z}_1}{\partial \tilde{\theta}_2} & \cdots & \dfrac{\sigma_{\theta_k}}{\sigma_{z_1}}\dfrac{\partial \tilde{z}_1}{\partial \tilde{\theta}_k} \\ \dfrac{\sigma_{\theta_1}}{\sigma_{z_2}}\dfrac{\partial \tilde{z}_2}{\partial \tilde{\theta}_1} & \dfrac{\sigma_{\theta_2}}{\sigma_{z_2}}\dfrac{\partial \tilde{z}_2}{\partial \tilde{\theta}_2} & \cdots & \dfrac{\sigma_{\theta_k}}{\sigma_{z_2}}\dfrac{\partial \tilde{z}_2}{\partial \tilde{\theta}_k} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\sigma_{\theta_1}}{\sigma_{z_n}}\dfrac{\partial \tilde{z}_n}{\partial \tilde{\theta}_1} & \dfrac{\sigma_{\theta_2}}{\sigma_{z_n}}\dfrac{\partial \tilde{z}_n}{\partial \tilde{\theta}_2} & \cdots & \dfrac{\sigma_{\theta_k}}{\sigma_{z_n}}\dfrac{\partial \tilde{z}_n}{\partial \tilde{\theta}_k} \end{bmatrix} \begin{Bmatrix} \dfrac{\theta_1 - \tilde{\theta}_1}{\sigma_{\theta_1}} \\ \dfrac{\theta_2 - \tilde{\theta}_2}{\sigma_{\theta_2}} \\ \vdots \\ \dfrac{\theta_k - \tilde{\theta}_k}{\sigma_{\theta_k}} \end{Bmatrix} + \underset{\sim}{\boldsymbol{\varepsilon}} \tag{20}$$

where the subscripts m and 0 are now omitted, for the sake of simplicity, and $\underset{\sim}{\boldsymbol{\varepsilon}}$ denotes the standardized error vector. Thus,

$$\underset{\sim}{\boldsymbol{z}} = \underset{\sim}{\mathbf{S}}\,\underset{\sim}{\boldsymbol{\theta}} + \underset{\sim}{\boldsymbol{\varepsilon}} \tag{21}$$

where $\underset{\sim}{\boldsymbol{z}}$, $\underset{\sim}{\boldsymbol{\theta}}$ and $\underset{\sim}{\mathbf{S}}$ are the standardized vectors of responses and parameters and the standardized sensitivity matrix, respectively.

The standardized parameters covariance matrix is given by the correlation matrix. If the chosen parameters are independent, then the covariance matrix is given by the identity matrix,

$$\operatorname{cov}\left(\underset{\sim}{\boldsymbol{\theta}}, \underset{\sim}{\boldsymbol{\theta}}\right) = \mathbf{I} \tag{22}$$

Assuming the error $\tilde{\boldsymbol{\varepsilon}}$ in eq. (20) to be independent of the parameters, then the output covariance matrix may be expressed as

$$\operatorname{cov}\left(\underset{\sim}{\boldsymbol{z}}, \underset{\sim}{\boldsymbol{z}}\right) = \underset{\sim}{\mathbf{S}}\,\underset{\sim}{\mathbf{S}}^T + \operatorname{cov}\left(\underset{\sim}{\boldsymbol{\varepsilon}}, \underset{\sim}{\boldsymbol{\varepsilon}}\right) \tag{23}$$

Eq. (23) may be expanded so that,

$$\operatorname{cov}\left(\underset{\sim}{\boldsymbol{z}}, \underset{\sim}{\boldsymbol{z}}\right) = \boldsymbol{s}_{\theta_1}\boldsymbol{s}_{\theta_1}^T + \boldsymbol{s}_{\theta_2}\boldsymbol{s}_{\theta_2}^T + \cdots + \boldsymbol{s}_{\theta_l}\boldsymbol{s}_{\theta_l}^T + \operatorname{cov}\left(\underset{\sim}{\boldsymbol{\varepsilon}}, \underset{\sim}{\boldsymbol{\varepsilon}}\right) \tag{24}$$

where $\boldsymbol{s}_{\theta_l}$ denotes the l^{th} column of the standardized sensitivity matrix $\underset{\sim}{\mathbf{S}}$. The term $\boldsymbol{s}_{\theta_l}\boldsymbol{s}_{\theta_l}^T$ on the right-hand-side of eq. (24) therefore represents the contribution of the l^{th} parameter to the standardized output covariance matrix. Hence, for parameter selection purposes, one would like to select those parameters that make the most significant contributions.

The covariance matrix of measured outputs may be expressed by its singular value decomposition as

$$\mathbf{A} = \operatorname{cov}\left(\underset{\sim}{\boldsymbol{z}}, \underset{\sim}{\boldsymbol{z}}\right) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H \tag{25}$$

As $\mathbf{A}$ is a square matrix, the right-singular vectors V are equal to the left ones, so that

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^H \tag{26}$$

From the right-hand-sides of eqs. (24) and (26), it is clear that the number of parameters that contribute to $\mathbf{A}$ must be equal to the number of non-zero singular values. The left-singular vectors corresponding to the non-zero singular values of $\mathbf{A}$ span the range of $\mathbf{A}$, so

$$\operatorname{range}\left[\mathbf{A}\right] = \operatorname{span}\left[\mathbf{U}_{\boldsymbol{\Sigma}\neq 0}\right] \tag{27}$$

Hence, the projection onto $\mathbf{U}_{\boldsymbol{\Sigma}\neq 0}$ of the contribution of each parameter θ_l to $\mathbf{A}$, i.e., each term on the right-hand-side of eq. (24), is then given by

$$\boldsymbol{s}'_{\theta_l} = \mathbf{U}_{\boldsymbol{\Sigma}\neq 0}\mathbf{U}_{\boldsymbol{\Sigma}\neq 0}^T\boldsymbol{s}_{\theta_l} \tag{28}$$

where $\mathbf{U}_{\boldsymbol{\Sigma}\neq 0}\mathbf{U}_{\boldsymbol{\Sigma}\neq 0}^T$ is the orthogonal projector onto the range of $\mathbf{A}$ (for further details see [16]).

Ideally, if a parameter θ_l makes a non-zero contribution, then $\boldsymbol{s}'_{\theta_l}$ must be given exactly by a linear combination of the columns of $\mathbf{U}_{\boldsymbol{\Sigma}\neq 0}$, so that $\boldsymbol{s}_{\theta_l}$ and $\boldsymbol{s}'_{\theta_l}$ are identical. In practice they will be different and the cosine-distance may be used to assess the closeness between $\boldsymbol{s}_{\theta_l}$ and $\boldsymbol{s}'_{\theta_l}$,

$$1 - \cos\Psi_l = 1 - \frac{\left|\boldsymbol{s}_{\theta_l}^T\boldsymbol{s}'_{\theta_l}\right|}{\left\|\boldsymbol{s}_{\theta_l}\right\|\left\|\boldsymbol{s}'_{\theta_l}\right\|} \tag{29}$$

where Ψ_l denotes the angle between $\boldsymbol{s}_{\theta_l}$ and $\boldsymbol{s}'_{\theta_l}$. The cosine-distance takes a value between zero and unity and, in practice, if less than a chosen threshold,

$$1 - \cos\Psi_l < \varepsilon_\Psi \tag{30}$$

then θ_l may be considered to be a contributing parameter.

The test for parameter θ_l in eq. (29) requires that $\operatorname{cov}\left(\underset{\sim}{\boldsymbol{z}}, \underset{\sim}{\boldsymbol{z}}\right)$ must be less than full rank, so there are columns of $\mathbf{U}$ corresponding to small (theoretically zero) singular values, i.e., $\mathbf{U}_{\boldsymbol{\Sigma}=0} \neq \emptyset$. Otherwise it is not possible to recognise wrongly selected parameters. This means that there must be more outputs than significant parameters.

5 Numerical example - Pin-jointed truss

The pin-jointed truss shown in Figure 1 has overall dimensions 5m×1m and is composed of 21 rod elements in total, each with a stiffness matrix given by,

$$\mathbf{K}^{(i)} = k_i \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{31}$$

for $i = 1, 2, \ldots, 21$ bar elements with a generic stiffness k_i. The elastic modulus, mass density and cross sectional area are assumed to take the values,

$$E = 70\text{GPa}, \quad \rho = 2700\text{kg/m}^3 \quad \text{and} \quad A = 0.03\text{m}^2$$

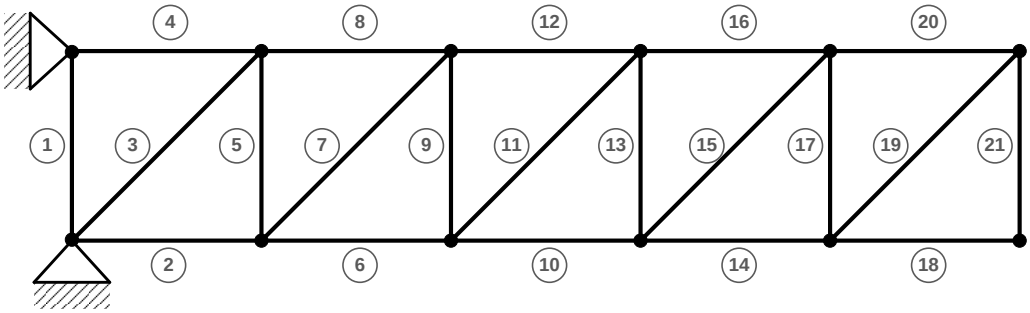


Figure 1: Pin-jointed truss.

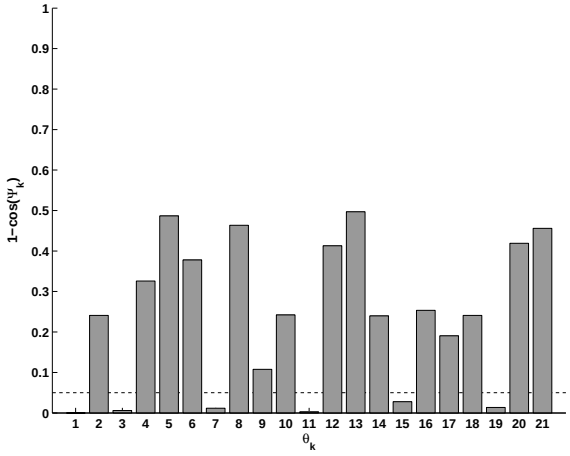
The five diagonal bars of nominal stiffness $(EA/L)_j = 1.485 \times 10^8\text{N/m}$ are each randomised to generate the experimental data set for updating. The true mean value of each is equal to the nominal stiffness and the standard deviations are given by $\sigma_{k_j} = 0.135\mu_{k_j}$, for $j = 3, 7, 11, 15, 19$, where 0.135 is the $\text{CoV}(k_j)$. For the purposes of parameter selection, the initial estimates of all the mean stiffnesses, $k_i,\ i = 1, 2, \ldots, 21$, are considered to be 70% of the reference values and the standard deviations are given by $\sigma_{k_j} = 0.27\mu_{k_3}$.

Parameter selection results are shown in Figure 2. It is seen that the correct parameters for updating are recognised correctly in each case of different sensitivity analysis.

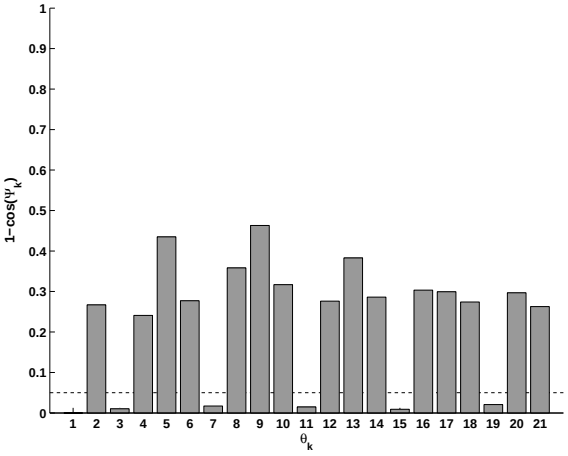
It can be seen from Figure 2 that the first bar element k_1 has zero cosine-distance. This happens because the boundary condition prevents any extension or compression of element 1, so that all the outputs are insensitive to it. When the constraints are removed, so the truss is in free-free conditions, the cosine-distance corresponding to parameter k_1 becomes finite and exceeds the threshold of 5%, as shown in Figure 3. It can be concluded that k_1 is not a randomised updating parameter.

Having correctly identified the randomised updating parameters, it is necessary to carry out the stochastic model updating. The initial values of the updating parameters are set to:

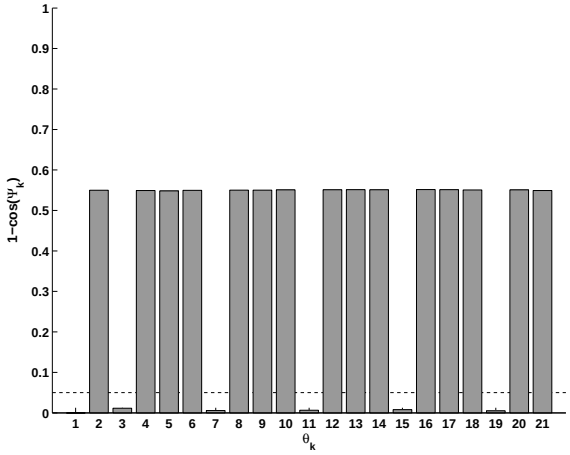
$$k_3 = 0.70\mu_{k_3},\ k_7 = 1.20\mu_{k_7},\ k_{11} = 0.90\mu_{k_{11}},\ k_{15} = 0.80\mu_{k_{15}},\ k_{19} = 1.15\mu_{k_{19}}$$



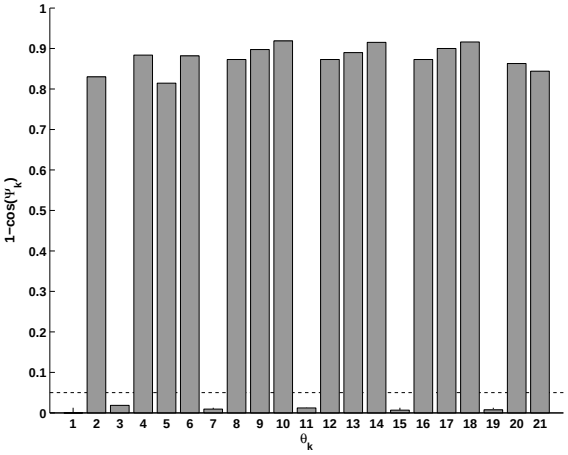
(a) Computed sensitivities: 1st 10 eigenvalues.



(b) Computed sensitivities: 1st 15 eigenvalues.

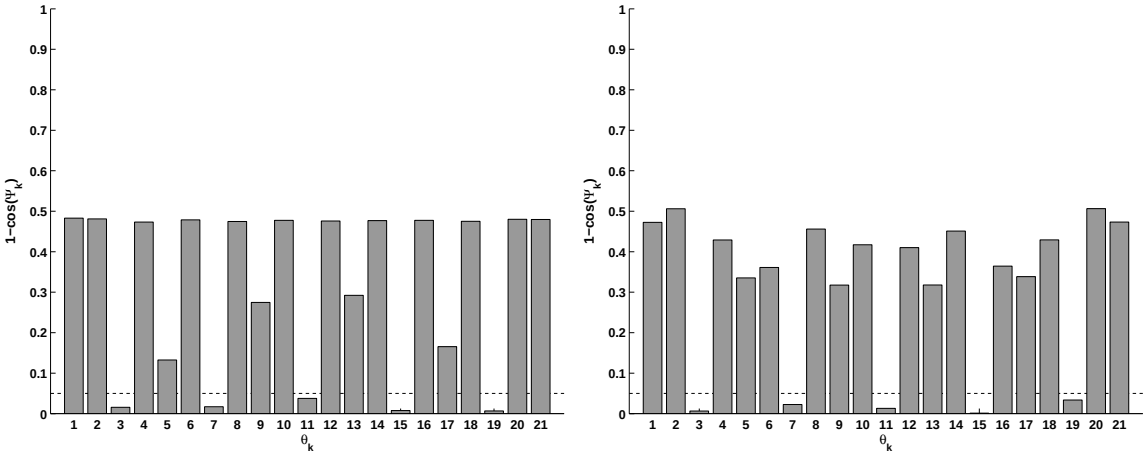


(c) Computed sensitivities: all eigenvalues.



(d) Computed sensitivities: all eigenvalues and eigenvectors.

Figure 2: Cosine distance (Pin-jointed truss).



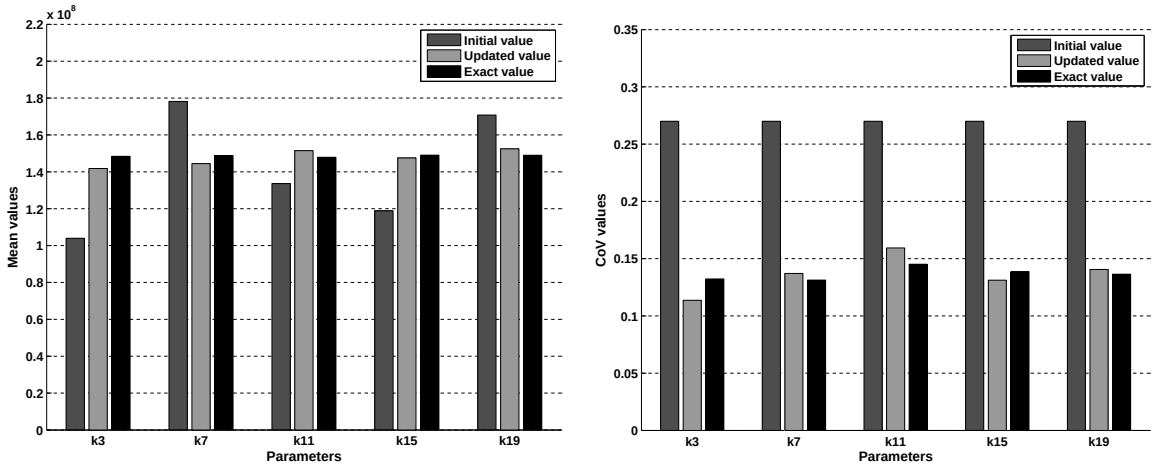
(a) Computed sensitivities: 1st 10 eigenvalues. (b) Computed sensitivities: all eigenvalues.

Figure 3: Cosine distance (Pin-jointed truss in free-free condition).

and

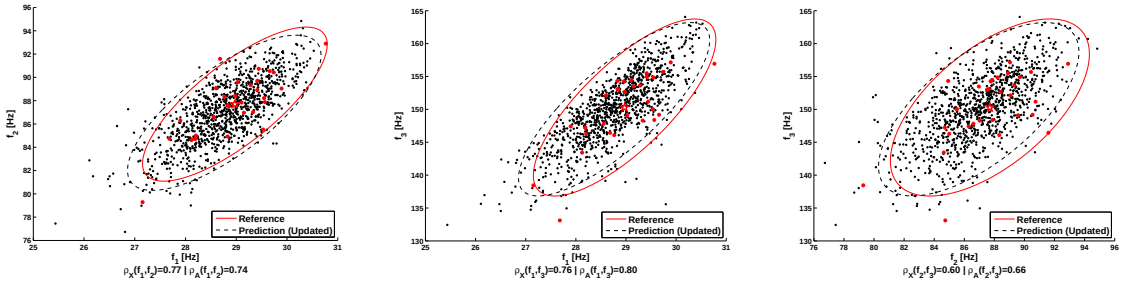
$$\mathrm{CoV}(k_j) = 2\frac{\sigma_{k_j}}{\mu_{k_j}}$$

Updating results are shown in Figures 4 to 7. It can be observed in Figures 4 and 5 that when the updating is carried out using the first 10 eigenvalues, then the exact mean values of the parameters and their covariances are closely approximated. The reconstructed output ellipses are in good but not quite perfect agreement with the measured data.



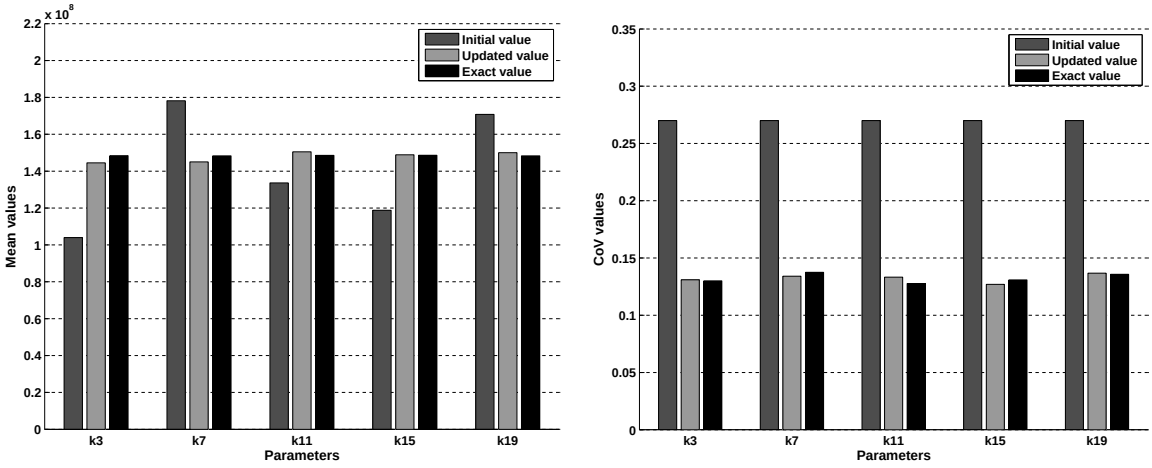
(a) Mean values. (b) Standard deviation values.

Figure 4: Identified parameters (Pin-jointed Truss - Eq. (17): using 1st 10 eigenvalues).



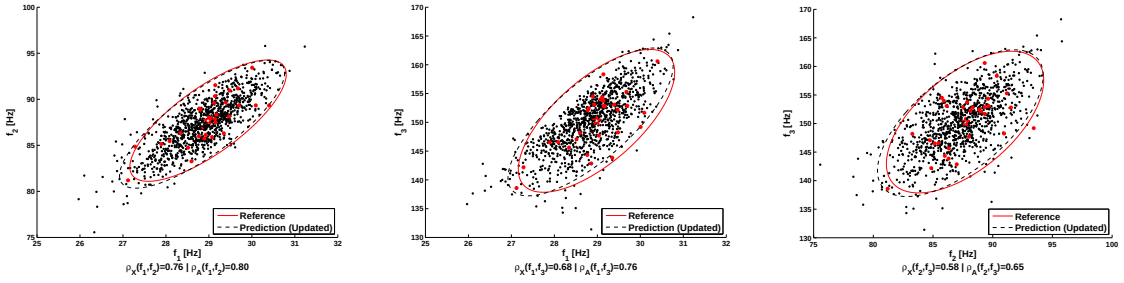
(a) f_1 vs f_2. (b) f_1 vs f_3. (c) f_2 vs f_3.

Figure 5: Frequency scatter plots (Pin-jointed Truss - Eq. (17): using 1st 10 eigenvalues).



(a) Mean values. (b) Standard deviation values.

Figure 6: Identified parameters (Pin-jointed Truss - Eq. (17): using 20 eigenvalues).



(a) f_1 vs f_2. (b) f_1 vs f_3. (c) f_2 vs f_3.

Figure 7: Frequency scatter plots (Pin-jointed Truss - Eq. (17): using 20 eigenvalues).

In Figures 6 and 7, one can observe that when all the 20 eigenvalues are used in model updating, the updated parameter means and covariances are in very good agreement with the values used to generate the data. Also, the output covariances are reconstructed almost exactly.

6 CONCLUSION

The stochastic model updating problem is briefly framed into the field of structural dynamics. A simple and efficient formula for updating the parameter covariance matrix is presented. This approach is computationally very efficient, as it uses only the measured output covariances and the transformation matrix obtained at the final deterministic step of updating the parameter mean values. However, the covariance updating requires the use of techniques related to parameter selection, as the covariance is not well updated if the updating parameters are wrongly chosen, even if the means may be updated correctly. It is shown that the measured output covariance matrix may be decomposed to reveal the contributions of each independent updating parameter. A vector is formed from a standard column of the sensitivity matrix and the cosine distance corresponding to the angle between this vector and its projection on the space defined by the columns of the covariance matrix is used to distinguish the updating parameters from other candidate parameters that might be deemed responsible for the observed output variability. A numerical examples is used to demonstrate the effective performance of the method in parameter selection and stochastic model updating.

ACKNOWLEDGEMENT

The first author acknowledges the financial support of the Portuguese Foundation for Science and Technology through UNIDEMI (PEst-OE/EME/UI0667/2014).

REFERENCES

[1] Maia, N. M. M., Silva, J. M. M., 1997. Theoretical and Experimental Modal Analysis. Mechanical Engineering Series. Research Studies Press Limited.

[2] Mottershead, J. E., Friswell, M. I., oct 1993. Model Updating In Structural Dynamics: A Survey. Journal of Sound and Vibration 167 (2), 347–375.

[3] Friswell, M. I., Mottershead, J. E., 1995. Finite Element Model Updating in Structural Dynamics. Solid Mechanics and its Applications. Kluwer Academic Publishers, Dordrecht.

[4] Silva, T. A. N., 2015. Development and Implementation of Model Updating Techniques in Structural Dynamics. Ph.d. thesis, Instituto Superior Técnico, Universidade de Lisboa.

[5] Marwala, T., 2010. Finite-element-model updating using computational intelligence techniques: Applications to structural dynamics. Springer-Verlag London.

[6] Hemez, F. M., Farrar, C. R., 2014. A Brief History of 30 Years of Model Updating in Structural Dynamics. In: Foss, G., Niezrecki, C. (Eds.), Special Topics in Structural Dynamics, Volume 6. Conference Proceedings of the Society for Experimental Mechanics Series. Springer International Publishing, Cham, pp. 53–71.

[7] Mottershead, J. E., Link, M., Friswell, M. I., 2011. The sensitivity method in finite element model updating: A tutorial. Mechanical Systems and Signal Processing 25 (7), 2275–2296.

[8] Fox, R. L., Kapoor, M. P., 1968. Rates of change of eigenvalues and eigenvectors. AIAA Journal 6 (12), 2426–2429.

[9] Ahmadian, H., Mottershead, J. E., Friswell, M. I., 1998. Regularisation methods for finite element model updating. Mechanical Systems and Signal Processing 12 (1), 47–64.

[10] Friswell, M. I., Mottershead, J. E., Ahmadian, H., 2001. Finite-element model updating using experimental test data: parametrization and regularization. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 359, 169–186.

[11] Khodaparast, H. H., Mottershead, J. E., Friswell, M. I., nov 2008. Perturbation methods for the estimation of parameter variability in stochastic model updating. Mechanical Systems and Signal Processing 22 (8), 1751–1773.

[12] Govers, Y., Link, M., apr 2010. Stochastic model updating - Covariance matrix adjustment from uncertain experimental modal data. Mechanical Systems and Signal Processing 24 (3), 696–706.

[13] Khodaparast, H. H., Mottershead, J. E., Badcock, K. J., may 2011. Interval model updating with irreducible uncertainty using the Kriging predictor. Mechanical Systems and Signal Processing 25 (4), 1204–1226.

[14] Mottershead, J. E., Link, M., Silva, T. A. N., Govers, Y., Khodaparast, H. H., 2015. The Sensitivity Method in Stochastic Model Updating. In: Sinha, J. K. (Ed.), Vibration Engineering and Technology of Machinery: Proceedings of VETOMAC X 2014. Vol. 23. Springer International Publishing, University of Manchester, UK, pp. 65–77.

[15] Silva, T. A., Maia, N. M., Link, M., Mottershead, J. E., mar 2016. Parameter selection and covariance updating. Mechanical Systems and Signal Processing 70-71, 269–283.

[16] Datta, B. N., 2010. Numerical Linear Algebra and Applications, 2nd Edition. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA.

SYMCOMP 2017
Guimarães, 6-7 April 2017

A NOVEL FINITE ELEMENT FORMULATION FOR THE BUCKLING ANALYSIS OF LAYERED COMPOSITE BEAM STRUCTURES

Hugo A.F.A. Santos

Área Departamental de Engenharia Mecânica
Instituto Superior de Engenharia de Lisboa
Instituto Politécnico de Lisboa
Rua Conselheiro Emídio Navarro 1, 1959-007 Lisboa, Portugal
e-mail: hsantos@dem.isel.pt

Keywords: Composite beams, buckling, Timoshenko theory, partial interaction, complementary energy, finite elements

Abstract. *A novel finite element formulation for the buckling analysis of two-layer composite beam structures will be presented. The beam elements possess a single flexible shear interface and each layer is modelled by means of Timoshenko's theory. The formulation relies on a hybrid variational principle of complementary energy only involving force/moment-like variables as fundamental unknown fields. The approximate field variables are such that all equilibrium differential equations hold in strong form, and the inter-element equilibrium as well as the Neumann boundary conditions are enforced by means of the Lagrangian-multiplier method. The accuracy and effectiveness of the proposed formulation will be demonstrated through the analysis of several numerical tests.*

1 INTRODUCTION

Due to their high load carrying capacity and preferential strength-to-weight ratio, composite beams are widely used in various industries, such as aerospace, automotive, nuclear, marine, biomedical and civil engineering. The mechanical behaviour of these structures largely depends on the type of connection between layers. If the layers are connected continuously by means of strong adhesives, the mechanical assumption of a perfect bond between the layers is reasonable. However, the layers are often connected non-continuously, by means of connectors, such as shear studs, nails, etc., which are not rigid. Therefore, some slip and uplift can occur at the interlayer. While the uplift is often small and can be neglected, the interlayer slip significantly affects the behaviour of composite elements. This phenomenon is called *partial* (or *incomplete*) *interaction* and is an important issue in composite structures [1]. In fact, the inclusion of the interlayer-slip effect in the theory of composite beams is essential for the optimal design and accurate representation of the actual mechanical behaviour of composite structures with partial interaction. There exist a large amount of literature where composite beams are analytically or numerically studied; see for instance [2, 3, 4, 5, 6, 7, 8] and references therein.

The strength of layered beams depends on their buckling resistance as well as cohesion between the layers. It is therefore of practical interest to develop analytical and/or numerical formulations to analyse their geometrically nonlinear behaviour. Analytical solutions to this problem can be found in, *e.g.*, [9, 10, 11]. Although analytical formulations can be regarded as relevant for the mechanical understanding of this problem, they are restricted to very simple applications. Therefore, numerical methods are often used due to their ability to handle sophisticated problems.

Numerical methods, in particular the finite element method, have been widely used in the analysis of composite beams with partial interaction. Displacement-based finite element formulations for Timoshenko composite beams were developed in, *e.g.*, [12]. However, these formulations may suffer from the shear-locking and slip-locking phenomena. Models that attempt to overcome these limitations within the framework of composite members with partial interaction were proposed in [7, 13, 14, 15, 16]. Alternative strategies to alleviate the slip-locking behaviour in the classical displacement-based finite element formulation were adopted in [17].

Hybrid and mixed finite element formulations can be used to naturally avoid locking effects, without the use of numerical tricks. Special types of these formulations are the so-called equilibrium-based formulations, often derived from complementary variational principles. In these formulations, the approximate fields are chosen so that the stress fields are in equilibrium or, in other words, internal equilibrium and continuous stress transmission between elements are satisfied exactly. These formulations have a special appeal for practical design engineers, despite the popularity of the conventional displacement formulations, due to the exact transmission of stresses across boundaries between adjacent structural members. This feature avoids the need for 'averaging' procedures

required to obtain unique nodal values of stresses when resorting to displacement formulations. Several equilibrium-based finite element formulations have been presented in the literature to the geometrically nonlinear analysis of framed structures [18, 19, 20, 21, 22].

The goal of the present work is to extend the finite element formulation proposed in [23] for geometrically linear two-layer composite beam structures, in which beam elements are assumed to possess a single flexible shear interface and each layer is modelled by means of Timoshenko's theory, to the buckling (geometrically nonlinear) analysis of composite beam structures. The new formulation relies on a hybrid variational principle of complementary energy only involving force/moment-like variables as fundamental unknown fields and the approximations are such that all equilibrium differential equations hold in strong form. The inter-element equilibrium and Neumann boundary conditions are enforced by means of the Lagrangian-multiplier method. Feasibility and effectiveness of the proposed formulation are numerically demonstrated through the analysis of several numerical tests.

2 BOUNDARY-VALUE PROBLEM

As previously mentioned, the aim of the present work is to study the buckling behaviour of composite beams with two material layers with a shear flexible interface, as illustrated in Figures 1 and 2. The Timoshenko (or first-order shear deformation) theory is adopted to describe the deformation of the beam layers. Transverse shear deformations are, therefore, allowed. The rotation angle and shear deformations are assumed to be identical in the two layers. In addition, the following assumptions are considered: (i) no uplift occurs between the two layers (*i.e.*, both layers have the same transverse displacement); (ii) slip can occur at the interlayer (*i.e.*, partial interaction is assumed); (iii) the layers are connected continuously by means of longitudinally distributed shear connectors; (iv) the material behaviour is linear elastic.

An initially straight and planar beam whose centroidal axis is parameterized by $x \in [0, L]$, with L the length of the beam, is considered. The centroidal axis is decomposed into an internal part, represented by $\Omega =]0, L[$, and a boundary part, identified by $\Gamma = \Gamma_N \cup \Gamma_D = \{0, L\}$, where Γ_N and Γ_D correspond to the Neumann and Dirichlet boundaries of the beam, respectively, such that $\Gamma_N \cap \Gamma_D = \emptyset$. The beam is subjected to a compressive axial force P on Γ_N, acting at the centroid of the beam for the full composite section, C_∞ - this assures that no pre-bending occurs; see Figure 3.

V and Q_s are the shear force and interlayer slip flux fields of the beam, respectively. N_i and M_i denote the axial force and bending moment fields of layer i, with $i = 1, 2$, acting at the centroid of each layer, C_i; see Figure 3. h_i corresponds to the distance from the centroidal axis of layer i to the interface. u_i denotes the centroidal axial displacement of layer i. The rotation angle of the cross-section of the layers is represented by θ. w is the centroidal transverse displacement of the beam.

The equilibrium of an infinitesimal beam element can be expressed by the following

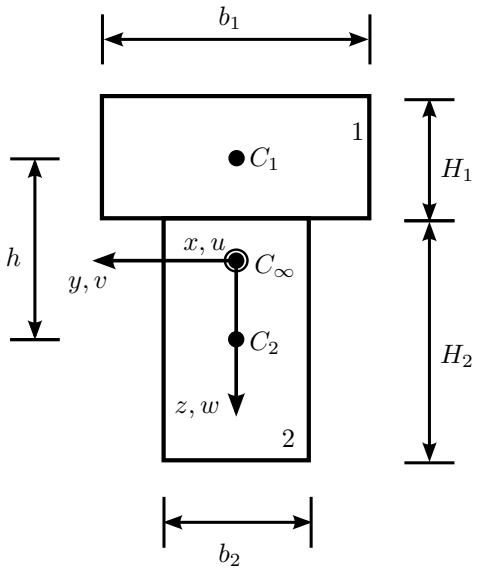


Figure 1: Geometric parameters of a partial interaction composite beam cross section

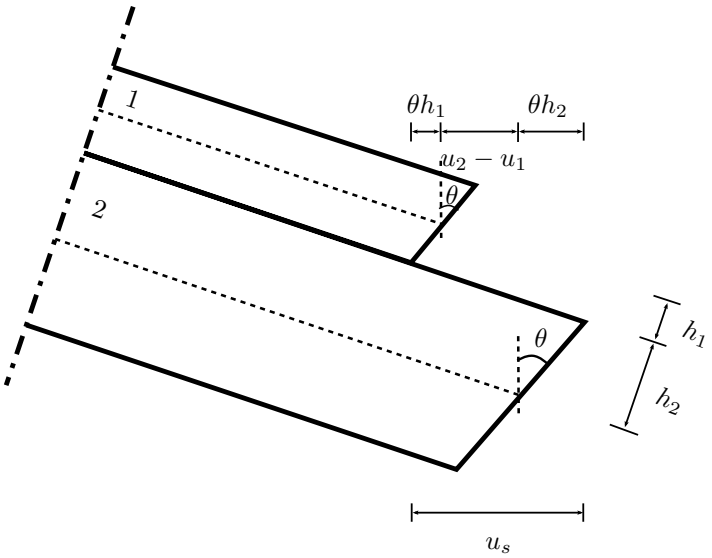


Figure 2: Kinematics of a partial interaction composite beam

differential equations in Ω

$$V' + (Pw')' = 0 \tag{1a}$$
$$M_1' + M_2' + Q_s h - V = 0 \tag{1b}$$
$$Q_s + N_1' = 0 \tag{1c}$$
$$Q_s - N_2' = 0 \tag{1d}$$

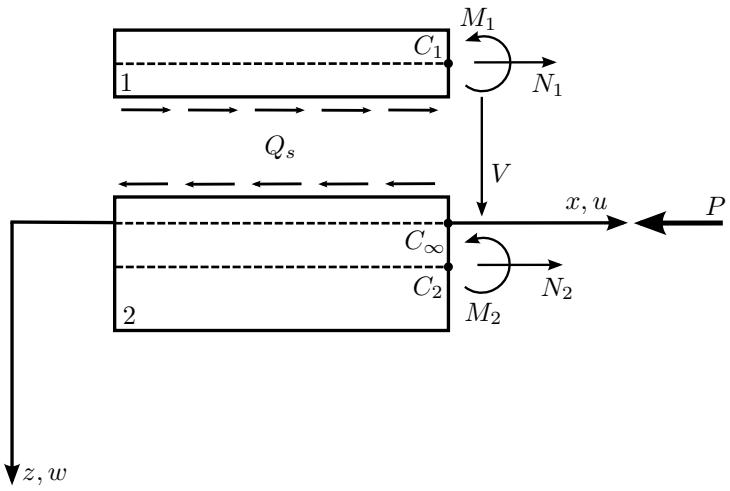


Figure 3: Internal forces and moments of a partial interaction composite beam and applied compressive axial force

representing equilibrium of shear forces, bending moments and axial forces, respectively, where $(\cdot)'$ stands for the derivative of $(\cdot)$ with respect to x.

The kinematical differential equations of the beam model under consideration arise in Ω as

$$\gamma = w' - \theta \tag{2a}$$
$$\kappa = -\theta' \tag{2b}$$
$$\varepsilon_1 = u_1' + \frac{1}{2}w' \tag{2c}$$
$$\varepsilon_2 = u_2' + \frac{1}{2}w' \tag{2d}$$

with γ being the shear deformation, κ the bending curvature and ε_i $(i = 1, 2)$ the axial deformation of layer i.

The cross-sectional compatibility relationship defines the interface slip field u_s as

$$u_s = u_2 - u_1 + \theta h \text{ in } \Omega. \tag{3}$$

The constitutive equations in Ω are taken as the following linear relationships

$$V = C\gamma \tag{4a}$$
$$M_1 = -E_1 I_1 \theta' \tag{4b}$$
$$M_2 = -E_2 I_2 \theta' \tag{4c}$$
$$N_1 = E_1 A_1 \varepsilon_1 \tag{4d}$$
$$N_2 = E_2 A_2 \varepsilon_1 \tag{4e}$$
$$Q_s = k_s u_s \tag{4f}$$

with $C = k_1 G_1 A_1 + k_2 G_2 A_2$ being the shear rigidity of the whole cross-section, where k_i is the shear correction coefficient of layer i, which depends on the cross-section geometry of the layer. E_i, G_i, I_i and A_i denote the Young's modulus, shear modulus, moment of inertia and cross-sectional area, respectively, of layer i. k_s stands for the interlayer slip modulus.

The Dirichlet boundary conditions of the problem are given on Γ_D as follows

$$u_1 - \bar{u}_1 = 0 \tag{5a}$$
$$u_2 - \bar{u}_2 = 0 \tag{5b}$$
$$w - \bar{w} = 0 \tag{5c}$$
$$\theta - \bar{\theta} = 0 \tag{5d}$$

The Neumann (or equilibrium) boundary conditions of the problem are given on Γ_N as follows

$$nN_1 - \bar{N}_1 = 0 \tag{6a}$$
$$nN_2 - \bar{N}_2 = 0 \tag{6b}$$
$$n(V + Pw') = 0 \tag{6c}$$
$$n(M_1 + M_2) = 0 \tag{6d}$$

with

$$n = \begin{cases} 1 & \text{if } x = L \\ -1 & \text{if } x = 0 \end{cases}$$

3 VARIATIONAL SETTING

The total potential energy functional associated with the beam model under study can be regarded as

$$\Pi_p(w, u_1, u_2, u_s, \theta) = \frac{1}{2}\int_\Omega \left(D\theta'^2 + E_1 A_1 u_1'^2 + E_2 A_2 u_2'^2 + C(w' - \theta)^2 + k_s u_s^2 \right) d\Omega \\ - \frac{1}{2}\int_\Omega P w'^2 \, d\Omega - [\bar{N}_1 u_1]_{\Gamma_N} - [\bar{N}_2 u_2]_{\Gamma_N} \tag{7}$$

with $D = E_1 I_1 + E_2 I_2$. It can be shown that, under the subsidiary conditions (5), the total potential energy renders a stationary principle. This energy functional can be transformed into the total complementary energy functional defined as

$$\Pi_c^*(M_1, M_2, N_1, N_2, V, Q_s, w) = \frac{1}{2}\int_\Omega \left(\frac{M_1^2}{E_1 I_1} + \frac{M_2^2}{E_2 I_2} + \frac{N_1^2}{E_1 A_1} + \frac{N_2^2}{E_2 A_2} + \frac{V^2}{C} + \frac{Q_s^2}{k_s} \right) d\Omega \\ - \frac{1}{2}\int_\Omega P w'^2 \, d\Omega - [N_1 \bar{u}_1]_{\Gamma_D} - [N_2 \bar{u}_2]_{\Gamma_D} - [V\bar{w}]_{\Gamma_D} - [(M_1 + M_2)\bar{\theta}]_{\Gamma_D} \tag{8}$$

Under the subsidiary conditions of equilibrium (1) and (6), the total complementary energy renders a stationary principle. As it can be seen, this energy functional does involve not only force/moment like quantities, but also the transverse displacement field. However, making use of (1a), this energy functional can be recast as a pure energy functional given by

$$\Pi_c(M_1, M_2, N_1, N_2, V, Q_s) = \frac{1}{2}\int_\Omega \left(\frac{M_1^2}{E_1 I_1} + \frac{M_2^2}{E_2 I_2} + \frac{N_1^2}{E_1 A_1} + \frac{N_2^2}{E_2 A_2} + \frac{V^2}{C} + \frac{Q_s^2}{k_s}\right) d\Omega - \frac{1}{2}\int_\Omega \frac{(V_0 - V)^2}{P}\, d\Omega - [N_1\bar{u}_1]_{\Gamma_D} - [N_2\bar{u}_2]_{\Gamma_D} - [V\bar{w}]_{\Gamma_D} - [(M_1 + M_2)\bar{\theta}]_{\Gamma_D} \quad (9)$$

with $V_0 \in \mathbb{R}$ a constant of integration. In addition, the Neumann condition (6c) can be rewritten as

$$nV_0 = 0$$

Let us now assume that the entire domain Ω is partitioned in subdomains $\Omega_e \subset \Omega$, such that $\Omega = \cup_{e=1}^{n_e}\Omega_e$ in which n_e represents the number of beam elements. If the inter-element equilibrium conditions and Neumann boundary conditions are relaxed within the framework of the complementary energy principle, then, the following augmented Lagrangian, or hybrid complementary energy, must be considered

$$L_c = \sum_{e=1}^{n_e}\Pi_{c,e} + \sum_{i=1}^{n_{int}}\left(\lambda_i^{N_1}[\![N_1]\!]_{\Gamma_i} + \lambda_i^{N_2}[\![N_2]\!]_{\Gamma_i} + \lambda_i^{V}[\![V]\!]_{\Gamma_i} + \lambda_i^{M_1}[\![M_1]\!]_{\Gamma_i} + \lambda_i^{M_2}[\![M_2]\!]_{\Gamma_i} + \lambda_i^{Q_s}[\![Q_s]\!]_{\Gamma_i}\right) \quad (10)$$

where n_{int} is the number of inter-element and Neumann boundaries and Γ_i is the inter-element boundary i. $[\![(\cdot)]\!]$ stands for the jump of $(\cdot)$ on Γ_i. $\lambda_i^{N_1}$, $\lambda_i^{N_2}$, λ_i^{V}, $\lambda_i^{M_1}$, $\lambda_i^{M_1}$ and $\lambda_i^{Q_s}$ are the energy-conjugate Lagrange multipliers of N_1, N_2, V, M_1 and M_2, respectively, defined on Γ_i.

The stationarity conditions of the hybrid complementary energy L_c are the equilibrium equations of the problem defined in Ω and Γ_N, the inter-element equilibrium conditions on Γ_{int} and, in addition, the inter-element compatibility conditions on Γ_{int}, with $\Gamma_{int} = \cup_{i=1}^{n_{int}}\Gamma_i$.

4 FINITE ELEMENT APPROXIMATIONS

The following trial finite element approximations are assumed for the bending moments

$$M_1^h = M_1^i\left(1 - \frac{x}{L}\right) + M_1^j\frac{x}{L} - 4M_1^k\frac{x(L-x)}{L^2} \quad (11a)$$

$$M_2^h = M_2^i\left(1 - \frac{x}{L}\right) + M_2^j\frac{x}{L} - 4M_2^k\frac{x(L-x)}{L^2} \quad (11b)$$

where the pairs M_1^i, M_2^i and M_1^j, M_2^j are the bending moments of layers 1 and 2, defined at $x = 0$ and $x = L$, respectively. M_1^k and M_2^k are the mid-span bending moments of layers 1 and 2.

The approximations for shear and axial forces are taken as

$$V^h = V^i\left(1 - \frac{x}{L}\right) + V^j\frac{x}{L} \tag{12a}$$

$$N_1^h = N_{10} - \int Q^h \, dx \tag{12b}$$

$$N_2^h = N_{20} + \int Q^h \, dx \tag{12c}$$

with V^i and V^j the shear forces defined at $x = 0$ and $x = L$, respectively, and N_{10} and N_{20} the axial force parameters defined at $x = 0$, where the approximation for shear flux Q_s^h is assumed as

$$Q_s^h = \frac{1}{h}\left(M_1^{h\prime} + M_2^{h\prime}\right) + V^h \tag{13}$$

A Galerkin approach is adopted, *i.e.*, the problem is numerically approached assuming the same trial and test approximation function spaces.

These approximations are such that all equilibrium differential equations hold in a strong form. It is also important to remark that the approximations for bending moments and shear forces are selected so that the corresponding inter-element equilibrium conditions are *a priori* satisfied. This avoids the need to enforce inter-element continuity of moments and shear forces by resorting to the Lagrangian multiplier method.

The discrete form of the hybrid complementary energy is therefore obtained as

$$L_c^h = \sum_{e=1}^{n_e} \Pi_{c,e}^h + \sum_{i=1}^{n_{int}} \left(\lambda_i^{N_1}[\![N_1^h]\!]_{\Gamma_i} + \lambda_i^{N_2}[\![N_2^h]\!]_{\Gamma_i} + \lambda_i^{Q_s}[\![Q_s^h]\!]_{\Gamma_i}\right) \tag{14}$$

where n_e is the number of beam elements and n_{int} is the number of inter-element and Neumann boundaries. Note that the terms corresponding to the enforcement of the inter-element shear force and moment equilibrium conditions were dropped, as they are no longer necessary due to the aforementioned arguments.

5 NUMERICAL TESTS

To validate and assess the accuracy and effectiveness of the proposed finite element formulation, a simply-supported beam is analyzed with various goals. The obtained results are compared with the exact analytical solutions given in [9]. The cross-section adopted has the following dimensions: $H_1 = 20$ cm, $H_2 = 30$ cm and $b = 30$ cm. The shear correction coefficients were taken as $k_1 = k_2 = 5/6$.

The beam length was taken as $L = 250$ cm. The material parameters of the beam were set as $E_1 = E_2 = 1200$ kN/cm^2, $G_1 = 80$ kN/cm^2 and $G_2 = 120$ kN/cm^2.

5.1 Simply-supported beam - Accuracy test

The goal of this test is to assess the accuracy of the proposed finite element formulation. With this purpose, the interlayer slip modulus was taken as $k_s = 0.243$ kN/cm^2, which corresponds to a connection with very low stiffness.

Uniform meshes of 1, 2, 3, 4, 5, 6, 7, 8 and 16 finite elements were considered. The numerical results obtained for the buckling loads of the beam are displayed in Table 5.1. Clearly, the computed results for the first four buckling loads converge (from above) to their corresponding exact analytical solutions, which were computed using the following expression, see [9],

$$P_{cr,T}^{(n)} = \frac{1 - \frac{\beta-1}{\beta+\frac{\alpha}{\xi^2}}}{1 + \frac{P_{cr,EB}^{(n)}}{C}\left(1 - \frac{\beta-1}{\beta+\frac{\alpha}{\xi^2}}\right)} P_{cr,EB}^{(n)} \tag{15}$$

with

$$\alpha = k_s\left(\frac{1}{E_1A_1} + \frac{1}{E_2A_2} + \frac{h^2}{E_1I_1 + E_2I_2}\right), \beta = \frac{E_1I_1 + E_2I_2 + \frac{E_1A_1E_2A_2}{E_1A_1+E_2A_2}h^2}{E_1I_1 + E_2I_2}, \xi = \frac{n\pi}{L}$$

where $n = 1, 2, 3, ...$

n_e	$P_{cr}^{(1)}$	$P_{cr}^{(2)}$	$P_{cr}^{(3)}$	$P_{cr}^{(4)}$
1	14996.3			
2	14996.3	44366.1		
3	14851.6	45473.1	69932.9	
4	14832.4	44366.1	72226.6	87639.0
5	14827.6	44120.3	70512.8	90238.6
6	14826.0	44044.5	69932.9	88709.0
7	14825.4	44015.0	69716.2	87968.1
8	14825.0	44001.6	69622.1	87639.0
16	14824.6	43985.2	69514.8	87280.7
Exact	14824.2	43984.2	69509.0	87263.5

Table 1: Critical loads of the simply-supported beam - Accuracy test

5.2 Simply-supported beam - Slip-locking test

To demonstrate that the proposed finite element formulation is free from slip-locking, the interlayer slip modulus was set to $k_s = 2430$ kN/cm^2, which corresponds to a very high stiffness connection. Uniform meshes of 1, 2, 3, 4, 5, 6, 7, 8 and 16 finite elements were considered. The computed results for the first four critical loads are displayed in Table 5.2.

As it can be seen, the approximated results converge in all cases to their corresponding analytical solutions. This proves the insensitivity of the proposed formulation to the slip-locking phenomenon.

n_e	$P_{cr}^{(1)}$	$P_{cr}^{(2)}$	$P_{cr}^{(3)}$	$P_{cr}^{(4)}$
1	36178.3	84102.6		
2	37440.6	84102.6	101521.8	
3	37694.3	79576.3	104473.4	110300.6
4	38231.0	79490.8	100046.1	112625.9
5	38431.9	79679.7	99701.2	109808.8
6	38585.4	79791.6	99645.2	109414.8
7	38674.9	79883.2	99644.4	109277.8
8	38736.1	79944.6	99661.8	109224.6
16	38868.9	80087.8	99734.2	109201.1
Exact	40131.9	81666.3	101138.8	110485.8

Table 2: Critical loads of the simply-supported beam - Slip-locking test

5.3 Simply-supported beam - Shear-locking test

Finally, the sensitivity to shear-locking is numerically assessed. To do so, and for different length-to-thickness L/h ratios, the first two critical loads of the two-layer Timoshenko composite beam are obtained with the present formulation on a 16 finite element mesh, $P_{cr}^{(i)}$, and compared with the analytical solutions provided by the Euler-Bernoulli theory, $P_{cr,EB}^{(i)}$, see Table 5.3. The interlayer slip modulus was set in this test to $k_s = 1 \times 10^6$ kN/cm^2, which corresponds to a connection that assures a full composite beam section behaviour. As can be observed, the numerical results for the two-layer Timoshenko beam model converge to the Euler-Bernoulli's solution as the beam becomes thinner. It can therefore be concluded that the proposed formulation is free from shear-locking.

L/h	$P_{cr}^{(1)}/P_{cr,EB}^{(1)}$	$P_{cr}^{(2)}/P_{cr,EB}^{(2)}$
10	0.8861	0.6802
100	0.9845	0.9812
1000	0.9856	0.9856
10000	0.9862	0.9857

Table 3: Critical loads of the simply-supported beam - Shear-locking test

6 CONCLUSIONS

This paper introduces a novel complementary energy functional for the buckling analysis of two-layered composite beam structures and its associated finite element formulation. The formulation only involves force/moment like quantities as the fundamental unknown variables. Its feasibility and effectiveness were numerically demonstrated through the analysis of several numerical tests, which show that, unlike classical displacement-based formulations, the proposed formulation is naturally free from both shear- and slip-locking phenomena. The presented formulation can be straightforwardly extended to the free vibration analysis of two-layered composite beams. Another appealing future development is the case of composite beams with more than two layers.

REFERENCES

[1] Oehlers, D.J., Bradford M.A. *Composite steel and concrete structural members: fundamental behaviour*, Pergamon Press, Oxford, 1995

[2] Wu, Y.F., Oehlers D.J., Griffith M.C. "Partial-interaction analysis of composite beam/column members" *Mechanics of Structures and Machines* Vol. **30(3)**, pp. 309-332, 2002

[3] Martinelli, E., Faella, C., di Palma, G. "Shear-Flexible Steel-Concrete Composite Beams in Partial Interaction: Closed-Form "Exact" Expression of the Stiffness Matrix" *Journal of Engineering Mechanics* Vol. **138(2)**, pp. 151-163, 2012

[4] Ranzi, G., Bradford, M.A. "Direct stiffness analysis of a composite beam-column element with partial interaction" *Computers and Structures* Vol. **85**, pp. 1206-1214, 2007

[5] Girhammar, U.A., Pan D. "Exact static analysis of partially composite beams and beam-columns" *International Journal of Mechanical Sciences* Vol. **49**, pp. 239-255, 2007

[6] Murakami, H. "A laminated beam theory with interlayer slip" *Journal of Applied Mechanics* Vol. **51(9)**, pp. 551-559, 1984

[7] Schnabl, S., Saje, M., Turk, G., Planinc, I. "Analytical solution of two-layer beam taking into account interlayer slip and shear deformation" *Journal of Structural Engineering* Vol. **133(6)**, pp. 886-894, 2007

[8] Martinelli, E., Nguyen, Q.H., Hjiaj, M. "Dimensionless formulation and comparative study of analytical models for composite beams in partial interaction" *Journal of Constructional Steel Research* Vol. **75**, pp. 21-31, 2012

[9] R., Xu, Y., Wu "Static, dynamic, and buckling analysis of partial interaction composite members using Timoshenko's beam theory" *International Journal of Mechanical Sciences* Vol. **49(10)**, pp. 1139-1155, 2007

[10] Challamel, N., Girhammar, U.A. "Variationally-based theories for buckling of partial composite beam-columns including shear and axial effects" *Engineering Structures* Vol. **33(8)**, pp. 2297-2319, 2011

[11] R., Xu, G., Wang "Variational principle of partial-interaction composite beams using Timoshenko's beam theory" *International Journal of Mechanical Sciences* Vol. **60(1)**, pp. 72-83, 2012

[12] Dall'Asta A., Zona, A. "Non-linear analysis of composite beams by a displacement approach" *Computers and Structures* Vol. **80**, pp. 2217-2228, 2002

[13] Salari, M.R., Spacone, E., Shing, P.B., Frangopol, D.M. "Nonlinear analysis of composite beams with deformable shear connectors" *Journal of Structural Engineering (ASCE)* Vol. **124(10)**, pp. 1148-1158, 1998

[14] Ayoub, A., Filippou, F.C. "Mixed formulation of nonlinear steel-concrete composite beam elements" *Journal of Structural Engineering (ASCE)* Vol. **126(3)**, pp. 371-381, 2000

[15] Dall'Asta, A., Zona, A. "Three-field mixed formulation for non-linear analysis of composite beams with deformable shear connection" *Finite Elements in Analysis and Design* Vol. **40(4)**, pp. 425-448, 2004

[16] Ayoub, A. "A force-based model for composite steel-concrete beams with partial interaction" *Journal of Constructional Steel Research* Vol. **61**, pp. 387-414, 2005

[17] Erkmen, R.E., Bradford, M.A. "Treatment of slip locking for displacement-based finite element analysis of composite beam-columns" *International Journal for Numerical Methods in Engineering* Vol. **85(7)**, pp. 805-826, 2011

[18] Santos, H.A.F.A., Moitinho de Almeida, J.P. "Equilibrium-based finite element formulation for the geometrically exact analysis of planar framed structures" *Journal of Engineering Mechanics* Vol. **136(12)**, pp. 1474-1490, 2010

[19] Santos, H.A.F.A., Pimenta, P.M., Moitinho de Almeida, J.P. "A hybrid-mixed finite element formulation for the geometrically exact analysis of three-dimensional framed structures" *Computational Mechanics* Vol. **48(5)**, pp. 591-613, 2011

[20] Santos, H.A.F.A. "Complementary-Energy Methods for Geometrically Non-linear Structural Models: An Overview and Recent Developments in the Analysis of Frames" *Archives of Computational Methods in Engineering* Vol. **18(4)**, pp. 405-440, 2011

[21] Santos, H.A.F.A., Almeida Paulo, C.I. "On a pure complementary energy principle and a force-based finite element formulation for non-linear elastic cables" *International Journal of Non-Linear Mechanics* Vol. **46(2)**, pp. 395-406, 2011

[22] Santos, H.A.F.A. "Variationally consistent force-based finite element method for the geometrically non-linear analysis of Euler-Bernoulli framed structures" *Finite Elements in Analysis and Design* Vol. **53**, pp. 24-36, 2012

[23] Santos, H.A.F.A., Silberschmidt, V.V. "Hybrid equilibrium finite element formulation for composite beams with partial interaction" *Composite Structures* Vol. **108**, pp. 646-656, 2014

SPIDER WEB SHAPE ENERGY ABSORBER

Khalid Almitani [1]*, Azeez Bakare [2] Abdulmalik Alghamdi [2]

1: Mechanical Engineering Department
Engineering
King Abdulaziz University
P. O. Box 80271, Jeddah 21589, Saudi Arabia
e-mail: kalmettani@kau.edu.sa

2: Mechanical Engineering Department
Engineering
King Abdulaziz University
P. O. Box 80271, Jeddah 21589, Saudi Arabia
e-mail: azeezbakare@yahoo.com, aljinaidi@kau.edu.sa

Keywords: Impact, Energy absorbing capacity, Quasi-Static, Spider web shape energy absorber, perfectly plastic deformation.

Abstract *In this paper, the energy absorption response of spider web energy absorber is carried out using Finite element analysis. It is centered on perfectly plastic deformation of spider web when subjected to Quasi-Static loading taking into consideration its crushing load displacement characteristics. The Numerical analysis displayed an initial collapse at a minimal load followed by amplitude of progressive increase in loads which signifies gradual perfectly plastic deformation.*

1. INTRODUCTION

Energy absorbers are devices that changes kinetic energy to other forms of energy [1]. Many studies have been carried out on the consequences of different crushing rates, various types of material properties, different forms of environmental temperature etc. on energy absorption capabilities of various types of designs. With increasing awareness of safety consciousness in everyday activities which can be extended to design of safety devices with the aim of reducing the effect of incidence related to people and property, it is necessary to carry out more studies on the effects of impacts (for example; earthquakes, accidents caused by air, road, rail and water transportation etc.) and the ways to reduce its consequence. The effect of impact can be greatly reduced with the use of energy absorbers.

Alghamdi [1] carried out an extensive overview of various shapes of collapsible energy absorber and their forms of deformation. Some of the common shapes studied are; frusta, honeycomb, circular tubes, square tubes etc. The energy absorption response of tapered and straight rectangular tubes were investigated by Nagel and Thambiratnam [2] under Quasi- static and dynamic loading using both ABAQUS and LS-DYNA finite element code. The finite element models were confirmed by comparing them with the existing models. The result showed the advantage of the application of tapered tubes for impact energy absorber. Othman and Jailani's [3] work was similar to [2] but their investigation focused on off-axis angle loading with the incorporation of polyurethane foam to fill the thin-walled tapered tubes. In another study, Erdin et al. [4], studied and compared the behaviour of a functionally graded thickness circular thin- walled aluminium tubes with a uniform thickness equivalents under quasi-static axial compression loading. Abramowicz [5], presented a paper on the on the crushing method of thin-walled mechanisms. With the use of LS-DYNA finite element code, an impact collapse simulation was carried out by Sun-Kyu Kim et al. [6] with the purpose of improving energy absorption power during crushing with the application of a controller. Thin walled members were studied with the purpose of application in vehicular structures. Minagawa et al. [7] carried out analysis on the effect of combination of both steel pipe and rubber as shock absorbers in base isolated bridges. Alghamdi [8] worked on re-inverted aluminium frusta, plastic deformation of the frusta was studied by investigating the effect of both wall thickness and angular changes with respect to absorbed energy. Various studies have been carried out on honeycomb, one of them is the work done by Said and Reddy [9]. Subjected to quasi-static loading, energy absorption of honeycomb made from aluminium was studied and they came up with the result that showed initial collapse followed by progressive folding collapse at peak load and successive loads respectively [9]. Similar to the study of honeycomb is the work carried out by Ince et al. [10]. Both experimental (by using a drop test unit) and numerical (by using ANSYS finite element software) investigation was carried out on impact conduct of crash boxes made from aluminium and steel.

Even though there is an ongoing work on the mechanical properties of spider web, there is virtually no work done on the usage of the unique shape of spider web as energy absorber. Therefore, the aim of this study is to carry out material behaviour of perfectly plastic deformation on spider web as energy absorber. In this paper, FEM was carried out on spider web as energy absorber. It is centred on perfectly plastic deformation of spider web shaped when subjected to quasi-static loading taking into consideration its crushing load displacement characteristics.

2. SPIDER WEB ENERGY ABSORBER

The spider web energy absorber, as shown in Fig. 1 below, was drawn using Ansys R15.0 DesignModeler with H1 = 200mm, H2 = 200mm, V1 = 200mm and V2 = 200mm with a spacing of 50mm from one individual web to the other. The design characteristics is shown in Table 1 indicating varying cross sectional diameter, its equivalent cross sectional area and volume of the absorber from 2mm to 5mm.

Cross sectional diameter (mm)	2	3	4	5
Cross sectional Area (mm^2)	3.1414	7.0682	12.566	19.634
Volume (mm^3)	10797	24293	43187	67480

Table 1. Geometrical characteristics of spider web as energy absorber.

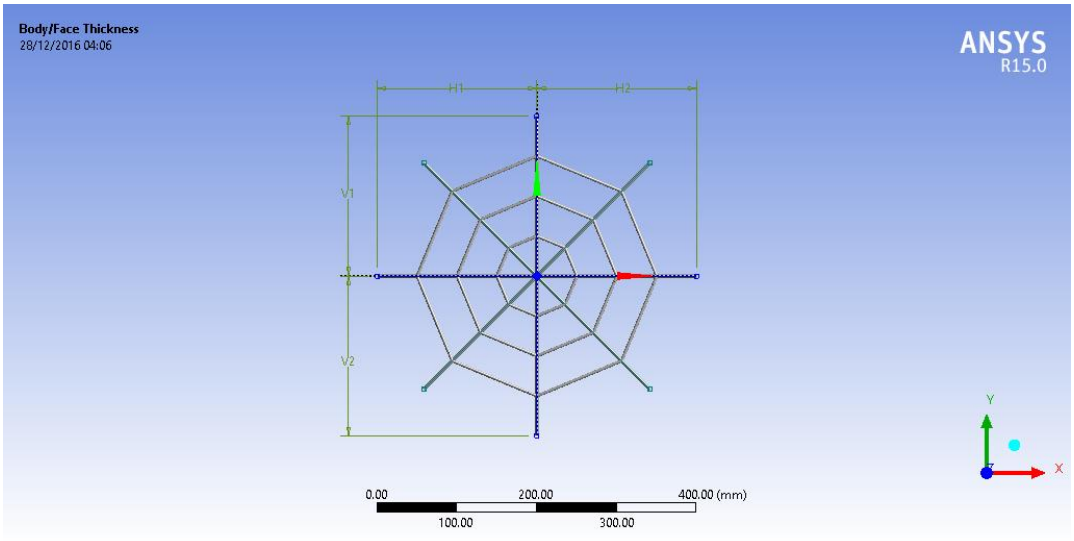


Fig. 1: A 2 Dimensional representation of spider web energy absorber.

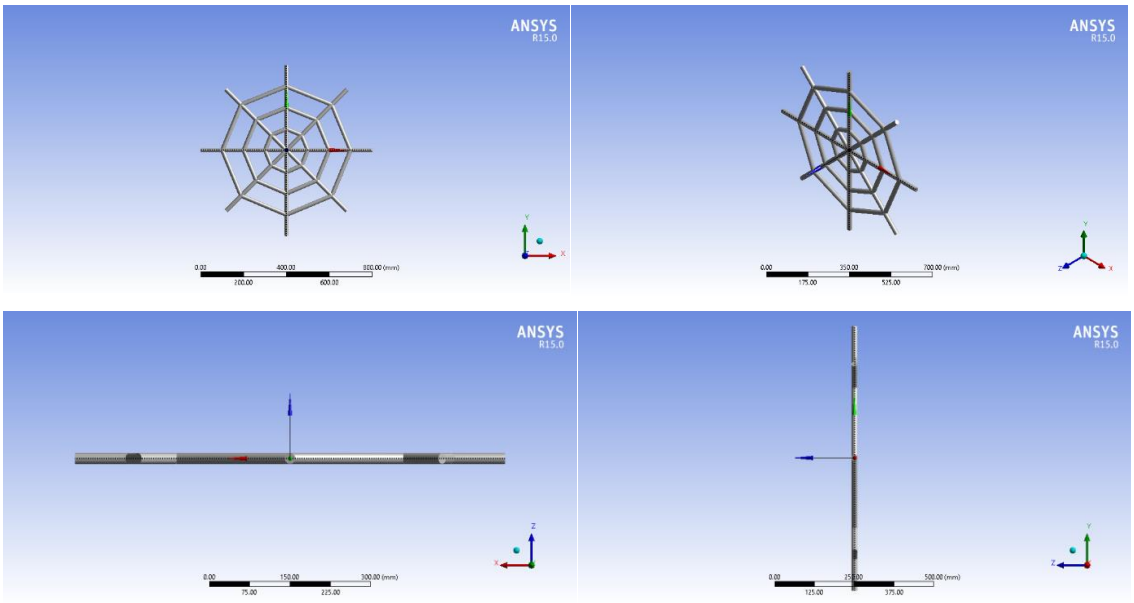


Fig. 2: Front, side, plan and isometric view of spider web energy absorber.

3. FINITE ELEMENT MODEL

Ansys R15.0 (Explicit Dynamics modeller) was used to model and study the deformation of the spider web energy absorber when subjected to various loading conditions while changing the cross sectional diameter of the web (from 2mm to 5mm), length of the web (from 250mm t0 400mm) and number of webs (from 3 to 8).

The characteristics of the model which is made up of a circular cross section is shown in Table 2 below. The web characteristics are a function of varying number of webs which is varied from 3 to 8.

Number of webs	Number of rod elements	Number of nodes	Number of edges	Number of vertices	Number of body
3	219	430	21	13	1

4	260	509	28	17	1
5	283	552	35	21	1
6	300	583	42	25	1
7	336	652	49	29	1
8	352	681	56	33	1

Table 2. The characteristics of the spider web model.

The load was subjected along the Z- axis as shown in Fig. 3 below, while the whole body was fixed at its extreme edges. Aluminium alloy and structural steel were used for the analysis. The analysis were carried out taking three conditions into account;

1. Varying number of web diagonal (from 3 to 8) while the circular cross section, length of the web and the number of lobes are fixed at 2mm, 400mm and 3 respectively

2. Varying length of the web (from 200mm to 400mm) while the circular cross section, the number of webs and the number of lobes are fixed at 2mm, 8 and 3 respectively.

3. Varying the circular cross section (from 2mm to 5mm) while the length, number of webs and the number of lobes are fixed at 200mm, 8 and 3 respectively.

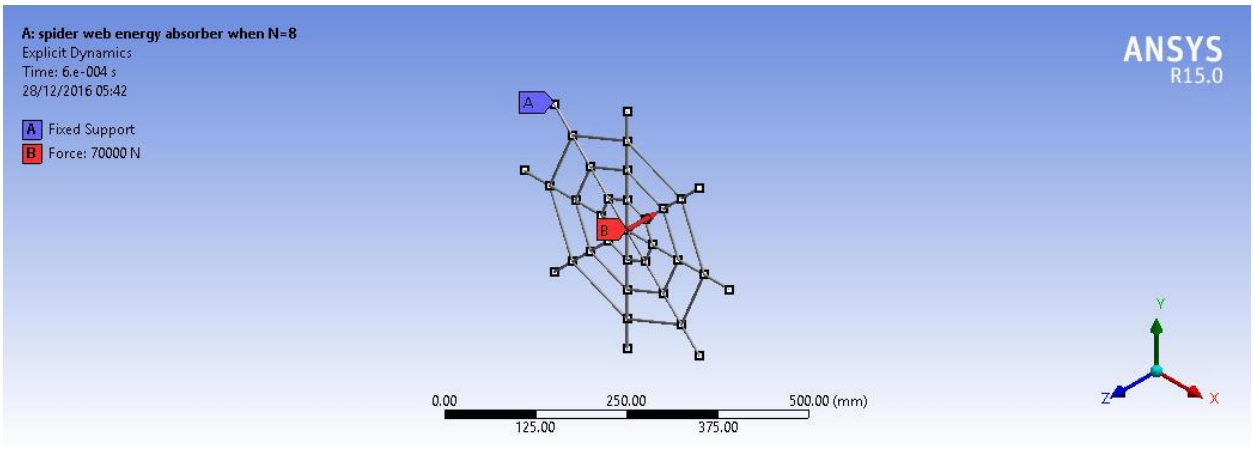


Fig. 3 Showing the direction of the applied force (B) and fixed boundary conditions (A)

The load was varied from zero displacement up to 10% of the overall size of the web structure.

Material properties for the structural steel and the Aluminium alloy are summarised in Table 3 below.

Material property	Structural Steel	Aluminium Alloy
Modulus of elasticity (GPa)	200	71
Poisson's ratio	0.3	0.33
Yield strength (MPa)	250	95
Density (Kg/m^3)	7850	2770
Specific heat (J/kg)	434	875

Table 3. Showing the material properties of the chosen models used for analysis

Fig. 4 below shows one of the typical solution when considering directional deformation at a subjected load of 40000N within a time frame of 60ms. The cross sectional diameter of this absorber is 4mm while the length of the web, the number of web and the number of lobes is 400mm and 8 and 3 respectively. The value of deformation at the above stated conditions is 30.616mm and the corresponding maximum stress is 367.21Mpa.

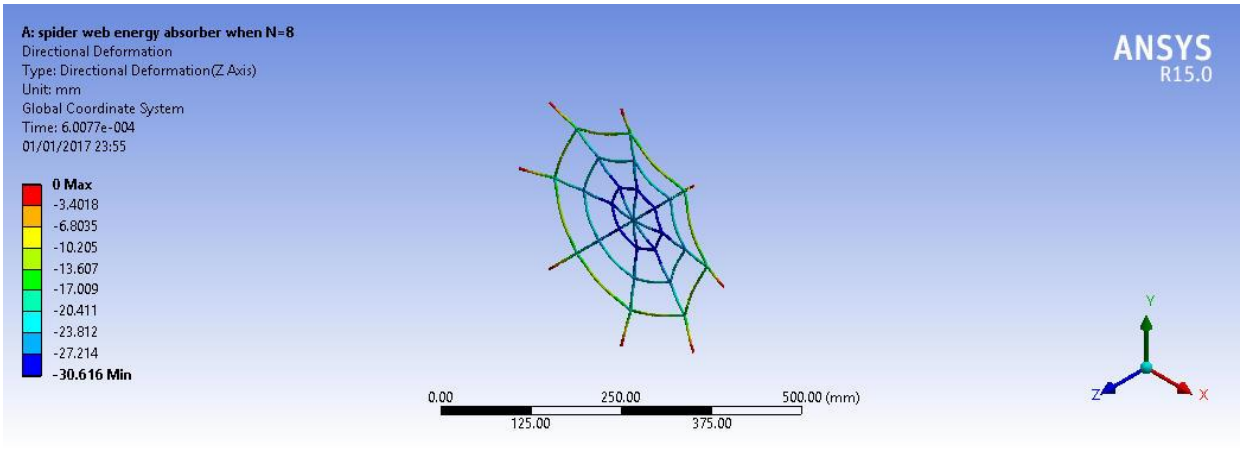


Fig. 4 showing a typical spider web under deformation.

4. RESULT AND DISCUSSION

After subjecting the spider web energy absorber to the conditions stated in the finite element

model analysis, the load versus displacement curve was plotted to show the response of the energy absorber to the applied load. Fig. 5 shows the result for structural steel when the number of web diagonal (N) is varied from 3 to 8 while the circular cross section(D) length (L) of the web and the number of lobes (M) were fixed at 2mm and 400mm respectively.

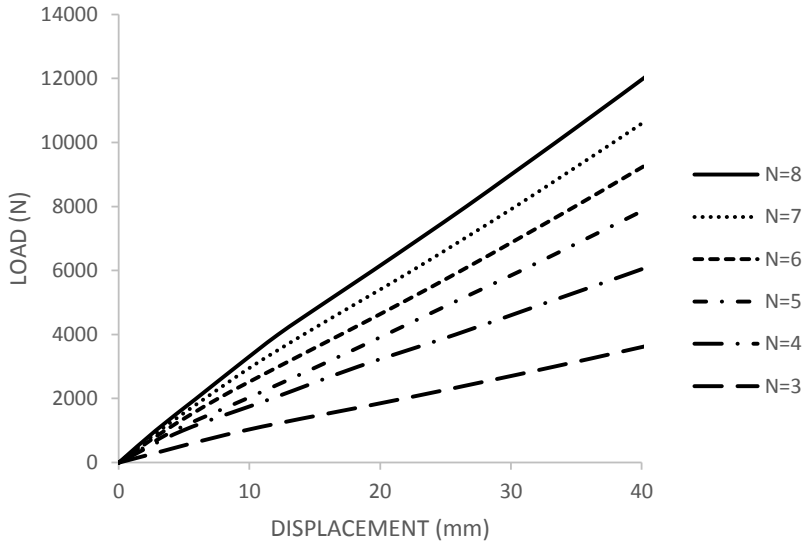


Fig. 5 Load displacement curves of the spider web made of steel at D=2mm, L=400mm, M=3 and N= number of webs.

It can be deduced from Fig. 5 that, as the number of web diagonal increases the load increases due to the increase in web material i.e. the absorption capacity increases thereby having the ability to absorb more impact load at shortest period of time.

Fig. 6, shows the result for structural steel (NL) when the length of the web varies from 200mm to 400mm while the circular cross section (D) and the number of webs (N) are fixed at 2mm and 8 respectively. It shows that as the length of the web decreases at a subjected load there is an approximately 20% average increase in displacement.

Also, Fig. 7 shows the result for structural steel when the circular cross section (D) varies from 2mm to 5mm while the length (L), number of webs (N) and the number of lobes (M) are fixed at 400mm, 8 and 3 respectively. It is an approximately 45% decrease in displacement. This means the absorption capability of the energy absorber increases with increase in cross sectional diameter due to the increase in materials since material is proportional to D.

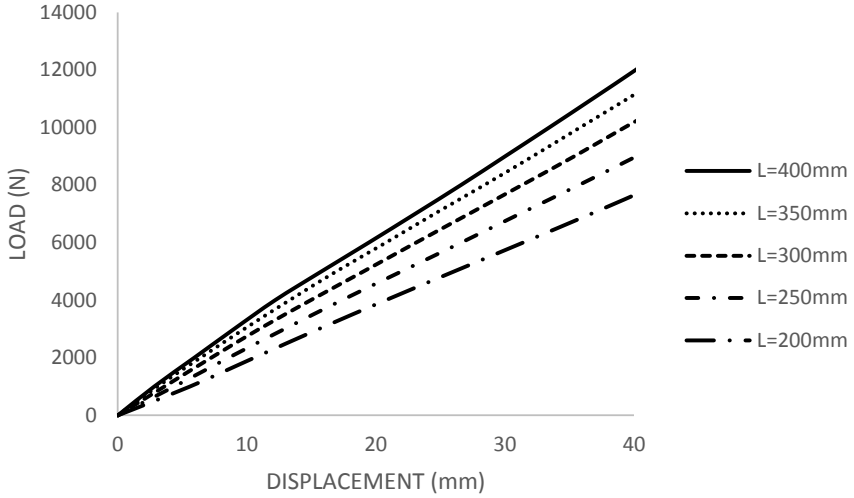


Fig. 6 Load displacement curves of the spider web made of steel at D=2mm, N=8, M=3 and L= length of webs.

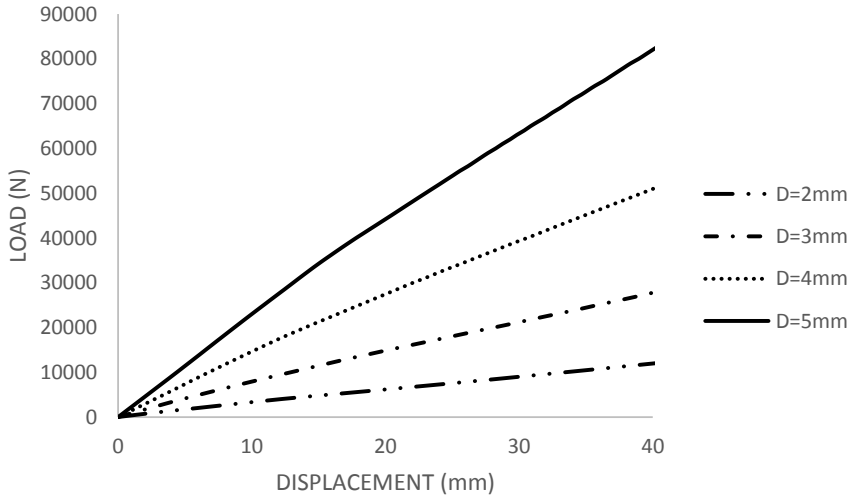


Fig. 7 Load displacement curves of the spider web made of steel at L=400mm, N=8, M=3 and D= circular cross section of webs.

Finally, Fig. 8 shows the two absorber when subjected to the same number of web diagonal (at 8), web length (400mm), circular cross section (2mm) and the number of lobes (3) but two different materials.

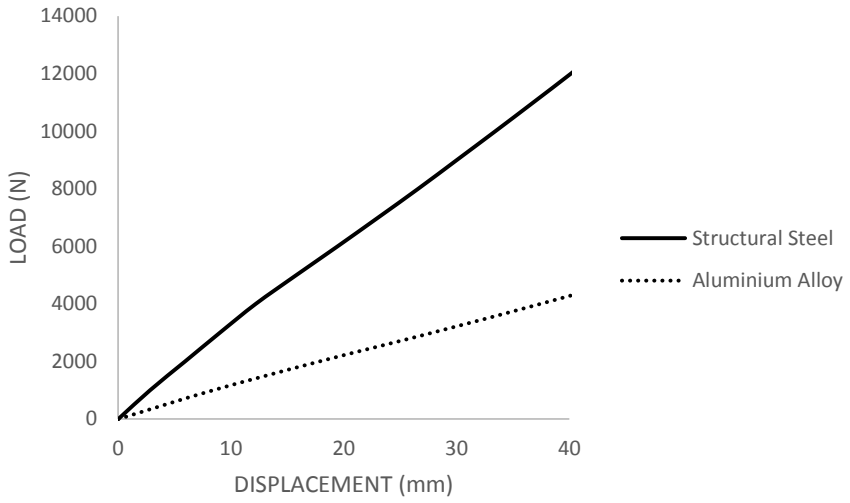


Fig. 8 Load displacement curves of the spider web between aluminium Alloy and steel at L=400mm, N=8, M=3 and D= 2mm.

It can be concluded from Fig. 8 that the absorption capability of structural steel is higher than that of aluminium alloy at the same subjected parameters due to the increase in yielding point.

5. CONCLUSIONS

In this study it can be deduced from the finite element analysis results that the absorption capability of spider web energy absorber is excellent, stable and predictable when crushing. Also, the comparison of absorption ability of structural steel with that of aluminium alloy showed that structural steel has a better absorption capacity than aluminium alloy.

6. ACKNOWLEDGMENT

Support given by King Abdulaziz University through L. Azrar Distinguished Adjunct Professor program is highly appreciated.

REFERENCES

[1] A.A.A. Alghamdi: *Collapsible impact energy absorbers: an overview*. Thin-Walled Structures 39 (2001), pp. 189–213

[2] G M Nagel and D P Thambiratnam: *Dynamic simulation and energy absorption of tapered tubes under impact loading*. International Journal of Crashworthiness (2004), pp. 389-399

[3] A. Othman and Azrol Jailani: *Off-Axis Crushes Simulation of Thin-Walled Tapered Tubes Inserted with Foam-Filled Structures.* APCBEE Procedia 9 (2014), pp. 395 – 400

[4] Muhammed Emin Erdin, Cengiz Baykasoglu and Merve Tunay Cetin: *Quasi-static Axial Crushing Behavior of Thin-walled Circular Aluminum Tubes with Functionally Graded Thickness.* Procedia Engineering 149 (2016), pp. 559 – 565

[5] W. Abramowicz: *Thin-walled structures as impact energy absorbers.* Thin-Walled Structures 41 (2003), pp. 91–107

[6] Sun-Kyu Kim, Kwang-Hee Im, Young-Nam Kim, Jae-Woung Park, In-Young Yang and Tadaharu Adachi: *On the characteristics of Energy Absorption control in Thin-walled members for the use of Vehicular structures.* Key Engineering Materials, Vol. 233-236 (2003), pp. 239-244

[7] Masaru Minagawa, Yuji Doi, and Fumio Nagashima: *Energy Absorbing capacity of Shock Absorbers combining Rubber and Steel Pipes.* Key Engineering Materials, Vol. 233-236 (2003), pp. 251-256

[8] A.A.A. Alghamdi: *Reusable Collapsible impact energy Absorber.* Key Engineering Materials, Vol. 233-236 (2003), pp. 257-262

[9] Dr. Md Razdai Said and Dr. T.Y. Reddy: *The energy absorption of Aluminium Honeycomb under Quasi-Static loading.* 4th International conference on Mechanical Engineering, Dec. 2001, pp. I 35-40

[10] F. Ince, H.S Türkmen, Z. Mecitoğlu, N. Uludağ, I. Durgun, E. Altinok and H. Örenel: *A numerical and experimental study on the impact behaviour of box structures*. Procedia Engineering 10 (2011), pp. 1736–1741

SYMCOMP 2017
Guimarães, 6-7 April 2017

AUTOMATION INSPECTION DIMENSIONAL AND PARTS OF RECOGNITION AND INDUSTRIAL MATERIALS

Charles Luiz S. de Melo[1*], Almir K. Junior[2], and Moises P. Bastos[2]

1: Manaus Institute of Technology – MIT
69088-130
e-mail: cluiz@uea.edu.br, web: http://www.mit-am.org.br/
e-mail: akimurajr@gmail.com

2: University of the State of Amazonas – UEA
Control engineering and automation
69020-000
e-mail: Moises_bastos@hotmail.com web: http://www.uea.edu.br

Keywords: System of computer vision, design of industrial parts, digital image processing techniques

Abstract *Most industrial visual inspection systems still dealing with a low variety of object classes. This is because even the most recognition techniques and representation scheme are not flexible enough to support major changes in the field of objects to recognize and scale. To avoid this type of restriction, our system incorporates digital processing techniques of image, a software able to do all sorts of for recognition and sizing and a fully compatible hardware with all the operational requirem ents of the software. From the analysis of the captured images, our system is able to choose, in a learning phase, the representation scheme best suited for each of the pieces, which facilitates and speeds up the identification and subsequent location of the same in the recognition phase. The advantages of applying these techniques representation were observed in several experiments, both with simulated real images as real-time images. The objective of this article is to describe the implementation of a computer vision system, able to choose the best strategy to recognize measure and locate planar objects, aiming at industrial applications.*

1. INTRODUCTION

According to [1], vision system is a computational system that uses computer vision techniques for various purposes, including inspection in industry usually intended to assurance and quality control. According to [5], consists but vision have automatic equipment composed of cameras, optical elements, appropriate lights on beyond the hardware and processing image software. The goal is to simulate human vision in inspection processes in facture lines. Images are processed and provide information for the system to make decisions and control manufacturing processes or other teammates, such as robots and product separation mechanisms with defects. In Brazil, this technology was introduced in the 80s, using image matching techniques, getting reasonable results, but with little accuracy. Subsequently, in the 90s, processing techniques of feature extraction based images were made by extracting characteristic numerical data existing in the image (e.g. the area of a play, circumference, etc.). There are many applications of vision systems, among which are the inspection of pharmaceutical products, cosmetic, automotive, food and beverage, electronics, graphic, among others. Typically used to control the production and statistical inventory of items produced, allowing identify any items manufactured with some(s) fail(s). Furthermore, the system information can be used to block production, enable correction processes, etc.

In a recent work [6], all this allows managing a line production in real time, facilitating the implementation of a statistical analysis and the automatic storage. As the work of employees who inspect products on a production line, this technology enables, preserving the human activity, to great advantage by not to lose performance in function due to fatigue or distraction. Therefore, it has become essential tool in the production processes by being able to inspect the lines production with high precision, speed, repeatability and consistency.

In a work [8], this technology creates competitive advantages for ensuring quality, increased productivity, preventing recalls, and significant gains in productivity and elimination of waste in the production process, thus reducing costs and improving the image from the company.

According to [9], it noted that the vision systems do not "see" the same Nature's Way as humans. According to [4], vision systems processing pixel images to extract attributes and make decisions based on information provided by human on the quality of the product concerned. Not we intend match the vision systems adaptability and human understanding, although much FAST and accurate. It is essential to be in mind that vision systems are applied where accepts or product failure is not based on subjective or non measurable attributes. In the system we describe below, we focus on their application and functionality on flat parts where the third dimension has not interference in the adoption of part quality.

The objective of this article is to describe the implementation of a computer vision system, able to choose the best strategy to recognize, to dimension and locate planar objects, aiming at industrial applications.

2. METHODS

As the needs of application and functionality offered in our system, which is State for the design of industrial parts, the scene to be captured will consist of pieces placed in arbitrary position in the visual field of two video cameras. These parts must be of planar and because

we are dealing with a theme that does not take into account, for recognition, the three-dimensional shape of the part, but only its silhouette. Image acquisition is made at various levels of colors, making up one after binarization in order to work only two levels of intensity (HSI INTENSITY) used for us to process the image, in Figure 1 shows this picture.

Figure 1. Sample Image Acquisition

For the project development we used the Labview tool for the development of the techniques of digital image processing. For the conversion of image acquisition used if the command shown in Figure 2, as in Figure 3 presents the result of this operation.

Figure 2. 8 bit Binarization step

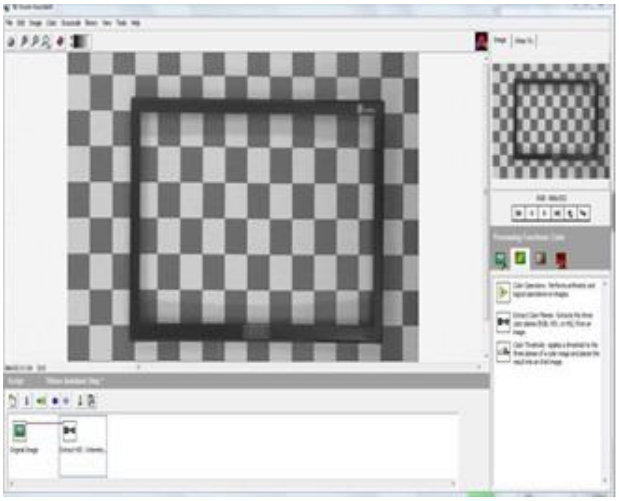


Figure 3. Binarized Image

According to [3], filtering techniques are transformations of pixel by pixel image, which does

not depend only on the gray level of a given pixel, but also the value of the gray levels of neighboring pixels.
The filtering process is done using matrix called masks (Figure 4), which are applied on the image.

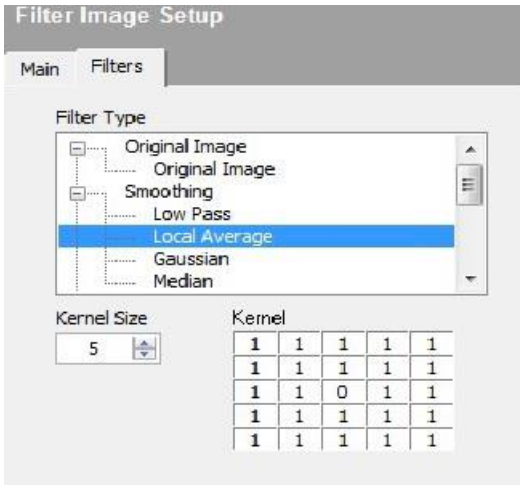


Figure 4. Application of Local Average Filter

So that we can recognize the board, we used scientific and academic techniques of digital processing images, which were the most common, the Gaussian filters, medium location, etc., as shown in Figure 4. According to [1], the local average filter application using a 5x5 matrix, it is necessary for the template material. It was used to remove the salt and pepper, creating the image sharpness.

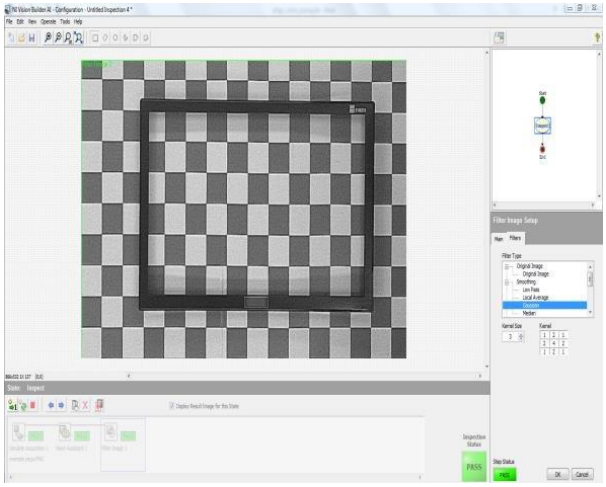


Figure 5. Application of Gaussian Filter

According to [7], the effect of the Gaussian filter is to smooth (smoothing, blur) picture, in much the same way that the median filter (mean filter). The result will be smoother as the higher the standard deviation (standard deviation) of used Gaussian, as in Figure 5 presents the result of this operation.
HighLight filter - according [3], used to enhance the detail and edge of images, to have done the recognition and dimensioning of the material (Figure 6).

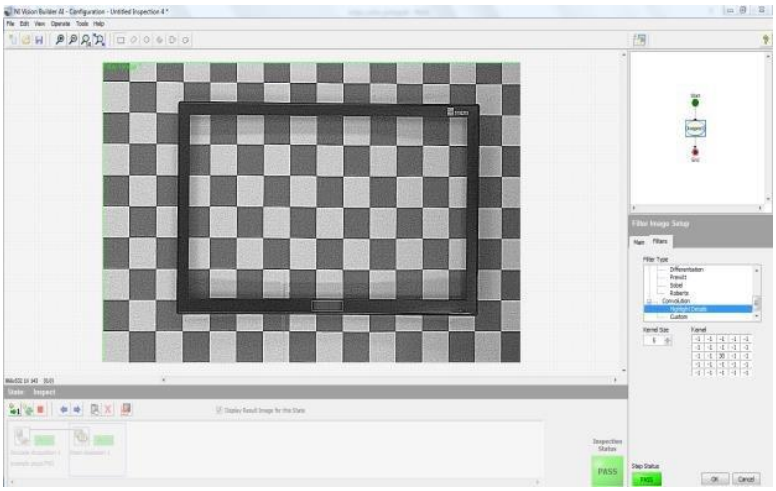


Figure 6. Application of HighLight filter

In general, the recognize patterns comes vision systems with their respective limits, which are entered by the user in the customization. In system operation, beyond the customization step - parameter setting step - are distinguished two phases: a learning phase and a recognition phase. The learning phase is used to create the internal representation of parts to be recognized, creating the bank system parts. In the phase of reknowledge, observed each piece undergoes treatment similar to that of the learning phase, in order to determine their representation; then a search is made in the system of the bank, confront an up of part characteristics in question with those of stored parts, under the same representation, in the learning.
In both phases of operation, one manager selects the technique most appropriate for treating each piece, aiming to maximize process efficiency.
One of the strategies that our saw are computer system used for the recognition and measurement of parts, as already mentioned, is based on contours. In this strategy, after drawn the contours of the image through to using digital processing image, captured generate up chains vectors elementary corresponding to each of them, and from the elementary vectors lists, the manager gets the values of global attributes correct the piece observed, and also the segments, protruding segments are those whose key feature the ability to identify unequivocally a part, its position and orientation. After the treatments done, the next step is the recognition, in which the following will be shown. For the Recognize step used if the command shown in Figure 7.

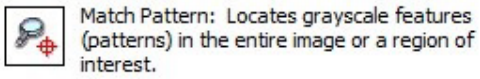


Figure 7. Recognize step

According [9], pattern recognition approaches the feature extraction technique of "objects" present in a digital image. The objects, parameters of the recognition system input are described by separate previously you pixel clusters of the background and will be analyzed for statistically. Every object in a recognition sis theme is described by its chartics or attributes. These characteristics are represented in a N dimensional space, where N is the number of features. Each object, but the way, within this vector space features, called feature vector.
To obtain the protruding segments in both the learning and recognition, the group part is divided into segments of a fixed comparing. Then it makes a transformation of the representation of contours of the parts of space (x, y) of pixels in mm using the calibration of the system software. The images are compared with those of stored pieces (standard parts) under this representation, the system of stock photography (Figure 8).

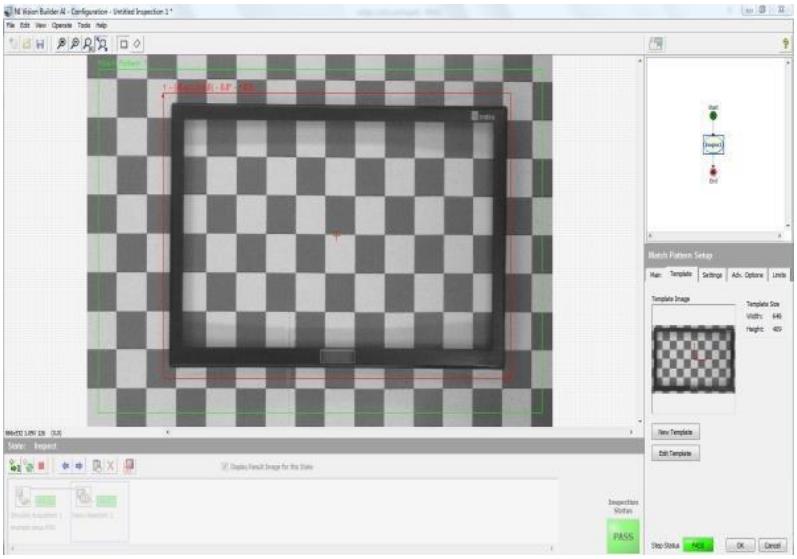


Figure 8. Image Recognize

In the representation via contours, the recognized will occur when there is a coincidence of the values of global attributes part observed with those of a given piece of system seat, standard piece as mentioned above, within the x-fair tolerance previously established by the general embracing specification and standard parts. Before the realization of mediction, Calibration is performed at the image, which pixels cross shape in real dimensions, which is used by the measurement function (Figure 9 and Figure 10).

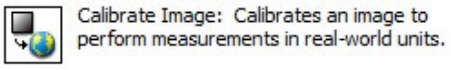


Figure 9. Calibration Step

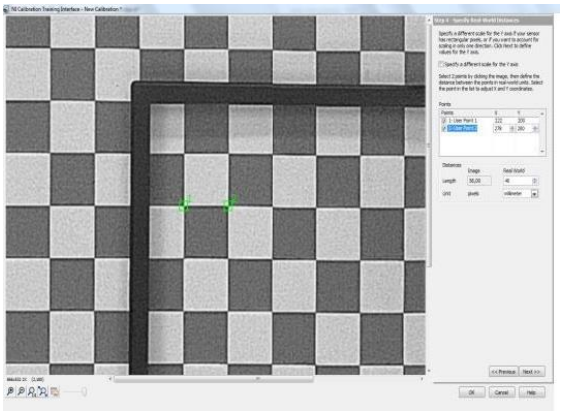


Figure 10. Image Calibration

The Caliper function (Figure 11) is the step that performs Inspection sizing, reiterating that everything is realized by comparison techniques in which is measured by the standard set by the software piece and inspection of tolerance for each material.

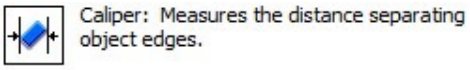


Figure 11. Automatic Caliper Step

The following will give some information about operation of the system: In general, the vision system is typically constituted two cameras, colored 4k resolutions with two 5 mm lens to the correct conditioning of the image.

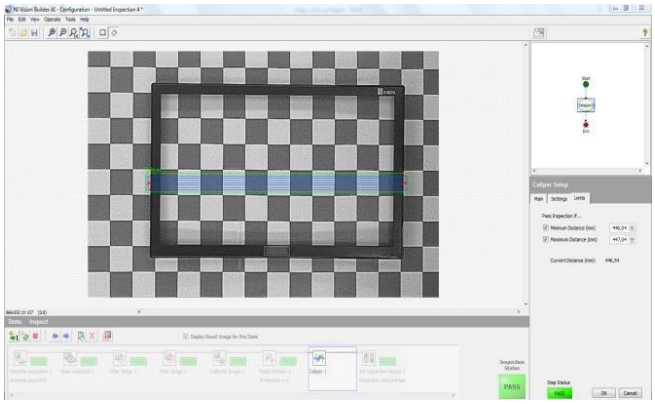


Figure 12. Calculation Lenses

The following will be described the calculation of how it's gone taught the best lens to match the cameras (Figure 12).

3. RESULTS

The lighting has to be positioned correctly, so as not to highlight the parts, preventing the contrast while preserving only the wishes of attributes. The transmission interface/scan images (known as "framegrabber"), one can use a USB transmission interface. As for processing, is based on a process high capacity of a desktop i7, LABVIEW software to process the images and detect the relevant features through visual treatments.

The system should be operated via a graphic interface with the user. When the installation in an industrial environment, the ultimate goal of system should be detect possible failures in parts. For this, you have two cameras for image acquisition of parts can this be color or black/white. For moving parts, have a neutral color ha mat and chess design in order to obtain greater contrast with the parts and facilitate the processing of images with illumination to facilitate processing. Making the analysis of the histogram and starting the extraction of part features in order to decide, through the system manager, what better way to re-knowledge.

The system must control the correct positioning (orientation) of the pieces on the e frontier, in the production line output. However, varies placement are fully tolerated, if the camera can capture the part of the image as a whole. The belt will be controlled through of a Programmable Logic Control (PLC).

Made the acquisition of the images for learning all the parts of interest, the user will have to provide some parameters to be used both in the learning phase and in recognition. Such off-line procedures are fundamental to the operation of our system.

It is important to note that all parameters, is to be adapted to the requirements of the degree of similarity between the pieces and storage of payload data. In addition to the above parameters, each piece has associated with it an identification that is supplied by the user in the learning phase. This identification may be associated with the part name engraved on a file. Just as in the learning phase, the models used in the system recognition phase are formed taking the parameters provided in the boot.

Therefore, the entire processing consists use the same set of parameters. There will need to register the system there is any change in the parameters, for example, to include new pieces to your bank parts. The following have been the images of the results with an example of PASS and FAIL of the same model COVER REAR a 32 "TV (Figure 13 and Figure 14).

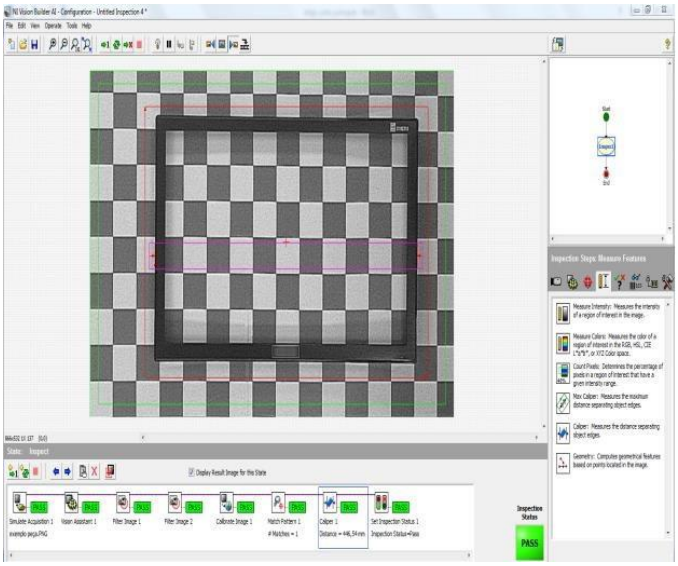


Figure 13. Test Result - PASS

3. CONCLUSÃO

In this paper, we describe the implementation of a two-dimensional binary computer vision system that incorporates multiple methods of scaling and inspection form, with a view to application in industrial environment. Our proposition to be to develop a system capable of recognize and locate a wide range of parts, using for this both their structural attributes, as its contours. The choice of technique will be given through the analysis of global attributes extract the image, and this method the big difference in our work.

The system was tested early in a bank of synthetic images, obtained via software, and later successfully employed in identifying and locating real parts from images captured by a camera of high resolution video accompanied by ultra low lenses distortion. It has worked to modernize and expand the flexibility of our system with large performance gains in an industrial environment.

We achieved great success, on the research and in development. The system was implemented and is working in the quality department of a large multinational company.

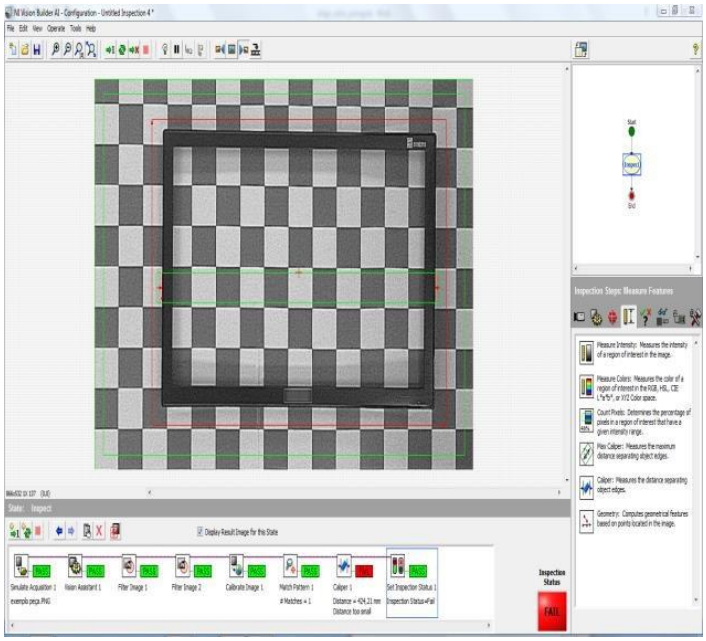


Figure 14. Test Result - FAIL

REFERENCES

[1] Abend, K., Harley, T. J., 1965. Classification of Binary Random Patterns, *IEEE Transactions on Information Theory*: 538-544.

[2] Asano, T., 1996. Digital Halftoning Algorithm Based on Random Space Filling Curve, *IEEE International Conference on Image Processing*, Vol 1, Lausanne, Sua, pp. 545-548.

[3] Almeida, L.B., 1994. The Fractional Fourier Transform and Time Frequency Representation, *IEEE Transactions on Signal Processing*: 3084-3091.

[4] Bunke, H., Sanfeliu, A., 1990. Syntactic and Structural Pattern Recognition: Theory and Applications.

[5] Evans, D. M. W., A Novel Algorithm for Computing the 1D Discrete Hartley Transform, *IEEE Signal Processing Letters*: 156-159.

[6] Faugeras, O., Luong, Q.T., 2001.The Geometry o Multiple Image: The Laws that Govern the Formation of Multiple Images of a Scene and some of Their Applications, MIT Press, Cambridge, MA, Estados Unidos.

[7] Nehme, D., Luis, E., 2000. Processa- mento Digital e suas aplicaes.

[8] Tanaka, E., 1995. Theoretical Aspects of Syntact Pattern Recognition, *Pattern Recognition*: 1053-1061.

[9] Reisfield, D.,Wolfson, H., Yeshurun, Y., 1995. Context Free Attentional Operators: *The Generalized Symmetry Transform, International Journey of Computer Vision* 14(2) 119-130.

[10] Zhang, D., Lu, G., 2002. A Comparative Study of Fourier Descriptors For Shape Representations and Retrieval, Proceedings Fifth Asian Conference on Computer Vision, Melborn, Austrlia,pp. 646-651.

COMPARISON BETWEEN 3D LASER SCANNING AND COMPUTED TOMOGRAPHY ON THE MODELLING OF HEAD SURFACE

Sousa, E[1,2]; Vieira, L[1,2,3]; Costa, DMS[2]; Costa DC[4]; Parafita R[4,5]; Loja, MAR[2,6,7]

1: ESTeSL - Escola Superior de Tecnologia da Saúde de Lisboa, Av. D. João II, Lote 4.69.01, 1990-096 Lisboa, Portugal.
2: GI-MOSM, ADEM, ISEL – Grupo de Investigação em Modelação e Optimização de Sistemas Multifuncionais, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal.
3: Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal.
4: Champalimaud Centre for the Unknown, Champalimaud Foundation.
5: Mercurius Health, Rua Braamcamp, 12, 3º E, 1250-050 Lisboa, Portugal.
6: ISEL - Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal.
7: LAETA, IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais, 1, 1049-001 Lisboa, Portugal.
e-mails: {eva.sousa@estesl.ipl.pt ; lina.vieira@estesl.ipl.pt ; dcosta@dem.isel.pt ; amelialoja@dem.isel.ipl.pt}

Keywords: 3D Laser Scanning, Computed Tomography, Modelling of Head Surface, Statistical Assessment of the Models

Abstract

The measurement of people physical parameters and proportions constitutes an important field of science, the anthropometry, since it is related to the characterization of the human size and constitution; it allows improving the design and sizing of systems and devices to human use.

To enable these measurements, different direct and indirect methodologies may be used depending on the particular aim of a specific study and on the eventual availability of data sources that can be used also for this purpose.

Because of this relevance, the present work intends to assess the influence of different acquisition and reconstruction methods in the modelling of a 3D head surface.

In order to assess the significance of the differences between acquisition and reconstruction methods a set of measurements between several anatomic references of a physical phantom were carried out. Statistical evaluation using the Friedman test for non-parametrical pared samples was considered.

We found, so far, no statistically significant differences between the several methods considered for acquisition and reconstruction.

INTRODUCTION

The measurement of people physical parameters and proportions constitutes an important field of science, as it enables to the understanding of anthropological information associated to human remains, a better acquaintance of our ancestors' heritage and, not least importantly, the improvement in design and sizing of systems and devices to human use.

Nowadays several approaches, ranging from low cost scanners to professional scanners, exist to reconstruct a precise 3D model of a human body. This problem is especially complex in non-rigid applications and to use these models in such applications, approaches to extract anthropometric data from human body scans are required. The extracted measures build the basis for further processing and thus automatic and reliable approaches are important. The literature proposes semi-automatic and automatic approaches like detecting landmarks and in its precise localization, although other methods recurring to new developed algorithms and prior data enable the modelling 3D surface without recurring to landmarks as done by Anguelov et al [1].

Within the methods to access surfaces one can use Scanners or Computed Tomography (CT) technique, being CT the gold standard, as it constitutes currently the main imaging technique to access surface for radiotherapy. Illustrating this we may refer the work developed by Gopan and Wu[2], and by Morton [3], who focused on the accuracy issues related to this imaging technique. The evaluation of spatial resolution of the applied imaging systems is essential for accurate and precise measurements. The advent of more complex imaging systems has posed an increasing challenge in our ability to assess their performance, including spatial resolution. Richard et al. [4]addressed the problem associated to iterative and statistical reconstructions which can exhibit nonlinear signal characteristics, which can affect system resolution properties differently than with standard reconstruction algorithms, and acquisition techniques.

Contrarily to CT, made of a sequence of bi-dimensional images, 3D Laser Scanning (3DLS) provide a three-dimensional sampling of an object, namely of the human body, characterizing it by a dense set of points in the 3D space, usually known as point's cloud. The topology of these points can be represented, for instance, through the constitution of a polygonal mesh or via parametric surfaces. It is noteworthy that the measurements obtained represent the human body at a certain time and that various scans may lead to slightly different results, as pointed by Paquet and Rioux[5].

Recent advances in three-dimensional scanning technology have enabled the generation of high-density point data sets to describe the surfaces of real objects (Bernardo and Loja [6][7]), including animate objects such as the human body. This means that it is now possible to produce computer-based models that describe in detail the topology and the geometry of an actual human body. Such models can be used to perform fast and accurate automatic measurements. However, such measurements cannot be made on raw point data, as very often the point's cloud is noisy and contains undesired information. To minimize these problems Douros et al. [8]proposed an algorithm to analyze the data and from it infer the topology (and subsequently the geometry), taking into consideration that skin surface presents further challenges, such as more accurate reconstruction algorithms.

With the present study the authors propose to evaluate and compare the CT images of craniofacial anthropometry of the PIXY phantom and the points cloud obtained via 3D laser scanning. These features are further compared to caliper measurements.

1. METHODOLOGY

As mentioned, the main aim of the present work is the comparison of different anthropometric head measurements by using different data acquisition sources, in order to assess from their significance so one can establish a viable departing point to an automated design of head devices.

To this purpose, one has considered the head of a physical phantom (PIXY) [9]and performed a set of direct and indirect measurements methods. The first method considered a direct, manual measurement, by using a Rosscraft caliper with the scale of 0-180mm and a precision of 0.1 mm mainly utilized in nutrition.

Two other approaches were carried by considering the data acquired via 3D laser scanning (3DLS) and via computerized tomography (CT).

To perform the 3D laser scanning of the phantom head, one has used a NextEngine system, which settings involved a medium range acquisition and a number of ten sectors with a density of 850 points per square inch. The system precision is 0.005 inch. With the 3D point cloud acquired, a subsequent mesh was generated in order to yield the surface reconstruction. This reconstruction stage was carried out automatically by the same system.

The two computed tomography were performed using a system Philips Gemini TF 16 with two different slices' thickness (1 mm and 2mm) while keeping other acquisition parameters constant: 20 mAs, 120kVp, pitch of 0.938, 512x512 pixel matrix, rotation speed of 0.5s per rotation. It is important to say that these settings are the usually considered in a normal planning CT, in the Nuclear Medicine scientific area.

The CT images were next used to build the 3D head mesh, by using two alternative image processing applications, ImageJ and Osyrix. The meshes were also obtained by different members of the research team.

1.1. 3D Reconstructions

The meshes built, were subsequently imported layer by layer, to a trial version of Geomagic Design, and converted into solid entities, enabling to obtain the required measurements. Figure 1 illustrates the solids corresponding to the 3DLS and CT acquisition methods.

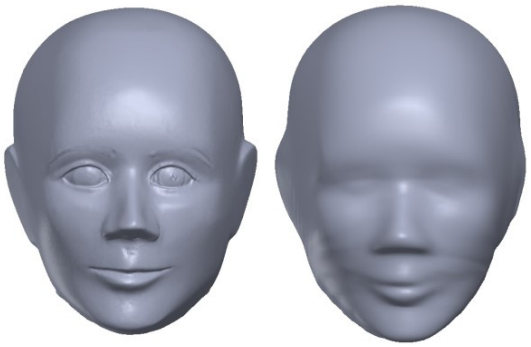


Figure 1: Solid representation of 3DLS (left) and CT (right) data.

From Figure 1 it is possible to conclude that for the settings used, there exists an evident difference on the level of detail obtained using 3DLS and CT.
The measurements carried out to characterize the anthropometry features of the physical phantom head, were based on the set of measurements carried out by Yokota [10]. The results obtained are presented in Table 1.

Table 1: Identification of craniofacial measurements [10].

Measure ID	Designation	Definition
1. BIGONIAL	Bigonial breadth	Straight-line between the right and left gonion on the jaw
2. BIOCBRMH	Biocular breadth	Distance between the right and left ectoorbitale
3. BIZBDTH	Bizygomatic breadth	Maximum horizontal breadth between zygomatic arches
4. CHINPROJ	Chin projection	XYZ coordinates between right tragion and mentona
5. HEADBRTH	Head breadth	Maximum horizontal breadth of the head
6. HEADLGTH	Head length	Maximum distance between the glabella and back of the head
7. LIPLGTH	Lip length	Distance between the right and left cheilion on the corner of the mouth
8. MENSUBNH	Menton–subnasal	Distance between the menton and the subnasal
9. NOSEBRTH	Nose breadth	Distance between the right and left alare
10. RTRAGX	Right tragion X	Distance between right tragion and back of the head plane
11. SBNSSELH	Subnasal-sellion length	Distance between the subnasal and sellion
12. SELLIONZ	Sellion Z	Distance between sellion and top of the head plane
13. SELTRAG	Facial projection	Distance in XYZ coordinates between sellion and right tragion

A better illustration of these measurements' meaning is shown in Figure 2 that presents a schematic representation extracted from Yokota[10] although in the present case, the physical phantom doesn't possess that much detail.

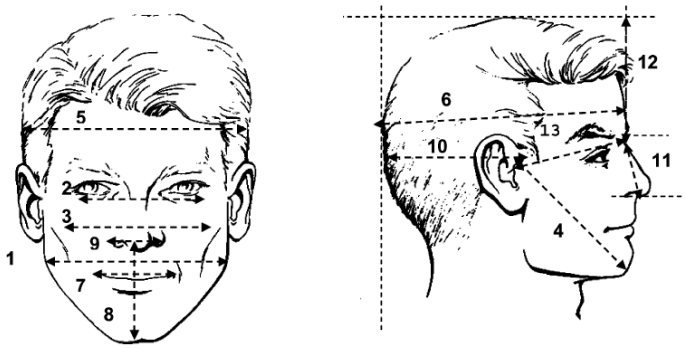


Figure 2: Schematic representation of craniofacial measurements. (Source: Yokota [10][10]).

To guarantee the maximum accuracy and consistency among the different phantom head reconstructions, all the four meshes were aligned, considering an approach based on the minimum global deviation among meshes. After this alignment, and after the selection of the points of interest, the measures were taken at each active layer, corresponding to a specific head model.

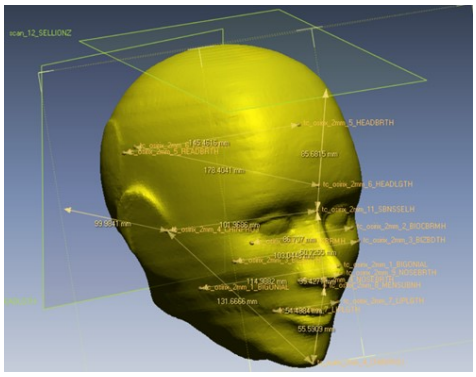


Figure 3: Measures extraction in 3DLS model.

This procedure allowed obtaining model by model the required measures. Figure 3 illustrates the extraction of some measures in the 3DLS model.
In the preliminary comparison of the measurements carried out to characterize the anthropometry features of the physical phantom head, one has calculated different relative deviations. These relative deviations were plotted using radar charts.

1.2. Statistical assessment of the models

As at that moment it wasn't intended to assume a single mesh as the golden standard, then one has assumed that the CT 1mm and 3DLS may assume that role.
Statistical inferential models were also applied in the comparison of the measurements executed in the different acquisitions (1 mm or 2 mm slices thickness), and of the different

imaging reconstruction applied methods, and also in the comparison of these with the physical measurements obtained with calliper. In order to execute further inferential statistical evaluation, it is necessary to investigate previously the distribution of the sample and classify that distribution as a normal on non-normal one. Due to the small size of the sample it is necessary to apply Shapiro-Wilk test to enable classifying the distribution of the samples obtained. Furthermore, in a subsequent phase, to evaluate if there were any statistical significant differences between the measurements collected of the same landmarks (when the phantom is reconstructed and/or acquired with different settings) is was applied Friedman for the multiple comparisons.
All statistical analysis was performed using Statistical Package for the Social Sciences (SPSS v.24) software. For all the groups included in the study, a significance level of 5% was used [11].

2. RESULTS AND DISCUSSION

In order to simplify the designation of the different reconstructions performed, the following acronyms and corresponding meaning were used: CTImg1 – ImageJ CT reconstruction, 1 mm slice; CTImg2 – ImageJ CT reconstruction, 2 mm slice; CTOsx1 – Osyrix CT reconstruction, 1 mm slice; CTOsx2 – Osyrix CT reconstruction, 2 mm slice; 3DLS – 3D Laser scanning.
The results here summarized include the measurements carried out using the different methodologies, their deviations and the further statistical analysis to understand whether there are statistically significant differences among the methodologies.

2.1. Craniofacial measurements

After the preliminary meshes alignment stage, one has proceeded to the measurements, which task was developed. Table 2 presents these results, and also the calliper direct measures.

Table 2: Measures obtained via different methods.

Measure ID	Caliper [mm]	CTImg1 [mm]	CTimg2 [mm]	CTOsx1 [mm]	CTOsx2 [mm]	3DLS [mm]
1. BIGONIAL	110	116.18	112.77	113.54	114.99	116.85
2. BIOCBRMH	85	86.57	86.64	86.71	86.71	86.69
3. BIZBDTH	100	104.11	100.78	109.84	103.04	106.34
4. CHINPROJ	130	132.65	133.57	132.51	131.67	132.30
5. HEADBRTH	110	146.17	140.95	144.78	145.46	145.45
6. HEADLGTH	170	179.45	177.16	177.67	178.40	178.14
7. LIPLGTH	55	54.46	54.66	54.48	54.49	54.51
8. MENSUBNH	50	54.85	54.05	55.19	55.59	55.11
9. NOSEBRTH	36	36.22	33.47	35.94	35.43	35.51
10. RTRAGX	85	99.85	99.78	99.80	99.98	99.93
11. SBNSSELH	50	50.45	50.33	50.36	50.73	50.72
12. SELLIONZ	120	85.69	85.73	85.70	85.68	85.69
13. SELTRAG	75	102.76	102.96	102.99	101.97	102.30

In Table 2, it is clear that ID measures 3 and 5 are the ones where the values obtained are more dissimilar. To allow a better perception of the relative deviations among different methods using different reference reconstructions, Figures 4-6 allow us to understand those deviations when considering different reference models. The relative deviations, when one considers the CTImg1 model as reference is presented in Figure 4.

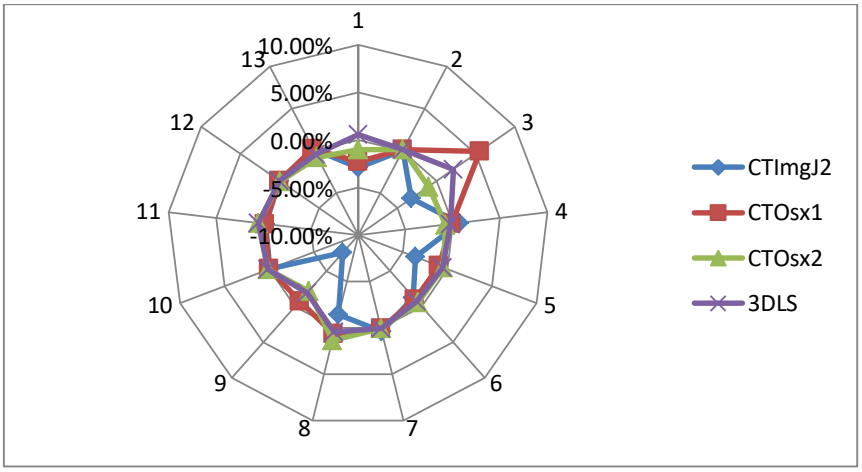


Figure 4: Relative deviations using CTImg1 as the reference measurement model.

In addition, the measures ID 3, 5 and 9 show the greater deviations than all the others. The CTImg2 and CTOsx1 models are the ones where the major values occur, when the golden standard is considered to be the CTImg1. Contrarily, the CTOsx2 and the 3DLS are the ones that perform more accordingly the standard, with minor deviations for the latter one.

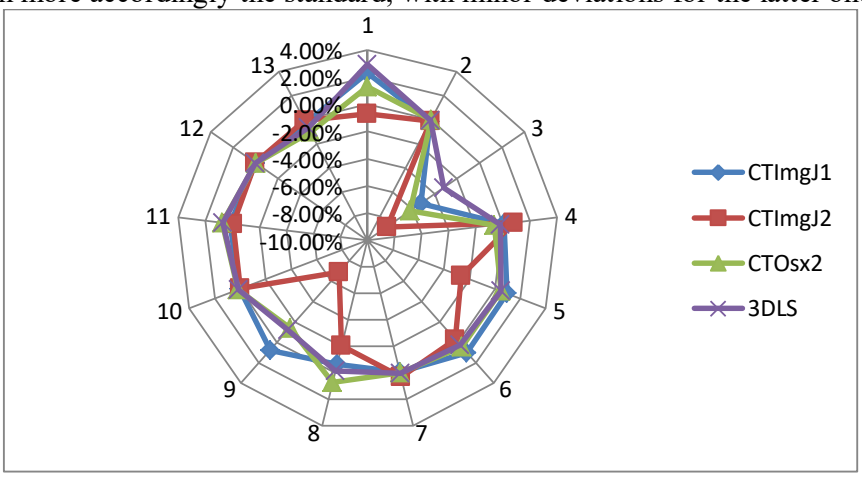


Figure 5: Relative deviations using CTOsx1 as the reference measurement model.

Considering now the CTOsx1 as the reference model, the relative deviations presented by the different models, are represented in Figure 5.
Figure 5, shows again the CTImg2 model as being the one where the major deviation values occur, and this happens especially for the ID measures 3, 5 and 9. The approach that provides a closer set of measurements when compared to the CTOsx1 reference is the 3DLS model.
Finally and considering the laser scanning results, Figure 6 depicts the deviations presented by the different CT reconstructions taking the 3DLS as the standard.

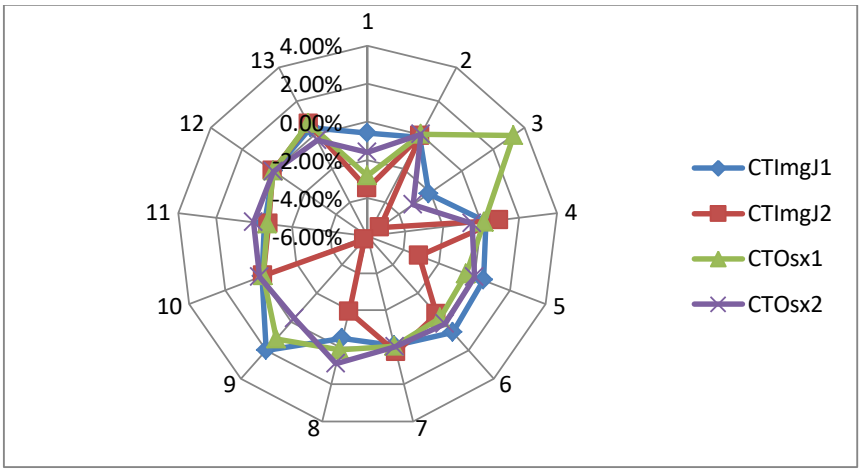


Figure 6: Relative deviations using 3DLS as the reference measurement model.

This last deviations' comparison presented in Figure 6, demonstrates the CTImg1 to be the one that is globally closer to the 3DLS standard. The CTOsx1 is also closer if one excludes the measures 3 and 9.

2.2. Statistical assessment results

After the comparison of landmarks measurements between the different acquisition and reconstructed methods by Friedman test, including all the reconstructed imaging for all methods and thickness of the slices, the different acquisition methodology such as CT and laser technology and comparing this with physical measurements there were found significantly statistical differences (p=0.000).
These differences were probably due to the different level of precision and accuracy that were obtained of all the other techniques when compared with the physical measurements obtained with calliper. Although it is possible to be certain about the reference points for measurements executed in all the reconstructed images, and all were measured in the same software, the same is not possible when doing the physical measurements, and this can introduce some inaccuracies.
After the comparison of all the reconstructed images measurements, there were not found any statistically significant differences, in the together comparison of the methods, allowing us to

conclude that al methods from a statistical point of view can be used as suitable for the reconstruction process (p=0.414), even when it is not applied the same thickness of slices in the acquisition protocol of CT (1mm and 2 mm), and when the acquisition technique is different (CT and 3DLS).
When maintaining the thickness of the slices and comparing the CT with 3DLS there were not verified statistically significant differences, neither for the comparison of CT slices acquired with 1 mm neither in the comparison between CTImg1, CTOsx1 and 3DLS (p=0.981); neither in the comparison of CTImg2, CTOsx2 and 3DLS (p=0.232). Despite the non-existence of statically significant differences the results show a greater concordance between 3DLS and the both reconstruction models when the CT was acquired with 1 mm slices, this can be due to the effect of the increase of the slices thickness in the degradation of the imaging spatial resolution and overall quality of the CT imaging [12].
It is also possible to conclude that even in lower spatial resolution CT acquisition, statistically significant differences are not observed when one uses different reconstruction methods.

3. CONCLUSIONS

There are no statistically significant differences between the several methods of acquisition and reconstruction considered in this work.
Therefore, and from a statistical point of view, all models used, can be considered as suitable for the reconstruction process and subsequent measurement and head anthropometry characterization.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the financial support of Project IPL, IDI&CA/SOFTIMOB, and Project LAETA—UID/EMS/50022/2013. The authors also wish to acknowledge Fundação Champalimaud the possibility of obtaining the CT images which were essential to this study.

REFERENCE

[1] Anguelov D; Srinivasan P, Pang H-C, Koller D, Thrun S, Davis J. “The Correlated Correspondence Algorithm for Unsupervised Registration of Nonrigid Surfaces”. Stanford AI Lab Technical Report SAIL2004-100.
[2] Gopan O, Wu Q. “Evaluation of the accuracy of a 3D surface imaging system for patient setup in head and neck cancer radiotherapy”. *Int J Radiat Oncol Biol Phys,* Vol.84, pp. 547-52, 2012.
[3] Morton AM. “Validation of 3D Surface MeasurementsUsing Computed Tomography”. Queen’s University, thesis for Master in Science for School of Computing, 2011.
[4] Richard S, Husarik DB, Yadava G, Murphy SN, Samei E. “Towards task-based assessment of CT performance: System and object MTF across different reconstruction

algorithms". *Med. Phys*. Vol.39(7), 4115-22, 2012.
[5] Paquet E, Rioux M. "Antropometric visual data mining: A ContentBased Approach". *NRC Publications Archive*, 2003.
[6] Bernardo GMS, Loja MAR. "Reconstruction and Analysis of Hybrid Composite Shells using Meshless Methods". *International Journal of Advanced Structural Engineering,* 2017, DOI:10.1007/s40091-017-0152-2.
[7] Bernardo GMS, Rodrigues JA, Loja MAR. "Towards an Expeditious as-is Surface Reconstruction". *Engineering Structures*, Vol.129, pp. 91-107, 2016.
[8] Douros I, Dekker L, Buxton BF. "An Improved Algorithm for Reconstruction of the Surface of the Human Body from 3D Scanner Data Using Local B-spline patches. Modelling People". *Proceedings. IEEE International Workshop*, pp. 29-36, 1999.
[9] A. Phantoms, "PIXY The Anthropomorphic Training / Teaching Phantom." pp. 1–3.
[10] Yokota M. "Head and facial anthropometryof mixed-race US Armymale soldiers for militarydesign and sizing: A pilot study". *Applied Ergonomics*, Vol. 36, pp. 379-383, 2005.
[11] Alexandre P, "*SPSS Guia Prático de Utilização Análise de Dados Para Ciências Sociais e Psicologia"*, Edições Sí. 2006.
[12] Lee W. G., "Principles of CT: Radiation Dose and Image Quality," *J Nucl Med Technol*, 2007.

ASSESSING PARAMETRIC UNCERTAINTY ON FIBRE REINFORCED COMPOSITE LAMINATES

A. Carvalho[1,2,3], T.A.N. Silva[2,4], M.A.R. Loja[1,2,5]

[1] ISEL - Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal.
[2] GI-MOSM, ADEM, ISEL – Grupo de Investigação em Modelação e Optimização de Sistemas Multifuncionais
[3] CEMAPRE, ISEG, Universidade de Lisboa, Rua do Quelhas, 6, 1200-781 Lisboa, Portugal.
[4] NOVA UNIDEMI, Faculdade de Ciência e Tecnologia, Universidade Nova de Lisboa, Campus de Caparica, 2829-516 Caparica, Portugal.
[5] LAETA, IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais, 1, 1049-001 Lisboa, Portugal.

e-mails: acarvalho@adm.isel.pt ; tan.silva@fct.unl.pt ; amelialoja@dem.isel.ipl.pt

Keywords: Parametric Uncertainty Characterization, Composite Laminates, Variability of Composite Behaviour

Abstract

When modeling a composite structure it is important to take into account its greater exposition to parametric variability, when compared to other types of materials more traditionally used. The possibility to tailor composite materials according to specific requisites is simultaneously a source of additional variability, which origin may be associated to material and to geometrical characteristics.

Regardless the origin of this variability, they will produce its effect in the structure response, thus it is very important to anticipate them and to quantify them as much as possible.

With the present work, it is intended to assess and quantify the influence of geometrical parameters variabilities on the composite structural response. Behind this characterization, linear static analyses were performed, considering that the layers' thicknesses and fibre orientation angles will be affected by uncertainty. A set of simulations and numerical results are presented and discussed.

1. INTRODUCTION

The growth verified in composites usage, in a global perspective, may be mainly attributed to the transportation and construction industries, although other areas such as the medical and health technologies are also becoming to arise. Within those, the manufacturing processes that are witnessing a higher development resin transfer molding and glass mat thermoplastics as well as long fibre-reinforced thermoplastics. According to a market research mentioned in the report for 2017 concerning the Composites Industry in US [1], this latter has grown 25 times, whereas the aluminium industry was 3 times bigger and the steel industry only grown 1.5 times. These numbers denote an important reality landscape on the increasing use of composite materials, thus confirming a continuous need for a deeper research to a more complete understanding of these materials.

The need for stronger materials is already leading to the development of glass – most often used reinforcement – with 2-3 times higher tensile strength to meet operation requirements such as the ones posed by wind blades, bicycle frames and diverse automotive and aerospace parts. Simultaneously, lightweight materials can become very attractive as they may simultaneously meet regulatory requirements for emission reduction, fuel economy and safety, if one considers for example the automotive and aerospace industries, and in such cases, carbon fibre reinforced polymers has been the primary beneficiary. However, the cost of carbon fibre still constitutes a disadvantage and the materials are not fully recyclable at the end of its life cycle.

The use of composites in the most diverse areas, thus pose different questions depending on the nature of the specific applications. Moreover, being materials with greater intrinsic constitution heterogeneity, and considering their manufacturing processes, it is expected to encounter a greater uncertainty when compared to homogenous traditional materials as metals for example.

Several published works can be found, attempting to characterize this uncertainty, using different approaches. Mesogitisa et al. [2] presented a review work about the multiple sources of uncertainty associated with material properties variation and boundary conditions variability. In their study, they presented experimental and numerical results concerning the statistical characterization and the influence of inputs variability on the main steps of composites manufacturing including process-induced defects.

In the context of more specifically focused works we may refer the one due to Noor et al. [3] which proposed a two-phase approach and a computational procedure for the prediction of the variability in the nonlinear response of composite structures associated with variations in the geometric and material parameters of the structure. To this aim, they considered in a first phase, a hierarchical sensitivity analysis, in order to identify the parameters that have a greater influence on the response quantities of interest. In the second phase, these parameters are taken to be fuzzy parameters, and a fuzzy set analysis is used to determine the range of variation of the response.

The problem of uncertainty propagation in composite laminate structures was studied by António and Hoffbauer [4]. They considered an approach based on the optimal design of composite structures to achieve a target reliability level. Using the Uniform Design Method

(UDM) a set of design points is generated over a design domain centred at mean values of random variables, aimed at studying the space variability. The most critical Tsai number, the structural reliability index and the sensitivities are obtained for each UDM design point. Then using those design points as input/output patterns, an Artificial Neural Network (ANN) is developed based on supervised evolutionary learning. That network is used to implement a Monte Carlo simulation procedure in order to obtain the variability of the structural response based on global sensitivity analysis. The use of Artificial Neural Networks was also considered by Teimouri et al. [5] to investigate the potential impact of manufacturing uncertainty on the robustness of commonly used Artificial Neural Networks (ANN) in Structural Health Monitoring (SHM) of composite structures, namely concerning to the thickness variation in laminate plies. An ANN SHM system was assessed through an airfoil case study based on the sensitivity of delamination location and size predictions, when the ANN is imposed to noisy input.

Mukherjee et al. [6] studied the influence of material uncertainties in failure strength and reliability analysis for single ply and cross ply laminated composite subjected to only axial loading. They have categorized the uncertainty at different scales, although in this study they only consider ply level uncertainty. These uncertainties are taken as basic random variables and the strength parameters of the composite are derived through uncertainty propagation considering both Tsai-Wu and Maximum stress criteria. Monte Carlo simulation is performed to quantify uncertainty effects.

This work presents a study on the uncertainty propagation of laminate geometric parameters, associated to each ply thickness and fibre orientation. Each of these input parameters has a specific effect on the simulated linear static response, and therefore on the characterization of its variability. To enable the simulation of uncertainty on the geometrical input parameters, a random multivariate normal distribution is used.

2. MATERIALS AND METHODS

2.1. Fibre reinforced composite materials

Fibre reinforced composite materials are often constituted by two phases: a continuous phase, usually known as matrix and responsible for the embedding of the reinforcement discontinuous one, which in the present case is constituted by long synthetic fibres.

Concerning to its characterization, these materials are considered orthotropic, thus presenting properties that are different in the three mutually perpendicular directions associated to a Cartesian coordinate system.

The typical configuration of these materials is illustrated in Figure 1, where an exploded view of an arbitrary stacking of a three-layered composite is presented.

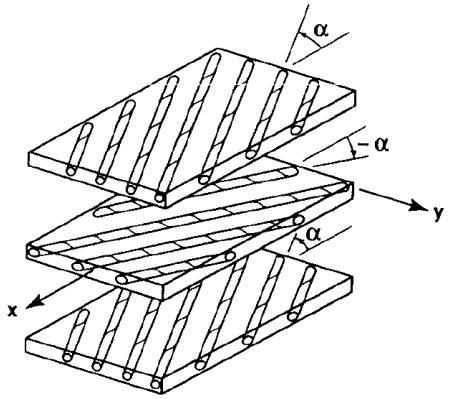


Figure 1. Exploded view of a three-layered fibre reinforced composite material (Jones [7]).

The x and y directions are associated to the laminate structure as a whole, being also visible the angles α between the x direction and the fibre longitudinal direction within each ply. The possibility of considering different materials in different plies allied to the possibility of varying the fibre angle within each ply enables a tailored material, able to an optimized response given known external generalized loads.

2.2. Constitutive relation and equilibrium equation

Due to the characteristics of the plate structures that will be analysed, the plate finite element model used in this work is based on the First Order Shear Deformation (FSDT), as in a previous work of the authors focused on plates made of functionally graded materials (Carvalho et al. [8]). Accordingly, the stress-strain relationships for each ply in the laminate coordinate system, becomes:

$$\begin{bmatrix}\sigma_{xx}\\ \sigma_{yy}\\ \sigma_{xy}\end{bmatrix} = \begin{bmatrix}\bar{Q}_{11} & \bar{Q}_{12} & \bar{Q}_{16}\\ \bar{Q}_{12} & \bar{Q}_{22} & \bar{Q}_{26}\\ \bar{Q}_{16} & \bar{Q}_{26} & \bar{Q}_{66}\end{bmatrix}\begin{bmatrix}\varepsilon_{xx}\\ \varepsilon_{yy}\\ \gamma_{xy}\end{bmatrix} \quad ; \quad \begin{bmatrix}\sigma_{yz}\\ \sigma_{xz}\end{bmatrix} = \begin{bmatrix}\bar{Q}_{44} & \bar{Q}_{45}\\ \bar{Q}_{45} & \bar{Q}_{55}\end{bmatrix}\begin{bmatrix}\gamma_{yz}\\ \gamma_{xz}\end{bmatrix} \tag{1}$$

where the transformed reduced elastic stiffness coefficients are given in literature (Reddy [9]), σ_{ij} stand for the stress components and ε_{ii} , γ_{ij} represent the normal and shear deformations. To overcome the through-thickness constant prediction of the transverse shear stresses, enabling to the model a closer response to the 3D Elasticity predictions, as usually done when considering this theory in an equivalent single layer approach, one considers a shear correction factor, which in the present study was set to 5/6.

To obtain the equilibrium equations that will allow carrying out the required linear static analyses, it is considered the potential energy functional, which can be written as:

$$\Pi = \frac{1}{2}\int_{\Omega} \boldsymbol{\epsilon}^T \boldsymbol{\sigma}\, d\Omega - \int_{S} p_z \boldsymbol{q}\, dS \tag{2}$$

wherein $\boldsymbol{\epsilon}$ and $\boldsymbol{\sigma}$ stand for the generalized strains and stresses vectors, the applied surface load is represented by p_z and $\boldsymbol{q}$ represents the generalized degrees of freedom vector.
After the functional minimization and some mathematical manipulations, the equilibrium equations for the whole discretized domain, become:

$$\boldsymbol{Kq} = \boldsymbol{f} \tag{3}$$

with $\boldsymbol{K}$ and $\boldsymbol{f}$ denoting respectively the plate structure elastic stiffness matrix and the generalized applied forces vector. After the boundary conditions imposed [9], it is then possible to obtain the nodal generalized displacements.

2.3. Simulation of the geometrical parameters uncertainty

Aiming to simulate the variability of a set of real specimens, one considers that the geometrical properties of a laminate composite material are uncertain. In the present work, one focuses on the study of the uncertainty propagation of the ply thicknesses and stacking angles. Each model parameter has a specific effect on the simulated responses, either static or dynamic, and therefore on the characterization of their variability. Thus, to simulate the uncertainty on the referred geometrical properties, one has used a random multivariate normal distribution, from which a set of model parameters $\boldsymbol{X}$ were sampled. Hence, the model parameters were sampled considering $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.e., $\boldsymbol{X}$ is distributed as a normal variable with the mean values $\boldsymbol{\mu}$ (Table 1) and the covariance matrix $\boldsymbol{\Sigma}$. Additionally, the correlation matrix, equal to the identity, is given to ensure that the independence between modeling parameters. Note that one uses a Latin hypercube sampling (LHS) with the ability to ensure the independence between variables [10] to sample from a multivariate normal distribution with 30 observations.

2.4. Forward propagation of the uncertainty

The referred sampling procedure was applied to build different samples, aiming at simulating several plates made of different combinations of properties that are used with different aspect ratios a/h (relation between plate edge and its thickness). The material properties of the used composite material are given in Table 1. Note that only the geometrical properties are considered uncertain for plates with $a/h = [20;\ 100]$. Tables 2 and 3 summarize the case studies analyzed. Note that the values considered or the case studies are based on the manufacturing uncertainty. With the samples for all the defined cases, one performed a finite element analyses to evaluate the maximum transverse deflection, followed by an assessment of the correlation coefficients obtained for all cases.

$\boldsymbol{E_{11}}$ [GPa]	$\boldsymbol{E_{22}, E_{33}}$ [GPa]	$\boldsymbol{G_{12}, G_{13}}$ [GPa]	$\boldsymbol{G_{23}}$ [GPa]	$\boldsymbol{\nu_{12}, \nu_{13}}$	$\boldsymbol{\nu_{23}}$	$\boldsymbol{\rho}$ [kg/m^3]
161	11.38	5.17	3.98	0.32	0.44	1500

Table 4. Carbon fiber prepreg lamite properties (IM7/8552UD Hexcel Composites).

Case	a/h	Stacking sequence	$\mu_{\theta_{ply}}$	$\sigma_{\theta_{ply}}$
1.a	20	$[0]_4$	nominal values	2º
1.b		$[0\ 90]_s$		
1.c		$[0\ 90]_2$		
2.a	100	$[0]_4$	nominal values	2º
2.b		$[0\ 90]_s$		
2.c		$[0\ 90]_2$		

Table 5. Case studies with uncertain stacking angles.

Case	a/h	Stacking sequence	$\mu_{h_{ply}}$	$CoV_{h_{ply}}$
3.a	20	$[0]_4$	0.131 mm	2.5%
3.b		$[0\ 90]_s$		
3.c		$[0\ 90]_2$		
4.a	100	$[0]_4$	0.131 mm	2.5%
4.b		$[0\ 90]_s$		
4.c		$[0\ 90]_2$		

Table 6. Case studies with uncertain ply thicknesses.

It is worthy to mention that one uses the same sample of parameters for all the case studies related to the stacking angles and other sample for the cases related to the uncertain ply thicknesses. This is done to enhance the comparison between cases.

3. RESULTS AND DISCUSSION

The results discussed in the present section are focused on the assessment of the impact of the parameter uncertainty on the maximum transverse displacement of a carbon reinforced composite plate. Based on the methodology presented in Section 2.3, the geometrical

properties were simulated using a sample of 30 observations, which is a sufficiently large sample size to support the significance of the results, keeping the problem at a reasonable size to deal with experimental test data. With the sampled modeling parameters, one has carried out a finite element analyses to build a sample of the maximum transverse displacement for each one of cases presented in Section 2.4. The finite element analysis was carried out using the plate finite element formulation described in Section 2.2, considering simply supported edges and a uniform pressure. Note that the reference to a ply number is related to the stacking sequence illustrated in Figure 2, here the first ply is the lower one and the fourth the one on the upper surface, where the load is applied.

This discussion is focused on the analysis of the correlation coefficients obtained for different plates and uncertain parameters.

Figure 2 shows the sampled values for a set of laminate plates modeled according to Case 1.a. As expected, the individual histograms show a Gaussian behavior for the stacking angles and they are uncorrelated, as shown by the scatterplots and the corresponding correlation coefficients. Note that the correlation coefficients related to the modeling parameters are close to zero (Figure 2) and therefore their independency is verified, which is consistent with the uncertainty simulation described in Section 2.3. Figure 3 shows the same plot, however for Case 3.a, for which the uncertain parameters are the ply thicknesses, instead.

Aiming at assessing the influence of each ply, one has computed several combinations, considering different sets of uncertain parameters, namely all the stacking angles are uncertain

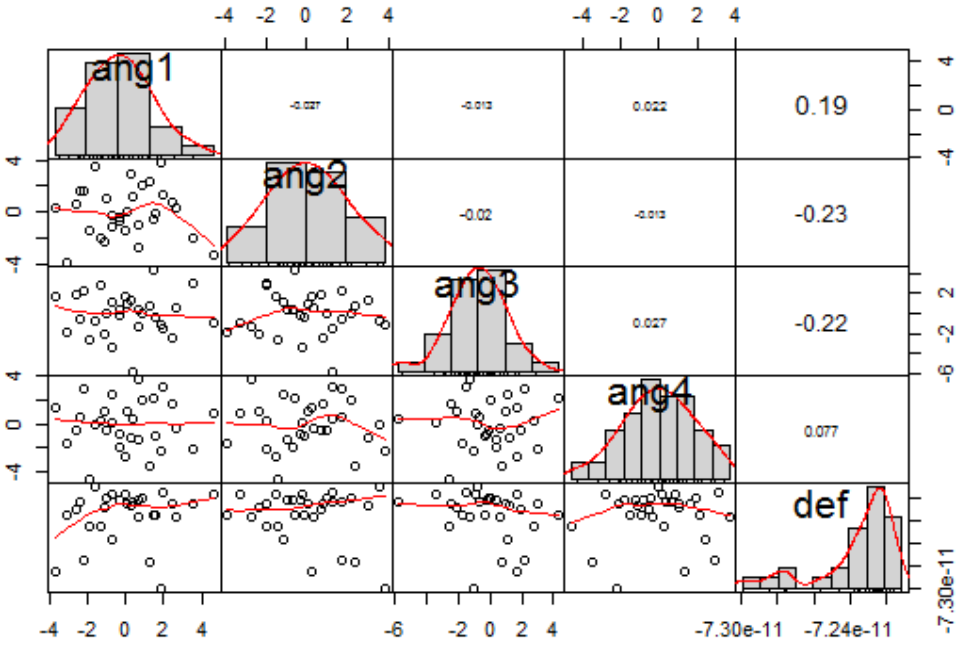


Figure 2. Matrix plot for the maximum transverse displacement (def) and for all the stacking angles (ang1-4) for Case 1.a.

(ang_all) and then that only one ply at a time has uncertain properties (ang1-4), as shown in Figure 4. Note that the sample of maximum transverse displacements of Figure 2 is the one for the combination with all the stacking angles uncertain (ang_all) in Figure 4.

Matrix plots as the presented ones were analyzed for all case studies (Tables 2 and 3), however, for sake of simplicity, one summarizes the results in Tables 4-9.

The correlation coefficients between samples for the maximum transverse displacement for almost all case studies are dominated by the uncertain properties of the fourth ply (Tables 4-9), with higher values for the cases related to uncertain ply thicknesses (Tables 7-9).

Comparing the correlation coefficients for Cases 1.a and 2.a (Table 4), one can observe that the values for θ_2 are higher than the ones for θ_1 and θ_3. It is interesting to observe an inversion of the correlation sign between Cases 1.a and 2.a (a/h=[20; 100]). This happens only for θ_2 and θ_3 for stacking sequence $[0]_4$ and it must be further evaluated (also compare Figures 4 and 5). The correlation values between θ_3 and θ_4 changes with the stacking sequences, from around zero for $[0]_4$ (Table 4) to almost 0.30 for $[0\ 90]_S$ (Table 5) and to different sign for $[0\ 90]_2$ (Table 6). Beside this, it is evident that the cases of Tables 5 and 6 are quite similar, despite the difference between stacking sequences. Note that the correlation coefficient for θ_1 is higher in these cases, reaching values similar to the one of θ_4 (Table 6). On the other hand, Table 5 shows that the correlation for $[0\ 90]_S$ presents higher values for all stacking angles, despite the value for θ_4 remains the highest.

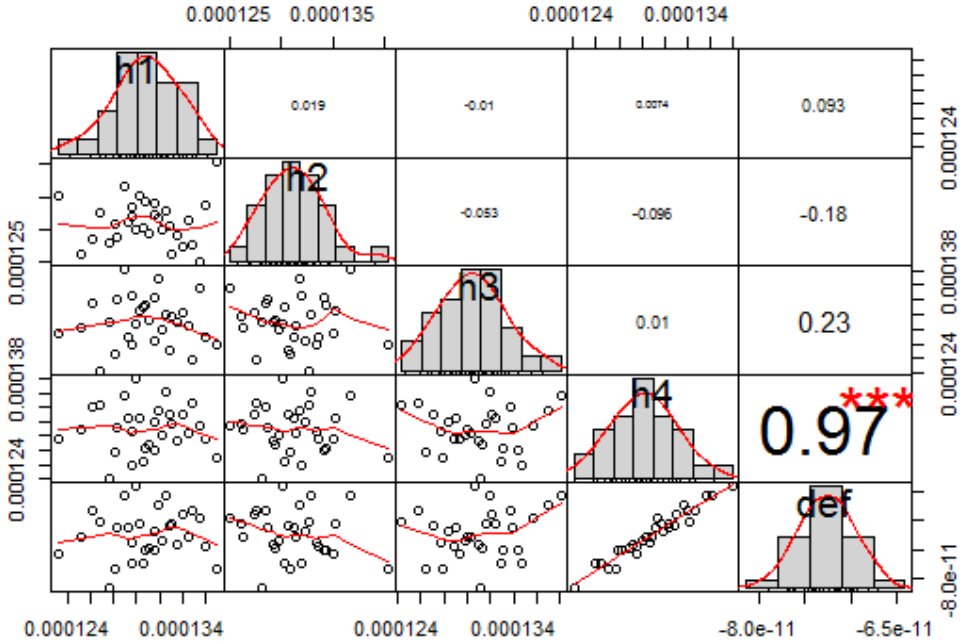


Figure 3. Matrix plot for the maximum transverse displacement (def) and for all the ply thicknesses (h1-4) for Case 3.a.

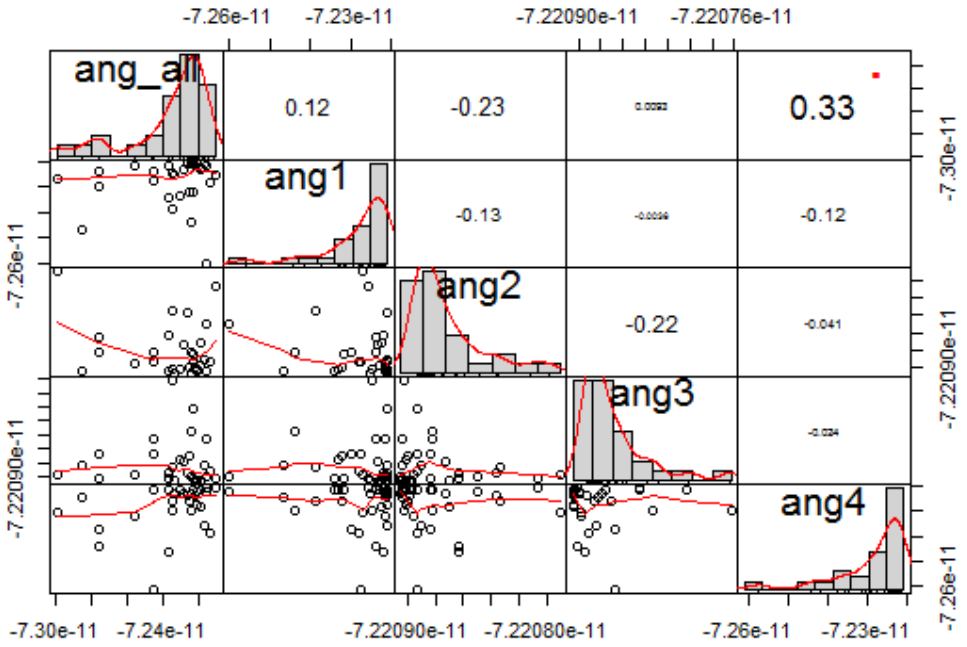


Figure 4. Matrix plot for the maximum transverse displacement considering different sets of uncertain stacking angles for Case 1.a.

In terms of the cases with uncertain ply thicknesses, the correlation coefficients for the thickness of the fourth ply (h_4) overcome all the others, with values near 1.0 (Tables 7-9).

Whereas, h_1 presents the lower values. Note that the correlation between combinations where only one of the ply thicknesses is considered to be uncertain at a time is not significant, despite the correlation between h_2 and h_4.

Comparing the cases with uncertain stacking angles (Cases 1 and 2) and the ones with uncertain ply thicknesses (Cases 3 and 4), for the different aspect ratios, there is more consistency with the distributions of the maximum transverse displacement for Cases 3 and 4 (Figures 6 and 7)., which are almost symmetric. Whereas for Cases 1 and 2 there are significant changes with the aspect ratios and stacking sequences (Figures 8 and 9).

θ_{all}	0.12	-0.23	0.01	**0.33**
	θ_1	-0.13	-0.01	-0.12
		θ_2	-0.22	-0.04
$[0]_4$			θ_3	-0.02
$a/h = 20$				θ_4

θ_{all}	0.16	0.18	-0.07	**0.35**
	θ_1	0.26	-0.09	-0.12
		θ_2	-0.18	0.04
$[0]_4$			θ_3	-0.05
$a/h = 100$				θ_4

Table 7. Correlation coefficients obtained with uncertain stacking angles for Case 1.a (left) and Case 2.a (right).

θ_{all}	**0.31**	0.17	**0.33**	**0.46**
	θ_1	0.17	0.29	-0.12
		θ_2	-0.15	-0.28
$[0\ 90]_s$			θ_3	0.27
$a/h = 20$				θ_4

θ_{all}	**0.32**	0.14	**0.33**	**0.49**
	θ_1	0.16	0.29	-0.12
		θ_2	-0.15	-0.27
$[0\ 90]_s$			θ_3	0.27
$a/h = 100$				θ_4

Table 8. Correlation coefficients obtained with uncertain stacking angles for Case 1.b (left) and Case 2.b (right).

θ_{all}	**0.35**	0.00	-0.10	**0.33**
	θ_1	0.00	0.26	-0.13
		θ_2	-0.19	0.19
$[0\ 90]_2$			θ_3	-0.19
$a/h = 20$				θ_4

θ_{all}	**0.36**	-0.01	-0.09	**0.33**
	θ_1	0.00	0.26	-0.13
		θ_2	-0.20	0.19
$[0\ 90]_2$			θ_3	-0.20
$a/h = 100$				θ_4

Table 9. Correlation coefficients obtained with uncertain stacking angles for Case 1.c (left) and Case 2.c (right).

h_{all}	0.10	0.17	0.24	**0.97**
	h_1	-0.01	-0.01	0.02
		h_2	0.04	0.09
$[0]_4$			h_3	0.02
$a/h = 20$				h_4

h_{all}	0.10	0.17	0.24	**0.97**
	h_1	-0.01	-0.01	0.02
		h_2	0.04	0.10
$[0]_4$			h_3	0.02
$a/h = 100$				h_4

Table 10. Correlation coefficients obtained with uncertain ply thicknesses for Case 3.a (left) and Case 4.a (right).

h_{all}	0.06	0.25	**0.45**	**0.88**
	h_1	-0.02	0.00	0.02
		h_2	0.05	0.10
$[0\ 90]_S$			h_3	0.01
$a/h = 20$				h_4

h_{all}	0.06	0.26	**0.45**	**0.88**
	h_1	-0.02	0.00	0.02
		h_2	0.05	0.10
$[0\ 90]_S$			h_3	0.01
$a/h = 100$				h_4

Table 11. Correlation coefficients obtained with uncertain ply thicknesses for Case 3.b (left) and Case 4.b (right).

h_{all}	0.10	0.18	0.24	**0.97**
	h_1	-0.01	0.00	0.02
		h_2	0.05	0.10
$[0\ 90]_2$			h_3	0.02
$a/h = 20$				h_4

h_{all}	0.09	0.18	0.22	**0.97**
	h_1	-0.01	0.00	0.02
		h_2	0.05	0.10
$[0\ 90]_2$			h_3	0.02
$a/h = 100$				h_4

Table 12. Correlation coefficients obtained with uncertain ply thicknesses for Case 3.c (left) and Case 4.c (right).

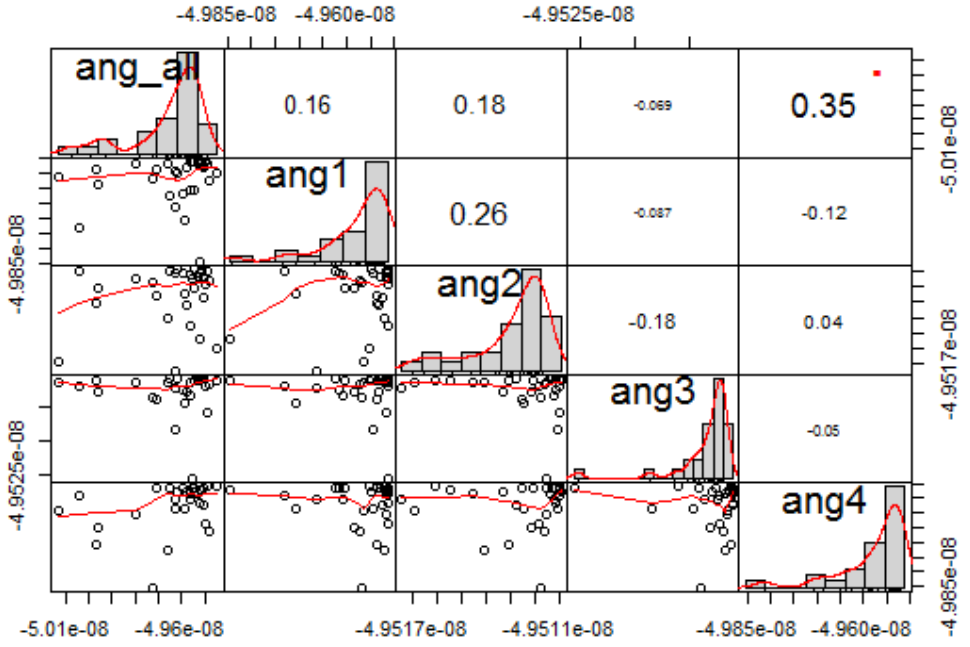


Figure 5. Matrix plot for the maximum transverse displacement considering different sets of uncertain stacking angles for Case 2.a.

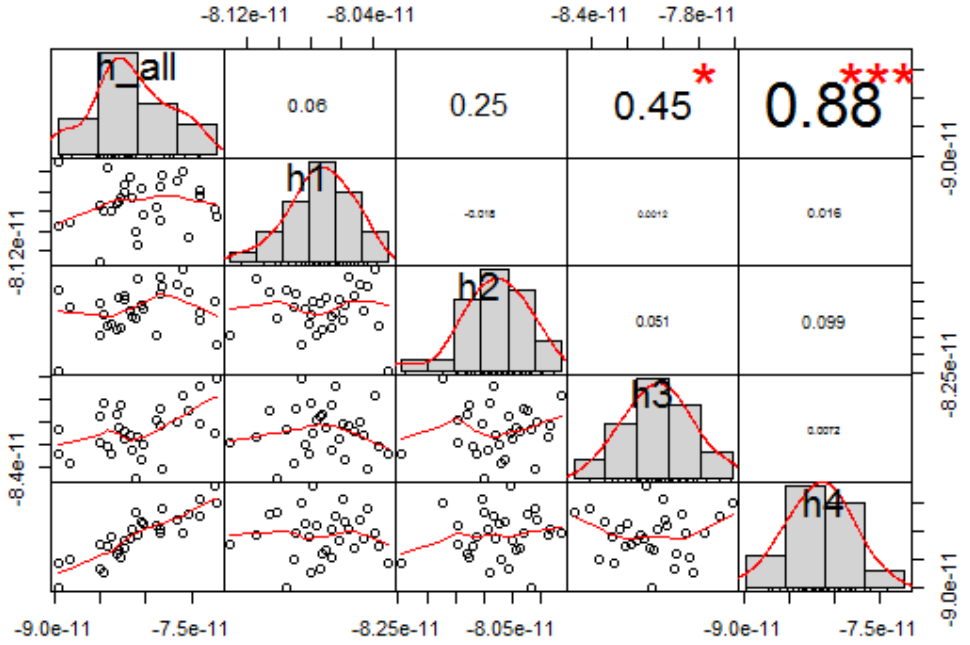


Figure 6. Matrix plot for the maximum transverse displacement considering different sets of uncertain ply thicknesses for Case 3.c.

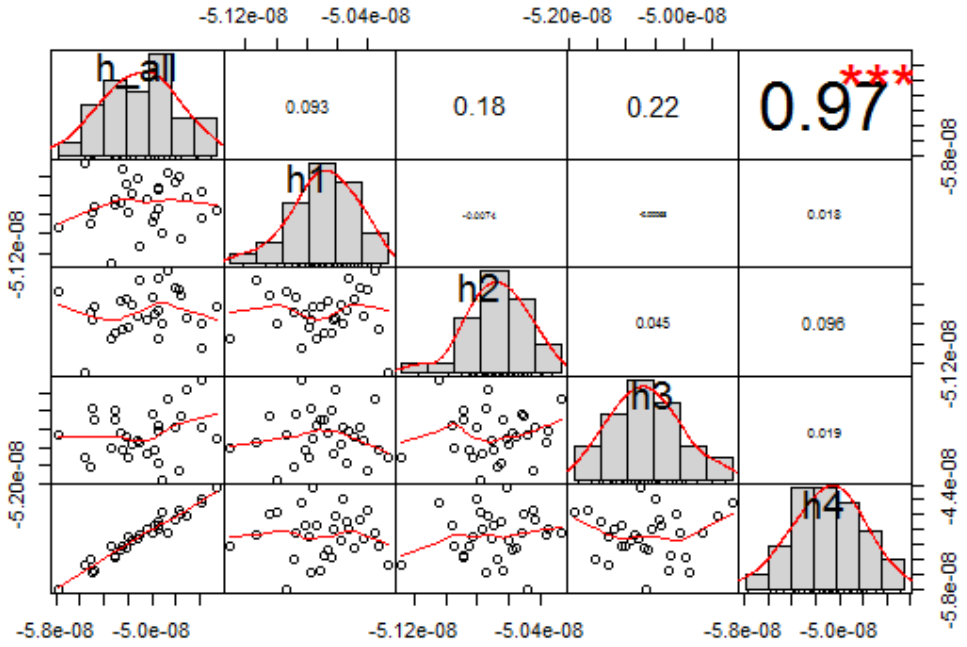


Figure 7. Matrix plot for the maximum transverse displacement considering different sets of uncertain ply thicknesses for Case 4.b.

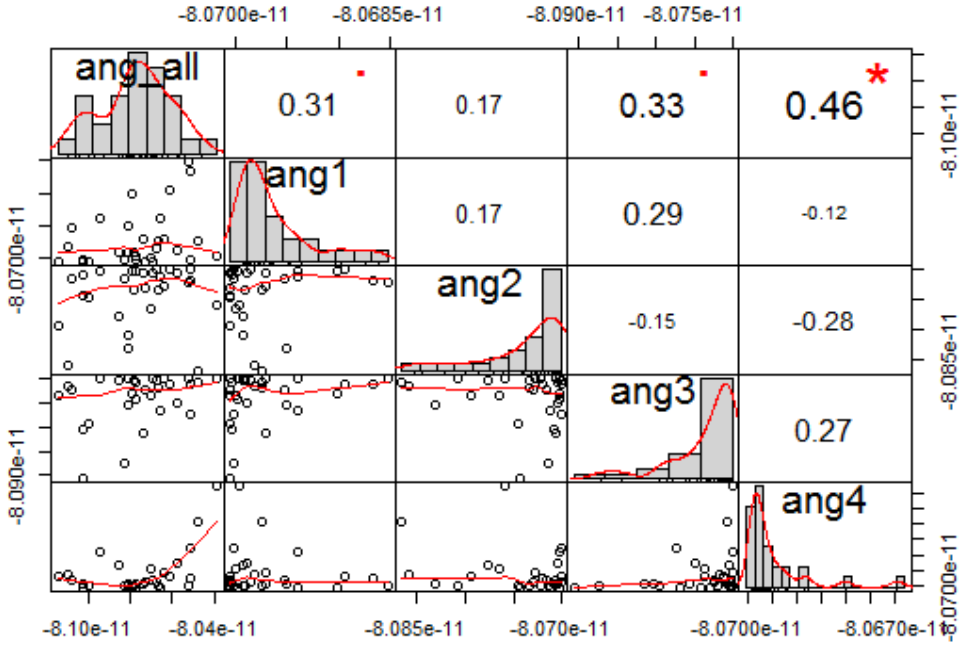


Figure 8. Matrix plot for the maximum transverse displacement considering different sets of uncertain stacking angles for Case 1.c.

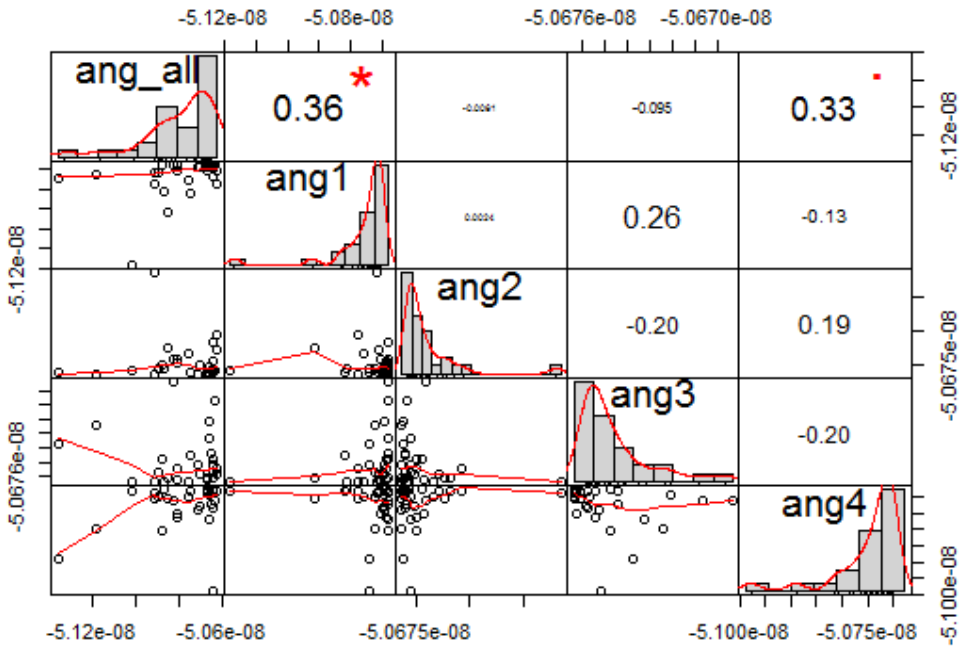


Figure 9. Matrix plot for the maximum transverse displacement considering different sets of uncertain stacking angles for Case 2.b.

4. CONCLUSIONS

This work presented a study on the uncertainty propagation of laminate geometric parameters, associated to each ply thickness and fibre orientation. The simulation of the uncertainty on these input parameters is carried out by considering a random multivariate normal distribution.

The specific effect that these input parameters have on the simulated linear static response of a certain composite structure, and therefore on the characterization of its variability, was analysed and preliminary conclusions were drawn.

From the results obtained, it is possible to observe that the variability of the maximum transverse deflection is more sensitive to changes on the external layers geometrical parameters, namely to the upper surface one. However the authors consider that further studies are necessary to complement the preliminary results here presented.

Nevertheless, under the present assumptions, it can be concluded that the approach here used, can be applied to any fibre reinforced composite structure, in order to characterize the contribution of each modeling parameter on the variability of the predicted responses

ACKNOWLEDGEMENTS
The authors wish to acknowledge the financial support of Project IPL, IDI&CA/COMPDRILL and of the Fundação para a Ciência e a Tecnologia through Project LAETA-UID/EMS/50022/2013, Project UNIDEMI-Pest-OE/EME/UI0667/2014 and Project CEMAPRE-*UID/Multi/00491/2013.*

REFERENCES

[1] ONLINE: State of the Composites Industry Report for 2017: http://compositesmanufacturingmagazine.com/2017/01/composites-industry-report-2017/ 01 April 2017.
[2] T.S. Mesogitisa, A.A. Skordosa, A.C. Long. Uncertainty in the manufacturing of fibrous thermosetting composites: A review. Composites Part A: Applied Science and Manufacturing, vol. 57, pp. 67-75, 2014.
[3] A.K. Noor, J.H. Starnes Jr., J.M. Peters. Uncertainty analysis of composite structures. Computer Methods in Applied Mechanics and Engineering, vol.185(2–4), pp. 413-432, 2000.
[4] C.C. António, L.N. Hoffbauer. Uncertainty assessment approach for composite structures based on global sensitivity indices. Composite Structures, vol. 99, pp. 202–212, 2013.
[5] H. Teimouri, A.S. Milani, R. Seethaler, A. Heidarzadeh. On the Impact of Manufacturing Uncertainty in Structural Health Monitoring of Composite Structures: A Signal to Noise Weighted Neural Network Process. Open Journal of Composite Materials, vol. 6, pp. 28-39, 2016.
[6] S. Mukherjee, R. Ganguli, S. Gopalakrishnan, L.D. Cot, C. Bes. Ply Level Uncertainty Effects on Failure of Composite Structures. Le Cam, Vincent and Mevel, Laurent and Schoefs, Franck. EWSHM - 7th European Workshop on Structural Health Monitoring, Jul 2014, Nantes, France. 2014.
[7] R.M. Jones. Mechanics of Composite Materials, 2nd edition. Taylor and Francis, Philadelphia, USA, 1999.
[8] A. Carvalho, T. Silva, M.A.R. Loja, F.R. Damásio. Assessing the influence of material and geometrical uncertainty on the mechanical behavior of functionally graded material plates, Mechanics of Advanced Materials and Structures, vol. 24(5), pp. 417-426, 2017.
[9] J.N. Reddy. Mechanics of Laminated Composite Plates. CRC Press, Boca Raton, Florida, USA, 1997.
[10] R.L. Iman, W.J. Conover. A Distribution-Free Approach To Inducing Rank Correlation Among Input Variables. Commun. Stat. B, vol. 11, pp. 311-334, 1982.

SYMCOMP 2017
Guimarães, 6-7 April 2017

SOUNDING ROCKETS MODELLING AND SIMULATION WITH MATHEMATICA

Paulo J. S. Gil[1*] and Tiago M. O. Pinto[2]

1: CCTAE, IDMEC, Departamento de Engenharia Mecânica
Instituto Superior Técnico
Universidade de Lisboa
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mail: paulo.gil@tecnico.ulisboa.pt

2: MSc student, Aerospace Engineering
Instituto Superior Técnico
Universidade de Lisboa
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mail: tiagomopinto@tecnico.ulisboa.pt

Keywords: Sounding Rockets, Mathematica, Simulation, Preliminary Design, Wind Profile

Abstract. *In this work, we develop a tool for preliminary design and trajectory simulation of passive-controlled model and sounding rockets, with the possibility of taken into account various effects, such as the wind. The objective was to be able to rapidly assess improvements on the rocket design or the importance of physical effects affecting the trajectory, such as the wind, and the accuracy of the models used to take them into account. The tool was developed in Mathematica© in order to explore the advantages of fast modelling of the relevant effects using analytical or semi-analytical approaches with various levels of simplicity or accuracy. In the context of the tool, we developed a data model for a simplified parametric description of sounding rockets with an arbitrary number of stages and many motors. The tool can be used as a platform to assess the relative importance of the perturbations affecting the trajectory and the quality of the models used to take them into account, simulation of trajectories given initial conditions, and rocket model development. As case study we focus on the effects of the wind and the modelling of the atmospheric boundary layer.*

1 INTRODUCTION

Passive-controlled model and sounding rockets have a considerable number of enthusiasts, usually aggregated under associations such as the American National Association of Rocketry (NAR) [1], that define rules, organize meetings and competitions, provide safety codes, and are recognized to certify their members. To ensure safety and success, it is helpful to determine the rocket's probable trajectory and its influencing factors.

Although tests are important, the main way to study the problem is by simulating the flight conditions and perform all kinds of analysis in order to foreseen what will happen during the flight as best as possible. However, being a very complicated problem to be solved with limited resources, especially in the context of small rockets, models are usually simplified. In these conditions it is particularly important to not only evaluate the models but also to improve and test them, mainly by means of simulation.

In this paper we develop a trajectory simulator in Mathematica© [2] for passive-controlled rockets. The emphasis is on being able to include the relevant effects influencing the flight, to easily experiment and develop new models describing relevant physical effects, such as the wind, and to experiment with the simplified dynamical description often used for model and sounding rockets, and missiles. A data model describing a simplified staged rocket was also developed. The data model is able to describe a fairly large set of rocket types. The objective is to take advantage of the large capabilities of Mathematica© to analyze the flight, and to be able to promptly test and develop new models describing or affecting the problem such as the effects of the wind.

Being a high level programming environment, a Computer Algebra System (CAS) such as Mathematica© offers many capabilities for analysis, such as a large set of built-in functions, functional programming, analytical calculations, and many more. When developing new approaches to a problem, the most time consuming tasks are usually not the processing time but how fast experiments, models, and changes can be performed. This should compensate the performance limitations of using an interpreted language.

In the remainder of this paper we start to discuss the design of rockets and how to simulate their trajectories. Then, we develop the data model describing the rockets and how the simulation environment is established. Next, we show how simulations are setup and how flexible they can be by using some features of the programming environment, such as functional programming. Finally, we show the capabilities of the simulator by presenting example applications, namely, a multi-stage rocket in its trajectory, the effects of the wind and other characteristics of the atmosphere, probable landing location with uncertain wind, and optimization of rocket design.

2 ROCKET DESIGN

2.1 Small rockets and their simulation

After the World War II people became interested in amateur rocketry, especially inventors and intellectuals, which led to several imprudent and unsupervised launches [3].

This lead to the development of solid-fuel motors to be safely used in small rockets that could be recovered and reused [3].

Model rockets are typically made of safe materials such as cardboard, plastic and balsa wood. They are propelled by a replaceable, small and pre-packaged solid fuel motor that has an ejection charge to deploy the recovery system. This consists of a parachute (or a streamer) attached to the nose cone.

High power rockets differ from model rockets in the propulsion power and weight. They have to fly under the safety code of the local governing organizations and only a qualified user can purchase a motor provided with such power [4]. Starting from these categories, the space is the limit with high-end amateur rockets and sounding rockets reaching altitudes of dozens of kilometers.

Currently, the determination of the probable trajectory under particular conditions is done with simulators. Two of the commonly used rocket flight simulators, that also include a rocket design tool, are *OpenRocket* [5] and *RockSim* [6]. The latter is a proprietary software which makes it impossible to validate its methods. Also, it may represent a significant investment for students and rocket hobbyists. These simulators make a series of assumptions commonly used to solve the problem namely use of the International Standard Atmosphere, models for drag, simplified attitude dynamics models, and others. These models are built-in in the software and cannot be easily and extensively improved [5]. For example, in the case of *OpenRocket*, the wind is determined by summing a constant speed along the altitude with a random, zero-mean turbulence velocity. We cannot change this model if we want to improve the result or meet specific conditions. Also, although these tools allow the design of rockets, it is very difficult or even impossible to analyse improvements on the design or test different models of possibly relevant physical effects.

2.2 Model rocketry

Model and high power rockets are an assembly of several pieces fitting together to produce the desired shape of the rocket. Basic model rockets consist of a nose and fins attached to a cylindrical body tube. The most common used noses are conical, parabolic or ogive shaped, and at their bottom exists the nose's shoulder to fit, internally, the nose to the body. To connect the stages of the rocket, or to change the body tube diameter, as an option, conical transition sections can be used. If these parts increase the rocket diameter, they are called *shoulders*, otherwise are *reducers* or *boat-tails*. The latter are used at the back of the rockets to decrease the base drag. The more motors a rocket has, the more pieces it will need to hold them in a fixed position and align the thrust force with the longitudinal axis of the rocket. To ensure this, engine mounts consisting of a mount tube, centering rings, and a motor block, must be used.

A model rocket's flight is usually divided in five phases [7]: (i) ignition and lift-off; (ii) engine burnout; (iii) coasting phase; (iv) apogee and ejection; and (v) recovery. Each phase must developed in harmony so the rocket achieves its goals, returning safely to the ground, and the simulator must be able to cope with all of them.

The launch pad includes a rod to guide the rocket while it gains the required speed to become aerodynamically stable. The ignition is usually triggered electrically increasing the safety comparing to a fuse in case of any unexpected problem. After the burnout and throughout the coasting phase, the rocket gains altitude and loses speed until the recovery system is deployed, safely, as the drag forces are significantly reduced. The rocket's descent can be significatively influenced by the wind. Usually, a parachute is used as a recovery system but a streamer can be used in order to keep the rocket in sight, since the fall becomes faster.

The propellant's burning rate, which is proportional to the chamber pressure, is not constant and changes according to the burning area [8]. Therefore, using different grain configurations, it is possible to build several thrust profiles to get the desired propulsive properties. Model rocket motors are divided in classes depending on their total impulse.

A model rocket is stable if its Center of Pressure (CP) is behind the Center of Mass (CM) (seeing from the nose to the tail of the rocket) [9]. The CP is the point where acts the resultant aerodynamic force produced by the air pressure and moves toward the nose with increasing angle of attack α. This force, when the angle of attack is different from zero, may be decomposed in axial and normal components where the latter creates a moment about the CM of the rocket. This moment produces a damped oscillating movement about the air flow direction until the rocket angle of attack returns to zero and the normal force vanishes, ensuring stability [9]. The static margin is the arm of the moment i.e., the distance between the CM and the CP. The larger the static margin, the higher the stability. However, if the static margin is too high, the rocket may become overstable and reaches a lower apogee because it turns sooner to the wind [9]. This effect, called weathercocking, is also enhanced if the velocity is low when leaving the launch pad. Hence the importance of taking the wind into account, especially in smaller rockets.

Just like in larger rockets, multi-staging can also be used in model and sounding rockets to minimize transporting structural mass no longer needed after the propellants started to be spent. In model rocketry there are two ways of staging a rocket: direct and indirect staging [10]. With direct staging the upper motor is ignited by the lower stage motor called the *booster*. These type of motors do not carry delay composition and ejection charge in order to ignite the upper stage nearly instantaneously after the burnout. Direct staging is the simplest and cheapest method because it does not require electronic devices to ignite the upper stage, unlike indirect staging [10].

With serial staging the lower stages must be equipped with fins (with increasing area on each added stage) to compensate the shift of the CM back to the tail [11]. These fins must be designed so that the stages become aerodynamically unstable in order to tumble and decrease speed during recovery. A larger fin area, however, tends to over-stabilize the rocket and increases the tendency to weathercock. For this reason, a multi-stage rocket should only fly with calm winds and never use more than three stages [12].

Parallel staging can be achieved if the rocket is provided with external boosters with a lower burnout time than the central motor. In this case, when the core motor and the

boosters are burning at the same time, it is called the *zeroth stage* [13]. Several motors can also be used together to provide the rocket with more thrust, denominated *clustering*, although usually limited to at most four engines due to reliability issues of the ignition [14].

2.3 Rocket dynamics

The motion of rockets can be studied relative to a reference frame fixed with the Earth, which is a natural selection as the atmosphere co-rotates with the planet. Using a plane Earth approximation is frequently enough as reference frame for small rockets. However, for sounding rockets, the effects from the rotation of the Earth (namely the Coriolis force) can be taken into account for extra precision. The equations of motion can be written as

$$m\vec{a} = \vec{F}_i + \vec{T} + \vec{D} + \vec{W}, \tag{1}$$

where m is the mass, $\vec{a}$ the acceleration relative to the Earth, $\vec{F}_i$ the Coriolis and other inertial forces when the rotation of the Earth is considered, $\vec{T}$ the thrust, $\vec{D}$ the drag forces, and $\vec{W}$ the weight. The thrust, that in uncontrolled rockets has a definite profile determined by the motor used, but which effectiveness depend not only on the the exhaust velocity of the gases but also on the external pressure and the motor nozzle area, is always align with the rocket's longitudinal axis as the rocket is considered not to be steerable. Since the rocket may fly with angle of attack until it stabilizes and the thrust becomes collinear with the true air speed vector, the effective propulsive force may be lower. A commonly adopted model for the effective thrust force $\vec{T}$ is

$$\vec{T} = T_p \cos\alpha \, \frac{\vec{v}_t}{v_t}, \tag{2}$$

where T_p is the propulsion from the motors' thrust profile, α the angle of attack, and $\vec{v}_t$ the true air speed vector.

The aerodynamic force is typically assumed to be merely the drag since the rocket only flies with considerable angle of attack during a short interval of time after the launch — or after a perturbation — that can be described as a low amplitude oscillation. The lift that is produced during these moments contributes essentially to stabilize the rocket, its average is usually considered to be zero, and its effect can be taken into account using a simple model (see below). Hence, the aerodynamic force can be written as

$$\vec{D} = \frac{1}{2}\rho v_t S C_D \vec{v}_t, \tag{3}$$

where S is the reference area (usually chosen as the maximum sectional area of the rocket or, after the apogee, the reference area of the recovery device) and C_D is the drag coefficient.

The attitude of small rockets can be treated in a simplified way. During the launch, when the rocket is constrained by the guiding rods, the drag and weight normal to the

launch direction are counterbalanced by the rods reaction and only the tangential forces to the rods contribute to the launch. After leaving the platform, we consider that any disturbance will produce a small angle of attack that force the rocket to oscillate about the true air speed direction but will be gradually dumped [9]. This damped oscillating movement is obtained by consider that the rocket is subjected to two moments [15]: a stabilizing (or restoring) moment and a damping moment. The stabilizing moment comes from the normal aerodynamic force acting on the CP which causes the rocket to rotate about the CM since these points must distance from each other by the static margin.

One source of the damping moment results from the aerodynamic resistance of the air while the rocket is rotating [15]. During the rotation, the angle of attack of each part of the rocket changes due to the tangential velocity of this motion. Like the stabilizing moment, only the normal force of this additional resistance contributes to dampen the rotation. The other contribution to the damping moment comes from the Coriolis acceleration due to the change of the gas flow through the nozzle [15], also called *jet damping*.

Once the moments acting on the rocket are known, we can find the governing equation for the motion of the angle of attack assuming the rotation is two dimensional (in the plane formed by the rocket's velocity and wind vectors) [16] which are equivalent to a damped harmonic oscillator. They can be *underdamped*, when the angle of attack decreases to zero, *critically damped* when the angle of attack returns smoothly to zero, or *overdamped*, when the disturbance decreases but always flies with an angle of attack different from zero. How fast this oscillations are is determined by the natural frequency of the rocket. The oscillations of the rocket become faster if the CM moves toward the nose. On the contrary, increasing the inertia of the rocket makes the oscillations slower, as expected.

The effect of dumping of the angle of attack after a gust of wind (perturbation) can be seen in Section 4.3.

2.4 Rocket parameters

In this work we will consider the general structure of a multi-stage rocket shown in Figure 1. Although a three-stage rocket is represented, this structure can be extended to

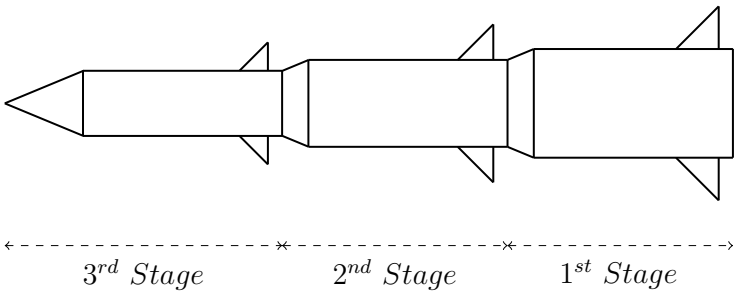


Figure 1: Generic multi-stage rocket.

any number of stages and was developed with the goal of being able to model the most common types of rockets. For each stage, we considered the following components: a connector (or nose, in the case of the last stage), a body tube, and fins. Internally, this structure only considers an autonomous description of the motors, or engines, since data from additional components can be joined with the body tube properties. The last stage includes the nose and in the first stage the rocket can also be supplied with an internal cluster of motors or external boosters in order to be able to describe parallel staging. The reference from which the position and properties of the components are specified is defined as being the top of the respective component.

The rocket's CM, moments of inertia, CP, and drag coefficient may depend and change according to the mass, shape and position of the components. There are cases where the rocket's properties are impossible to determine directly due to lack of information. Therefore, it is important to split the high level description of the rocket from its parts. For example, at high level, we only want to know the location of the CM or the moments of inertia. They vary with time since mass will be exhausted in a way specified by the motors. On the other hand, these parameters also depend on what kind of structure is being used in the rocket. So we envisioned two steps: first, select the motor and body of each stage, which will determine how the CM and moments of inertia of each stage vary with time; second, once all the components of the rocket are defined, we can define the whole rocket and its characteristics.

The properties of each stage and motor can be determined separately and diversely. For example, if the motor is from a supplier, usually some properties are known; tests can be done to determine thrust and mass moments; on the other hand, if the materials are known, calculations based on the geometry can be done to estimate the mass moments. Similarly for the body of the stage. The point is that we must, and can, deal with each stage separately from the assembly of the rocket and its characteristics as a whole.

Separate functions can be defined to obtain these parameters. Each function will have a set of assumptions, depending on the case. New, more precise, functions can be developed if possible and needed. For example, we can define a function that determines de CM for a motor that makes the assumption that the mass of the propellant is expelled homogeneously. But, depending on the shape of the combustion chamber, we can develop models that determine the location of the CM more accurately. These models can be developed independently of how the rocket is simulated since this is done at different levels, as we will see in Section 4. The properties of the whole rocket, e.g. the mass moments, can always be obtained from the properties of the stages (and motors, connectors, cones, and fins) and their relative positions within the complete rocket.

To estimate the aerodynamic characteristics of model and high power rockets we adopted the widely used Barrowman Method [17]. It has a series of assumptions, such as steady, irrotational and subsonic flow, the fins must be flat plates, and others, that restrict the type of rockets to describe but its relative simplicity is an advantage and can be used in many types of rockets. This is especially relevant during and shortly after lift-off

which, not considering wind perturbations, is when the CP and the moments controlling the attitude play an important role. The Barrowman Method divides the rocket in simple geometric parts, aligned with the adopted general description presented in Figure 1.

Since sounding rockets can be supersonic the Barrowman Method cannot be used for them. In this case, and since the CP is used to determine the angle of attack, we consider that rocket oscillations are already damped before reaching the transonic regime. At such high speeds the rocket has to fly without angle of attack to minimize the drag forces applied on its structure. One other limitation of this model is to allow only 3, 4, or 6 fins.

The drag force opposes the true speed direction and exists due to several effects of the flow around the rocket. The drag coefficient contributes to the aerodynamic force and it may change throughout the flight. It can be estimated by adding all the drag coefficients of the external components of the rocket and, in each one of them, the drag may come from different sources: pressure, skin friction, base, and wave drag.

Pressure drag arises due to the distribution of normal forces on the components. On the other hand, the skin friction drag is a tangential force generated by the viscosity of the flow that creates a boundary layer on the surface of the rocket [18]. Therefore, assuming a flight with approximate zero angle of attack, the nose, connectors and fins generate these two kinds of drag and the body tubes of each stage only produce skin friction drag.

Skin friction drag changes according to the flow regime. It is higher in turbulent flow than in laminar flow (although turbulence, in some cases, is desirable in order to delay separation and decrease the pressure drag). As the air hits the rocket's surface, the flow is laminar and rapidly turns to turbulent with increasing Reynolds number. Between this two regimes appears a laminar-turbulent transition when critical Reynolds is reached (Re_{cr}) and the flow becomes fully turbulent above transition Reynolds (Re_{tr}).

From Mach number $M > 0.3$, the flow no longer can be assumed as incompressible since there are changes in density and temperature and shock waves begin to increase the drag. The Prandtl-Glauert rule is commonly used to relate incompressible and compressible coefficients for slender and planar bodies [19].

After the apogee, or when it is desired, the recovery system of the rocket is deployed (if applicable) and the main source of drag comes from the parachute or streamer. The respective drag coefficient depends on the shape of the recovery device.

All the previous mentioned effects are modeled in a more or less precise way, depending on the effort that can be made and the required precision. The more precise the requirements are (and therefore the model to use), more effort is required. A tool that can help define models or assess the results of different models can help limiting the required effort.

3 ROCKET DESCRIPTION

3.1 Rocket assembly

As mentioned in Section 2.4, we envisioned defining the rocket in a generic (generic enough in order to be able to describe many rockets, although evidently not all) and

logical way, with the low level entries defined directly or being the output of particular functions. In Mathematica© all data can be defined as a list whose elements can be numbers or lists, that also can have lists, and so forth. The only restriction is that we must know the meaning of each slot. If done correctly, the number of slots can even be variable, that is, some lists can have a variable number of elements (e.g. the number of stages of a rocket).

In the case of a rocket, the *Rocket Assembly List* concerning each complete rocket was defined in a logical way in order to easily access and understand its elements, and to allow any number of stages. Every parameter inside this structure may be defined directly by the user (from measurements or other available data) and some of them can be determined from specific functions that can be defined using the models and methods mentioned in Section 2.

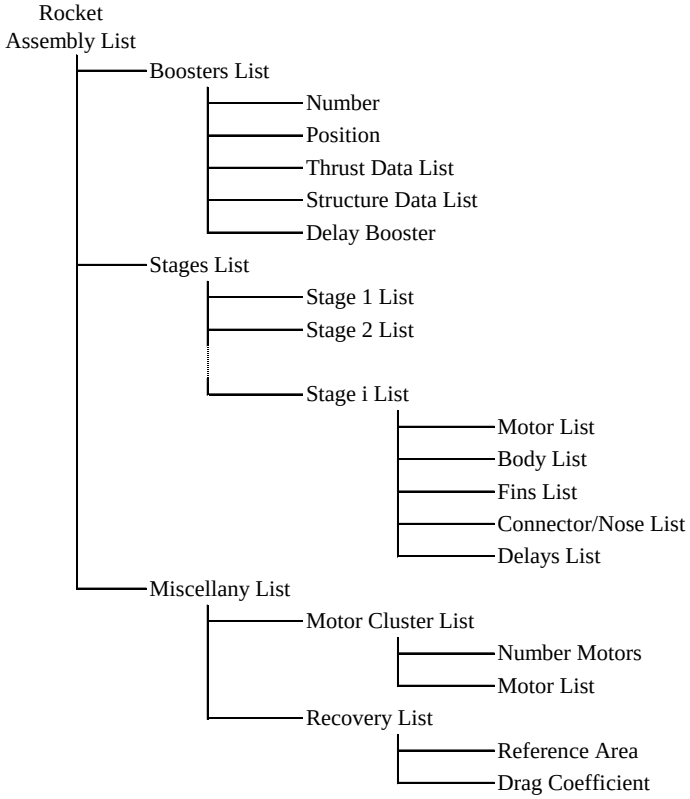


Figure 2: Rocket assembly list structure, the highest level description of a rocket.

Figure 2 shows the scheme in which the assembling of the rocket's data is structured. The sublist of stages contains as many lists as the number of stages. Each one of these lists, having in mind the generic multi-stage rocket presented in Section 2.4, holds other

lists for the rocket's components (*Motor List*, *Body List*, etc.) plus a list (*Delays List*) to indicate the time intervals between events (explained later, below). The lists of components appearing in each stage (Figure 2) will be presented in detail next.

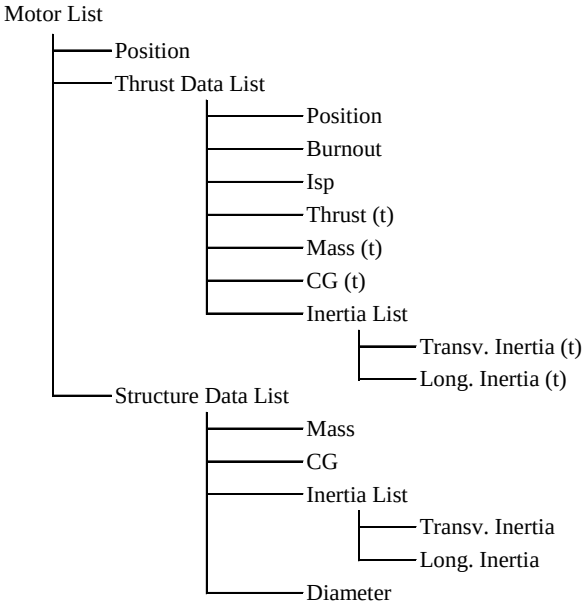


Figure 3: Motor list structure. Some of the entries depend on time.

The *Motor List* is described in Figure 3. Its first entry is the distance between the top of the motor and the top of the body tube of the corresponding stage. Likewise, in the sublist *thrust data list*, the first element is the distance between the top of the combustion chamber and the top of the motor. The moments of inertia should also be given relative to the top of the combustion chamber in order to have a similar fix reference. Otherwise, considering the propellant's moment of inertia, its reference would change over time. Although at this time the simulator does not require the longitudinal moment of inertia, we keep an entry for this property because a future version may require it.

The *Motor List* contains time dependent functions since the propellant is burning over time. Although this dependence is explicitly indicated by the (t) in Figure 3, these function are actually what in Mathematica© is designated by *Pure Functions*, a structure with one or more slots where the variables are located, independently of the symbol that represents each of them. This adds flexibility because other functions can appear that can also have dependencies on the same variables and we can unify the dependency under one symbol, selected once. This is one of the advantages of the functional programming of Mathematica© that we take advantage in this work.

The *Body List* is defined as depicted in Figure 4. There is no explicit time dependence

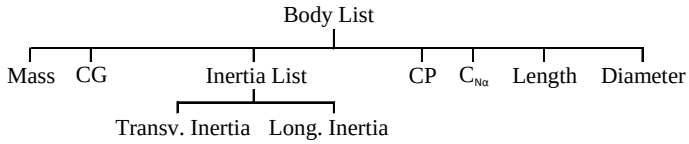


Figure 4: Body list structure. There is no time dependence other than the one when the stage is discarded.

as the variation of mass is totally accounted from the motor, that is one of the reasons the motor is accounted separately from the body (the other the ability of interchange different bodies and motors). From the whole rocket point of view, the only dependence on time induced by a stage rocket body is when latter is discarded. At that instant there will be discontinuities on the CM location, mass, etc., determined by the instant of stage separation.

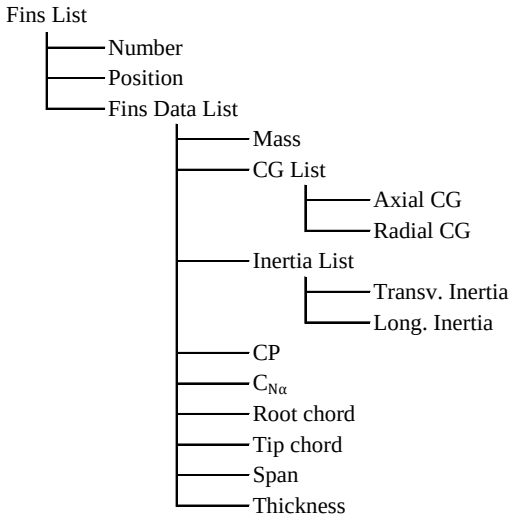


Figure 5: Fins list structure.

In the *Fins List*, showed in Figure 5, there is an additional CM in order to find the distance of the fin's CM from the rocket's CM, which is required for the parallel axis theorem. Also, the CP and the normal force coefficient derivative, C_{N_α}, in this list are assumed to be the values resulting from the total ensemble of fins existing in each stage.

The *Connector List* (or *Nose List*) is structured as depicted in Figure 6. The diameter is not included as it is inherited from the connected stages.

In the *Delays List* (Figure 7) the delay of the body represents the time when the respective stage's body tube is discarded after the burnout; the delay of the connector

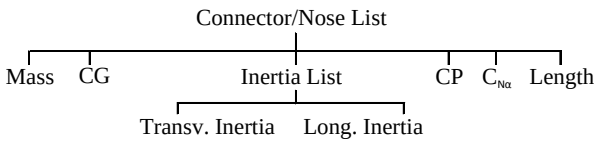


Figure 6: Connector or nose list structure.

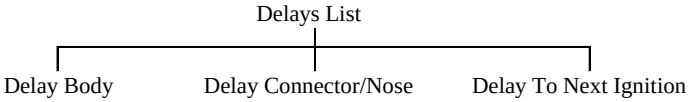


Figure 7: Delays list structure.

gives the instant when this component is discarded after the previous delay (if there is no connector, this delay must be zero); the third delay gives the time interval until the next stage is ignited. When the delays list belongs to the nose, the sum of these three delays represent the coasting time until the recovery system is deployed if the ejection charge option is selected in the simulation of the trajectory.

Since the boosters are external components, the *Structure Data List*, within the *Boosters List* (see Figure 2), contains additional elements: the lengths of the cylindrical tube and nose structures, the CP and the C_{N_α}. The last two must give the respective property for the complete structure of the booster. The boosters must also be identical and, when included, the user should pay attention for the new rocket's CM not surpass the CP. The delay of the boosters represents the time between their burnout and the instant they are discarded, which, at most, should be the instant when the first stage's body tube is also discarded. If the rocket does not have boosters, their number must be zero and when this data is processed the other entries within the *Boosters List* will be ignored.

Inside the *Miscellany List* (see Figure 2), the *Cluster List* allows defining a cluster of identical motors around the first stage's motor and with equal space between them. The *Recovery List* contains the properties of the recovery device (reference area and C_D) to be used during the descending phase of the simulation.

All sublists can be previously and autonomously defined for use when required. This adds modularity, ease of storage and use, and each one can be the output of, or used as, a function, i.e. each list doesn't have to be set element by element but as the result of a function defining the list under some desired rules, and be set to allow variation of one parameter but at the same time respecting the constraints of the problem.

3.2 Motors data

The case of the motors for model rockets illustrates the comment at the end os Section 3.1. We did store some model rocket motors data in another database, constructed

with the principles of data definition in Mathematica©. This is particularly useful for model rocket motors because they are built and sold commercially by companies and data is available and does not change for each model motor.

The data structure is a function whose entries (strings), organized in a certain way, are the named characteristics of the motor; one of the entries identifies the model of the motor. The structure is flexible as it can provide one or more of the stored characteristics. As example of use of the *MotorData* function, with the arguments

```
MotorData[``C6-0'', {``Thrust'',``Isp'',``Mass'',``CG'',``Inertia''}],
```

produces as output, for the model motor *C6-0*, a list of the thrust, specific impulse, mass, center of gravity, and moments of inertia, pure functions of time (from ignition to burnout of the motor). Other motors can be stored in order to automatically change all the motor's properties within the structure of the rocket only by specifying the name of the motor. These new motors may be defined by any mathematical model the user may require, which provides freedom to change a given property and observe its influence in the simulation.

For all the model rocket motors already in the database, their properties are determined considering an end burning combustion, a delay charge (for the ones which possess a delay time) and an ejection charge. The thrust profiles (the *thrust* output of the function exemplified above) are obtained by an interpolated function fitting the points given in the data sheets from the manufacturers that can be found e.g. at [20].

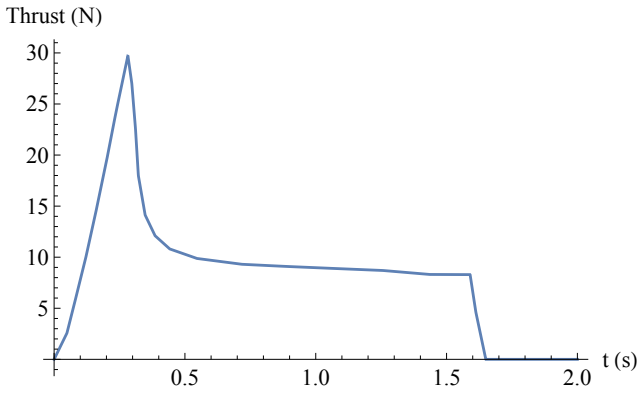


Figure 8: Thrust profile for the D12 motor.

As an example, Figure 8 represents the obtained thrust curve as function of time of the *D12* motor [20].

Although the main motors' properties are found from the thrust profile and other info provided in the data sheets by the methods described in Section 2, they depend on other data which can only be determined by measurements taken from the motors' dimensions.

Therefore, this data must be inserted in the database in order to automatically compute the properties of the motors.

3.3 Rocket characteristics

Once all the data is defined in the *Rocket Assembly List* we can generate the characteristics of the complete rocket using the function

```
RocketProperties[RocketAssemblyList]
```

that builds all the properties of the complete rocket as a function of time, including discontinuities when used stages are discarded, by combining the data of the several stages regarding the developed structure that gathers the data of the rocket and considering the instants of the ignitions and ejections of each stage.

For example, the parallel axis theorem and composition of bodies, the total CM, and moments of inertia are generated as functions of time from the corresponding functions for the stages (also functions of time) taking into account for each stage its location and when is ignited. The generated list of general rocket properties can be seen in Figure 9.

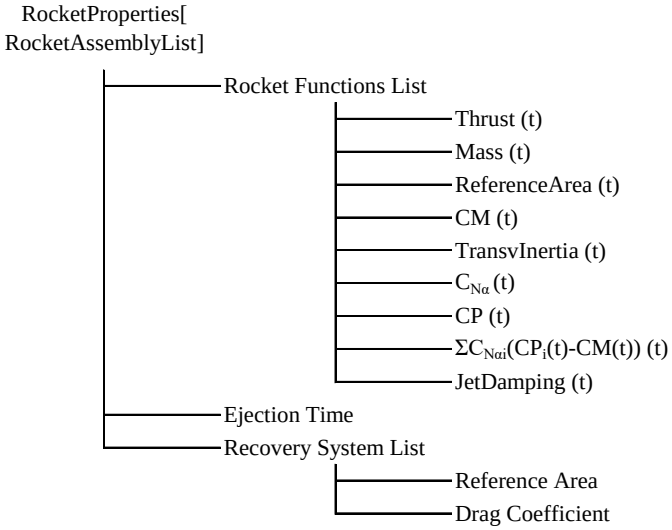


Figure 9: Complete rocket data structure resulting from the *RocketProperties* function.

4 TRAJECTORY SIMULATOR

4.1 Simulator description

Once the rocket is defined we can develop the trajectory simulator. The goals are to (i) determine the rocket trajectory and respective flight data given known/estimated inputs;

(ii) provide flexibility to the user to change the default trajectory and rocket's models; (iii) enable to simulate from micro rockets up to sounding rockets' models; (iv) forecast the landing region from a Monte Carlo simulation; (v) optimize the rocket characteristics. This tool was developed in Mathematica© [2] to take advantage of the functional programming and logical handle of variables, that was the rationale for defining the data structures defining the rockets in Section 3.

The simulator's main function is *RocketTrajectory* which integrates the trajectory equations and receives all the required inputs, including the result of *RocketProperties* that includes every parameter of the rocket as a Mathematica© pure function over time, as well as other parameters, such as the launch location, information about the wind, or even functions, e.g. a function modeling the drag. The simulation will be something like

```
RocketTrajectory[RocketProperties[RocketAssemblyList],...
                 ...,(other inputs: launch location, etc.),...,
                                 ...,CDModel[RocketAssemblyList],{x,y,z},t,Opts].
```

As we can see, the *RocketAssemblyList* is still appearing explicitly, wherever is necessary, and its entries can still be manipulated (e.g. changed in external cycle), or it could be defined previously. In this example the drag model also depends on the *RocketAssemblyList*; nevertheless, it still is a pure function on the coordinates, since it depends on the properties of the atmosphere (see below).

The most relevant inputs other than the rocket structure are the atmospheric density, the wind and the C_D models, together with the geographic location of the launch. All this is included explicitly in the input in order to be possible to change the models used. For example, one can use a simple constant for the drag coefficient or can use a more sophisticated model depending on the Mach number, and other parameters. Therefore it becomes easy to experiment and use new models required by any specific simulation.

One important feature is the inclusion in the input of the symbols representing the variables to be used in the simulation, namely in the integration of the equations of motion. This becomes necessary since the models used in the input can depend on those variables e.g. if we take compressibility effects into account, the C_D has a dependence on the atmospheric density that depends on the temperature that depends on the position, therefore the variables must be present in the input and included on the input function describing the C_D for identification. This makes possible to use analytical models that can appear explicitly on the input of the simulation. All functions must agree on what symbols represent position and time (in this case), which are the variables used in the propagation of the trajectory which is what the simulator does.

The output from the simulation are interpolated functions of the position (three components) and angle of attack as functions of time. The use of interpolated functions has the advantage of being differentiable in Mathematica© (therefore in the present case we can obtain the velocity and acceleration, and other flight data), save storage space, and being treated as analytical and hence in a adequate form for post processing.

The output is computed in three steps: one for the constrained trajectory due to the guiding rod, another for the climbing phase and the last one for the recovery trajectory. The angle of attack starts to be computed only in the second phase, since it is when the rocket suffers the first effect from the wind. The three parts of the trajectory are afterwards merged together.

If there is wind, the rocket needs to speed up until it becomes stable. Therefore, a guiding rod is required to restrain the trajectory during the launch in the desired direction. The rod's length is needed as input in order to determine the instant after which the rocket's trajectory becomes unrestricted and the model of the angle of attack starts to be computed. It is also recommended that the rocket's static margin is previously determined to assure it stays always positive (stable rocket).

During the simulation the rocket is treated as a particle since there are only three degrees of freedom and the model developed for the attitude is only used to estimate the loss of thrust along the true speed direction. Therefore, since the rocket's nose is always a steady point in the rocket's reference frame (unlike the CM), it is used to give the rocket's position over time and, consequently, the initial nose distance to the ground is required as input.

The rockets used in the simulator must always have, if any, three or more fins, boosters, or cluster motors — equally spaced — because the case of when the rocket has two components of each type has different transversal moments of inertia. This case was not considered since it is incompatible with the model developed for the angle of attack that assumes the rocket has only one transversal moment of inertia.

As default, the recovery system is deployed at the apogee. However, there is an optional argument that orders the simulator to start to compute the recovery trajectory at the ejection instant determined by the data within the rocket structure.

4.2 Atmospheric effects

The atmosphere is an important element influencing the trajectory since it is the fluid in which rockets fly. To simulate trajectories, in particular of model rockets, as these reach low apogees, it can be important to take into account the launch local conditions namely the weather conditions. To deal with sounding rockets, it is also required to know the atmosphere at high altitude.

We use an atmosphere model based on the Committee on Space Research (COSPAR) International Reference Atmosphere 1986 (CIRA-86), an empirical model ranging from 0 km to 120 km height when posterior refinements are included [21]. We also corrected some wrong entries of the original model, and organize data according to the structure depicted in Figure 10. Also, since pressure and temperature at the surface are not provided, the corrected model is improved by including these properties found by extrapolation at 0 km height. With this structure, pressure and temperature are easily interpolated in altitude, month and latitude.

This model is stored in a database in order to be easily explored and get access to

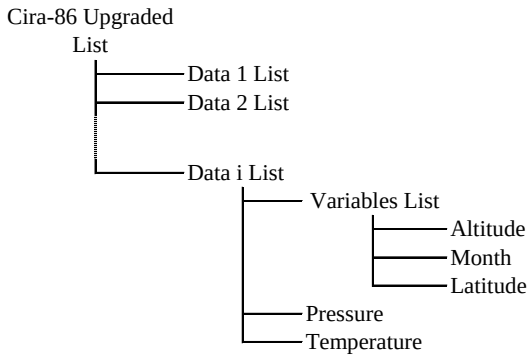


Figure 10: CIRA-86 upgraded version structure.

its properties and details. Also, this database allows the definition of other atmospheric models and new functions representing extra atmospheric parameters. The atmospheric database is a *listable* function (i.e., it is parsed to each element in a list), named *AtmosphereData* thus, like motors, different models can be defined and used. We can access its data by providing the name of the model and the required properties. For example,

```
AtmosphereData[``Cira86'', {``Temperature'', ``Density''}],
```

gives a list with the temperature and density functions for the CIRA-86 model. If another model is stored, we just have to change the name of the model to get these properties. Also, if we want to know the units of each property (or another information previously saved), we must type

```
AtmosphereData[``Cira86'', {``Temperature'', ``Density''}, ``Units''].
```

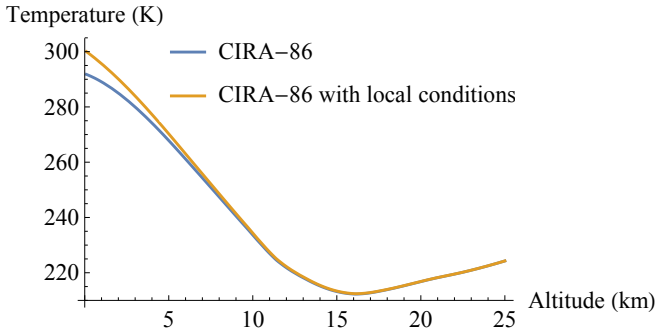


Figure 11: Temperature profiles for the CIRA-86 model.

At the launch site, the local conditions must be taken into account, especially for small

rockets. Therefore, we define a correcting function that corrects CIRA-86 at launch for the local conditions and asymptotically (exponentially, at a selected rate) tends to the original model when altitude increases. This correction can be applied to any atmospheric model that the user decides to use. An example of this correction is shown in Figure 11 for the temperature. After getting the adjusted pressure and temperature functions, the density is determined taking local weather conditions into account.

We also developed a wind model that takes into account the effects of the Atmospheric Boundary Layer (ABL) and the high altitude wind data provided by CIRA-86. This model requires the local wind speed at a certain altitude and the terrain's specification (described by the surface roughness length).

4.3 Effect of gusts

We also implemented a model of wind gusts, to simulate sudden gusts of wind that can happen during the launch, to evaluate possible undesirable effects. The effect of these perturbations is implemented in the *RocketTrajectory* function by changing the solution from the equation of the angle of attack to a new one with its initial angle condition determined by the rocket and wind gust's velocities at the perturbation's instant. Since the angle of attack in the simulator only affects the thrust force, the effect of wind gusts can only change the trajectory if the new solution gives a non zero angle during any burning.

The wind gusts' input is defined by a list containing as many lists as the number of gusts (see Figure 12). Each one of these lists contains the perturbation's instant (after

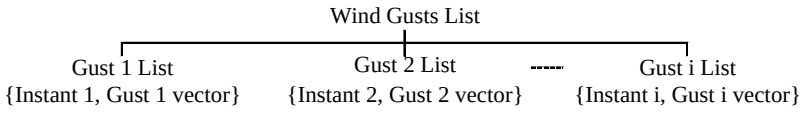


Figure 12: Wind gusts list structure.

the ignition) and the gust speed vector. The effect of a wind gust in the angle of attack at 2 s after the ignition is depicted in Figure 13.

4.4 Drag

We developed a drag coefficient model to be used in the simulator. If desired, another model can be developed which may depend on time, the coordinates variables and their derivatives. It is also possible to define a constant drag coefficient.

The developed model requires the list containing the structure of the rocket to determine the drag coefficients for each one of the external components and get the total C_D, depending on Mach, Reynold's number and time, with respect to the rocket's reference area. This drag model considers that all the components contribute to the skin friction

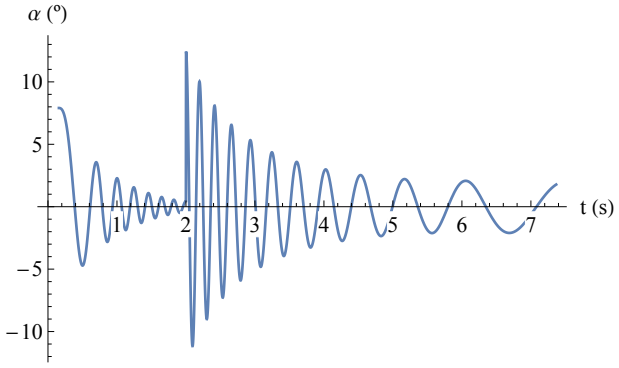


Figure 13: Angle of attack until the apogee with a wind gust perturbation.

drag but the pressure drag comes only from the nose, connectors and fins (due to the normal surface facing the flow). Also, in this model the rocket's nose is the only wave drag source.

4.5 Effects of the wind and atmospheric boundary layer

The lowest layer of the troposphere is directly influenced by the ground surface characteristics. Therefore, a wind speed profile is worth to be implemented since its effect will be noticed, especially in the case of small rockets. The thickness of the Atmospheric Boundary Layer (ABL) can change from about a hundred meters high to a few kilometers of altitude, varying in time, and with geographic region [22]. Above the ABL stays the *free atmosphere*, a more stable region where the frictional influences of the surface can be ignored and the wind is nearly geostrophic [22]. Regarding atmospheric stability, the ABL is classified in three different classes: neutral, stable and unstable. A neutral atmosphere implies an adiabatic lapse rate and no convection which is the case of a partially or highly cloudy atmosphere that may reduce the insolation at the surface [22]. Stable conditions occur mostly at night but can also appear when the ground surface is colder than the surrounding air. An unstable atmosphere is formed in clear weather during the day when there is high radiation from the sun causing ascending heat transfer. These conditions influence the wind profile and must be taken into account due to convective effects. The surface type also influences these phenomena.

4.6 Simulator's test and validation

We tested and validated the simulator by comparing the results with problems with known analytic solutions, namely a vertical rocket launch with constant thrust without atmosphere [23] to test the integration and a rocket in free fall with atmosphere reaching terminal velocity due to drag. No problems were detected.

Real rocket launches can provide more information about the performance of the sim-

ulator. It was not possible at this time to perform real tests but instrumenting a rocket or taking measurements from the ground to estimate the trajectory through triangulation will allow a stronger validation of the results.

5 APPLICATION: MODEL ROCKET WITH WIND

To demonstrate what the simulator can do we selected a model rocket and used its launch to test the models that take the wind and the atmospheric boundary layer into account.

5.1 Rocket launch

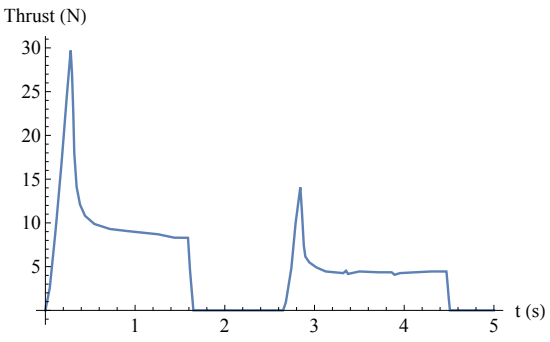
The results from the assembling of a small two-stage model rocket with $m = 114\,\mathrm{g}$ are presented in Figure 14.

Since model rockets reach lower velocities and apogees, they are suitable to observe the effects of the wind and the ABL on the trajectory simulation. We can see that the first stage's structure (not considering the connector) is discarded at the first burnout because of the first discontinuity. During the coasting phase (lasting 1 s), we see the connector's discharge represented by the second discontinuity. Except from these sharp changes, the rocket maintains its properties during coasting, leading the mentioned figures to match with Figure 14a when there is no thrust between the burnings. The reference area is defined as the largest cross section of the rocket (without considering boosters, for other cases) at each instant. This definition agrees with Figure 14f, where it shows the rocket's reference area changing nearly 2 s after the launch, which is the instant when the connector is discarded and the rocket loses its largest sectional area. Comparing Figures 14c and 14d, we conclude this rocket is always stable since the CM, instead of the CP, is closer to the nose all the time. Hence, the rocket is well assembled and able to be used in a trajectory simulation.

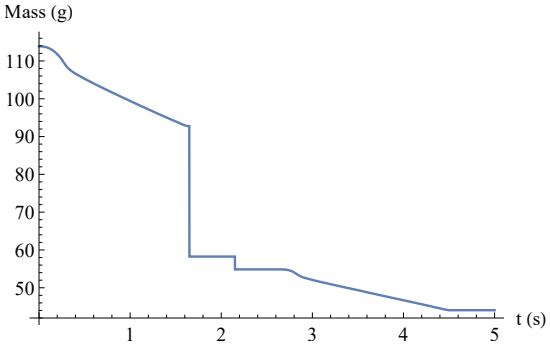
The obtained vertical model rocket launch trajectory with wind is represented in Figure 15. In this trajectory the weathercock effect is observable. Since the wind blows northeastward, the rocket climbs in the opposite direction (southwestward) to follow the true air speed. Although this effect increases smoothly with altitude due to the ABL, we can see a slightly discontinuity in the upward flight path around 200 m of altitude. This happens because the weathercock is enhanced during the coasting time, between the first burnout and the second ignition, as the rocket decreases its velocity.

During the recovery phase, the rocket slowly descends toward the downwind direction. Hence, it surpasses the launch site and lands 380 m north and 382 m east away from that place (coordinates taken from the simulator). The influence of the ABL is also observable during the recovery phase in which the trajectory describes a smooth curvature since the wind speed is decreasing as the rocket falls. If the wind was constant with altitude, the recovery trajectory would describe a linear path.

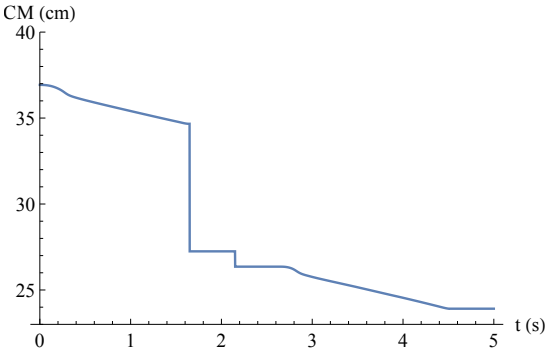
We can also compare the flight profiles from the model rocket launch with the same



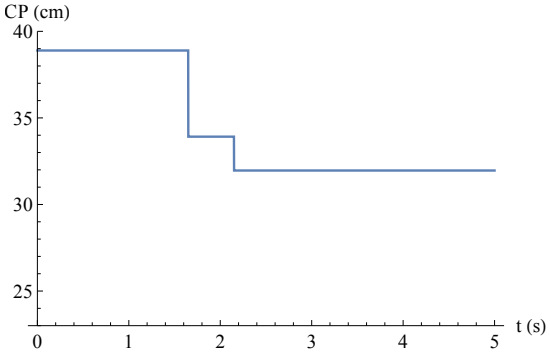
(a) Rocket's thrust.



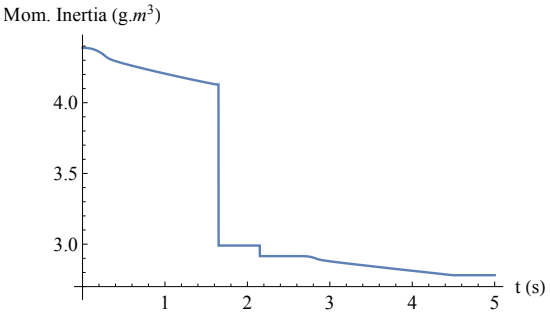
(b) Rocket's mass.



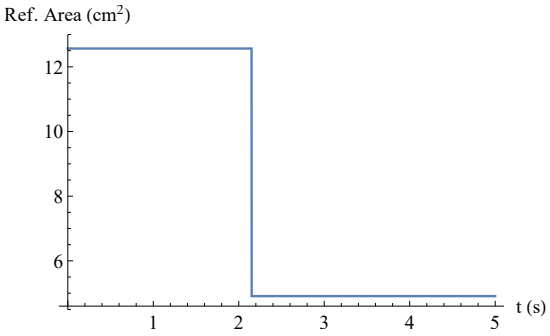
(c) Rocket's CM relative to the nose tip (reference).



(d) Rocket's CM relative to the nose tip (reference).

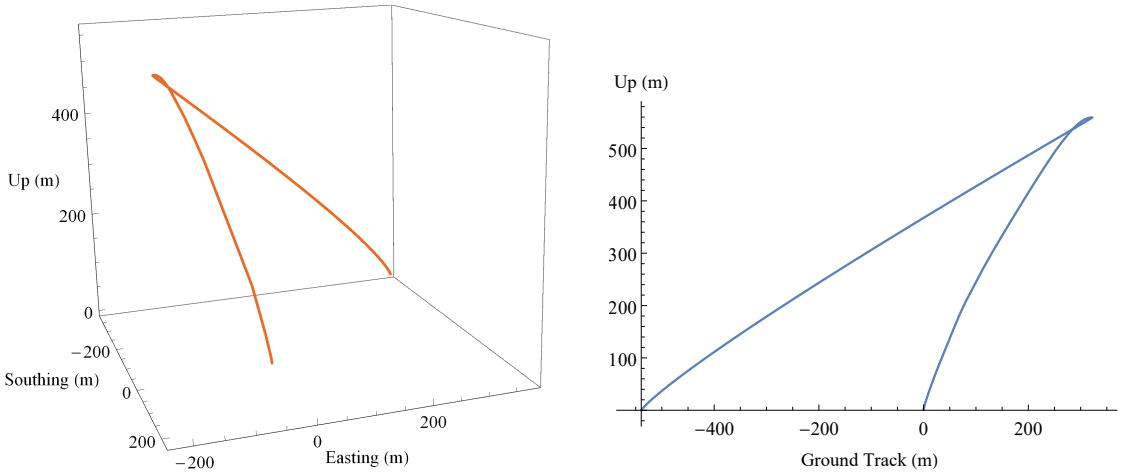


(e) Rocket's transv. mom. of inertia relative to the CM.



(f) Rocket's reference area.

Figure 14: Model rocket properties' functions.



(a) 3D trajectory.

(b) Trajectory profile.

Figure 15: Model rocket trajectory with a vertical launch and a neutral ABL; Parachute ejection at the apogee.

local wind speed used in the previous sections but changing the atmospheric stability. Figure 16 presents three wind profiles, under three different stable atmospheres, for a measured wind speed of $4.3\,\mathrm{m\,s^{-1}}$ at 6 m of altitude.

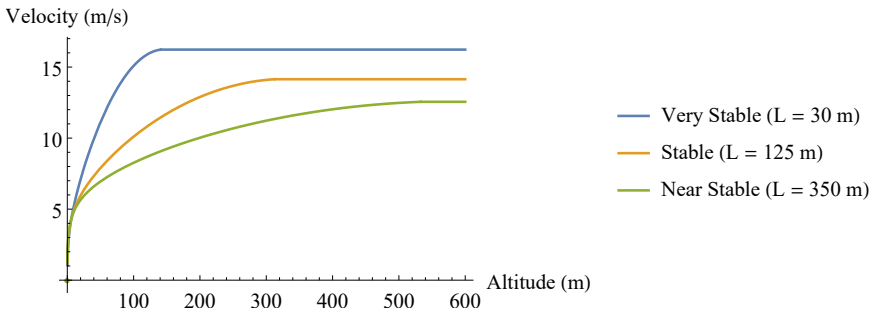


Figure 16: Wind profiles in a stable atmosphere for a measured wind speed of $4.3\,\mathrm{m\,s^{-1}}$ at 6 m of altitude.

We can see that as the atmosphere becomes more stable, the wind speed gets stronger at higher altitudes and the ABL thickness decreases. This effect mainly affects the descent phase of the rocket and results show that the rocket diverges more from the landing site under stable conditions as it corresponds to the strongest wind speed profile.

5.2 Stochastic simulations

The structure of the simulator allows to use stochastic methods to analyse the event or even to help design a better rocket. We can easily put the simulation function inside other function that allows to vary the inputs, generating many Monte Carlo type simulations. The only difficulty is that Mathematica© is an interpreted language, meaning that it is very time consuming to generate many simulations.

We illustrate this feature with a single-stage high power rocket in order to present results from a more powerful rocket. In Figure 17 we show the landing probability dis-

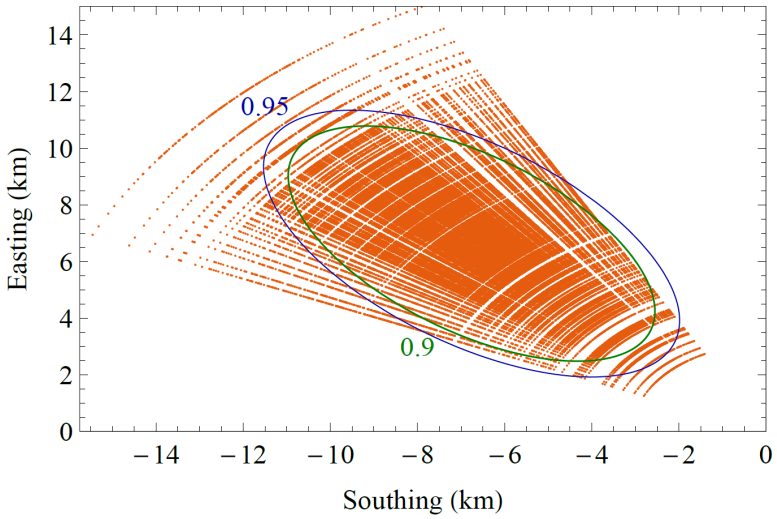


Figure 17: Landing positions and respective confidence ellipses from the Monte Carlo simulation.

persion foreseen for the case of considering wind with random variations in the direction, value of local wind, and the surface roughness.

It is possible even to vary the parameters of the rocket, effectively changing its properties and generating different rockets. In Figure 18 we can easily observe which design reaches higher within the allowed variations and what varying parameter has more impact in the results.

Even though these are simple examples, they provide a glimpse of the many possibilities at our disposal after developing a Monte Carlo simulation to our tool.

6 CONCLUSIONS

The present work focused on the development of a trajectory simulator tool for fin-stabilized model and sounding rockets, that supports many types of rockets, and is capable of addressing many physical effects influencing the problem. The tool was developed using

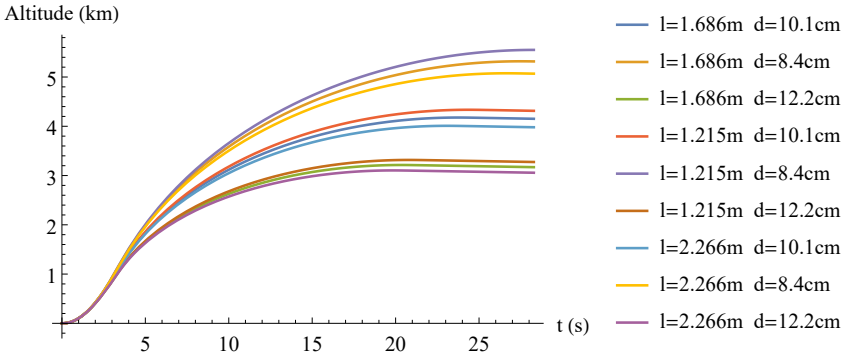


Figure 18: Altitude over time from nine launches with different rocket's length and diameter configurations.

Mathematica© and supports from model to sounding rockets with any number of stages. We focused on presenting the tool's flexibility to customize the default models and to compute the outcome from the inputs' uncertainties (or perform an optimization) from a Monte Carlo simulation.

The developed data structure describing the rocket was presented and is adaptable to any type of common rockets. We also illustrate how models such as the one for the atmospheric can take into account local conditions at the launch and other atmospheric features, such as the atmospheric boundary layer and the wind.

We illustrate the possibilities of the framework with examples, including the ability to refine the rockets' design by optimizing its characteristics. We also show how functions that depend on the variables of the simulation can be included on the simulation input provided that the symbols of the variables are also included on the input. The most difficult part was to integrate the data structure model and the atmosphere and other models in a logical way and without any hidden calls. We were able to achieve this by including all the required elements in the input of the simulation function, but for more complex cases it can be difficult to manage. It must be evaluated in the future if it is reasonable to invoke some functions, such as the atmosphere model, quietly (i.e. not explicitly in the input of the trajectory simulator) to be used by the simulator. It will nevertheless be necessary to assure control of the simulation by allowing only a set of specific functions with a certain format or inputs, otherwise it will be difficult not to change the simulator function many times.

Another possible development is to include a control law and try to study the outcome of perturbations such as gust of wind.

REFERENCES

[1] National Association of Rocketry. Nnational Association of Rocketry. `http://www.nar.org`. Accessed March 15, 2017.

[2] Wolfram Research, Inc., Mathematica, Version 10.2, Champaign, IL, 2015.

[3] W. Colburn. Where Did Model Rocketry Really Start? *Peak of Flight Newsletter*, (314), June 2012.

[4] National Association of Rocketry. High Power Rocketry. `http://www.nar.org/high-power-rocketry-info`. Accessed September 3, 2015.

[5] Sampo Niskanen. Development of an Open Source model rocket simulation software. Master's thesis, Helsinki University Of Technology, 2009.

[6] P. Fossey. *RockSim Program Guide*. Apogee Components, 2003.

[7] T. Milligan. Phases of a Model Rocket's Flight. *Peak of Flight Newsletter*, (117), December 2003.

[8] G. H. Stine. *Forty Years of Model Rocketry – A Safety Report*. National Association of Rocketry, 1997.

[9] J. Barrowman. Stability of a Model Rocket in Flight. Technical Information Report 30, Centuri Engineering Company, 1970.

[10] T. Milligan. How To Multi-Stage Rockets Work – Part 1. *Peak of Flight Newsletter*, (98), February 2003.

[11] V. Estes. *Multi-Staging*, 1963. Estes Industries Technical Report TR-2.

[12] T. Milligan. How To Multi-Stage Rockets – Part 2. *Peak of Flight Newsletter*, (99), February 2003.

[13] A. Tewari. *Atmospheric and Space Flight Dynamics*. Birkhauser, 2007.

[14] Estes Industries. *Cluster Techniques*, 1967. Technical Report TR-6.

[15] V. I. Feodosiev and G. B. Siniarev. *Introduction to Rocket Technology*. Academic Press Inc., 1959.

[16] F. Beer, E. Johnston, D. Mazurek, P. Cornwell, and E. Eisenberg. *Vector Mechanics for Engineers: Statics and Dynamics*. McGraw-Hill, 9th edition, 2010.

[17] J. Barrowman. Calculating the Center of Pressure of a Model Rocket. Technical Information Report 33, Centuri Engineering Company.

[18] S. F. Hoerner. *Fluid-Dynamic Drag*. Published by the author, 1965.

[19] E. Rathakrishnan. *Theoretical Aerodynamics*. Wiley, 2013.

[20] John Coker. ThrustCurve Home. `http://www.thrustcurve.org/`. Accessed March 15, 2017.

[21] ANSI/AIAA. *Guide to Reference and Standard Atmosphere Models*, February 2009.

[22] R.B. Stull. *An Introduction to Boundary Layer Meteorology*. Kluwer Academic Publishers, 1988.

[23] H. Curtis. *Orbital Mechanics for Engineering Students*. Elsevier, 2nd edition, 2010.

SYMCOMP 2017
Guimarães, 6-7 April 2017

A HIGH-ACCURATE SPH-MOOD METHOD

Xesus Nogueira[1*], Luis Ramirez[1], Stéphane Clain[2], Raphael Loubère[3], Ignasi Colominas[1], Luis Cueto-Felgueroso[4]

1: Department of Applied Mathematics, Universidade da Coruña, Spain
2: Universidade do Minho, Portugal
3: Univeristé de Toulouse, France
4: Universidad Politécnica de Madrid

e-mails: xnogueira@udc.es ; luis.ramirez@udc.es ; clain@math.uminho.pt ; raphael.loubere@math.univ-toulouse.fr ; icolominas@udc.es ; luis.cueto@upm.es

Keywords: Smoothed Particle Hydrodynamics, Compressible flow, Moving Least Squares

Abstract *The Smoothed Particle Hydrodynamics (SPH) method is based on a Lagrangian formulation and it has been widely used in CFD applications. It was developed in the 1970's for astrophysical applications [1,2], and since then it has been applied to a great variety of problems. Most of SPH formulations are based on the artificial viscosity approach, but these formulations present a number of problems, particularly when applied to flows with shock waves. In these cases, the method loses accuracy near shocks and contact discontinuities, and smeared shock fronts and strong glitches near contact discontinuities may appear [3].*

Here we present an accurate, stable and low-dissipative SPH Riemann-based method, based on the formulation introduced by Vila [4]. Explicit artificial viscosity is substituted by the use of Riemann solvers, and high-order is achieved by using Taylor reconstructions for the evaluation of the numerical fluxes at the midpoint between two interacting particles. The derivatives required for the reconstructions are computed using Moving Least Squares. The stability of the numerical scheme is achieved by using the Multidimensional Optimal Order Detection (MOOD) paradigm [5,6].

REFERENCES

[1] L.B. Lucy, A numerical approach to the testing of the fission hypothesis, Astronomical Journal, 82, 1013-1024, 1977.
[2] R.A. Gingold, J.J. Monaghan, Smoothed Particle hydrodynamics-Theory and application to non-spherical stars, Monthly Notices of the Royal Astronomical Society,181, 375-389, 1977.
[3] F.V. Sirotkin, J.J. Yoh, A Smoothed Particle Hydrodynamics method with approximate Riemann solvers for simulation of strong explosions, Computers and Fluids, :418-429, 2013.
[4] J.P. Vila, Particle Weighted Methods and Smooth Particle Hydrodynamics, Mathematical Models and Methods in Applied Sciences, 9(2), 161-209, 1999.
[5] S. Clain, S. Diot, R. Loubère , A high-order finite volume method for systems of conservation laws- Multidimensional Optimal Order Detection (MOOD), Journal of Computational Physics, 230:4028-4050, 2011.
[6] S. Diot, S. Clain, R. Loubère, Improved detection criteria for the Multidimensional Optimal Order Detection (MOOD) on unstructured meshes with very high-order polynomials, Computers and Fluids, 64: 43-63, 2012.

SYMCOMP 2017
Guimarães, 6-7 April 2017

ON THE CONVERGENCE OF NEWTON'S METHOD FOR EIGENVALUES OF SYMMETRIC TRIDIAGONAL MATRICES

Rui Ralha[1]*, Carlos Campos[2]

1: Universidade do Minho, Portugal
2: Instituto Politécnico de leiria, Portugal

e-mails: r_ralha@math.uminho.pt ; carlos.campos@ipleiria.pt

Keywords: Accurate eigenvalues, bisection, Newton's method

Abstract

For symmetric tridiagonal matrices, eigenvalues computed with the fast state-of-the-art LAPACK codes are not always as accurate as the data warrant.

We have observed that the routine DSTEBZ, which is an implementation of the bisection method, is the only one that consistently delivers such accuracy, when the appropriate stopping criteria is enforced. Because bisection delivers only one correct bit per iteration, it is natural to combine bisection with methods that have a better asymptotic convergence rate.

In this paper we consider the use of Newton's method for this purpose and present some theoretical results on its convergence. Numerical examples are given.

SYMCOMP 2017
Guimarães, 6-7 April 2017

INVESTIGATION OF NANOELECTRONIC AND NANO-OPTOELECTRONIC CIRCUITS USING MATLAB AND SIMULINK

João F.M. Rei[1]*, James A.M. Foot[1] , Gil C. Rodrigues[1] and José M.L. Figueiredo[1]

1: Universidade do Algarve, Portugal

e-mails: a40652@ualg.pt ; a40650@ualg.pt ; a39017@ualg.pt ; jlongras@ualg.pt

Keywords: Nanoelectronics, nanooptoelectronics, terahertz, resonant tunneling diode, laser diode, simulation

Abstract *Nanoelectronics and nanophotonics are emerging as major technologies. Devices and circuits based on these technologies are expected to be lightweight, highly efficient, low energy consuming, and are cost effective to produce. Several applications exploiting the interaction of light-emitting and light-sensing nanostructured materials have been proposed. Resonant tunneling diodes (RTDs) and optoelectronic circuits employing RTDs structures are object of abundant investigation towards the implementation of electronic-to-optical and optical-to-electrical conversion modulates for application in future ultra-wide-band wired and wireless communication systems. The RTD shows wideband negative differential conductance (that is, electrical gain) and are among the highest frequency operation nanoelectronic semiconductor devices capable to produce oscillation up to 2 THz, at room temperature. Moreover, due there high non-linear current-voltage characteristics they can give rise to circuits with new functionalities. They have potential applications including low energy consumption high frequency signal generation. When integrated with semiconductor laser diodes (LDs) the combination can work as an efficient and compact microwave to lightwave transducers that take advantage of the combination of the RTD and LD high non-linearities. The recent attention on these optoelectronic systems (iBROW H2020 EU Project) led to the need of the development of numerical and symbolic mathematical tools to design and investigate these novel electronic and optoelectronic circuits. Here we present RTD and RTD-LD oscillator simulation packages based on MATLAB (GUI) and SIMULINK (models) that allows signal level analysis such as transient and frequency responses. The modeling packages being implemented solve the coupled RTD electronic circuit differential equations (Liénard system) and the laser diode rate equations with a fixed-step 4th order Runge-Kutta algorithm for quick and quite accurate simulations. The simulations range from the gigahertz (nanosecond) to terahertz frequencies (picosecond) and are relatively fast lasting few minutes, show good agreement with experimental results. We also evaluate and discuss the simulated behavior of numerical equivalent experimental tested RTD electronic and optoelectronic circuits.*

SYMCOMP 2017
Guimarães, 6-7 April 2017

TOWARDS DAMAGE QUANTIFICATION CAUSED BY DRILLING IN FIBRE COMPOSITE LAMINATES

M.S.F. Alves[1,2], I.C.J. Barbosa[1,2,3], I.M.F. Bragança[1,2,3], M.A.R. Loja[1,2,3]

1: ISEL - Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal.
2: GI-MOSM, ADEM, ISEL – Grupo de Investigação em Modelação e Optimização de Sistemas Multifuncionais
3: LAETA, IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais, 1, 1049-001 Lisboa, Portugal.

e-mails: {MarisaSofia_94@hotmail.com, ines.barbosa@dem.isel.pt, ibraganca@dem.isel.pt, amelialoja@dem.isel.ipl.pt}

Keywords: Fibre Reinforced Composites, Damage, Thermography, Digital Image Processing.

Abstract

Drilling operations are a very frequent need when one needs to connect different components in a structure, regardless the nature of the materials involved. When these components are total or partially made of composite materials, this manufacturing operation requires a particular care, as the heterogeneous character of composites makes them especially prone to delamination or fibre pulling-out. These undesirable situations, not only contribute to reduce the stiffness/strength of the material in the drilled region neighbourhood, but they may even lead to a subsequent damage propagation and failure.

It is therefore important to improve drilling processes efficiency from the damage minimization perspective. To this purpose, the characterization of the drilling parameters influence in a cross-relation to the composite material characterization may be a relevant contribution.

With the present work, one intends to present preliminary results associated to the experimental procedure that will enable the characterization of the influence of drilling parameters and of material and geometrical specimen parameters on the drilling affected areas. The information acquisition of the affected areas is carried out through thermographic digital images. A set of illustrative cases is presented and preliminary conclusions are drawn.

1. INTRODUCTION

In view of the increasing use of composite materials in today's world, composite materials are one of the most interesting groups of materials in our technological society, as they allow the combination of specific properties that are not achieved with another group of materials, such as, the combination of low weight and high strength [1].

Laminated composite materials with long fibre reinforcement in unidirectional or woven fabric are used in numerous applications ranging from the design, construction and repair of components and systems in the most diverse industries. The constant need for assembly of components requires, in most cases, drilling processes, which in addition to weakening the area in which it is carried out, can cause damage to the material and consequently accelerate its degradation until fault occurs. To enable the characterization of damage occurrences the most used non-destructive tests are thermography, x-ray, and c-scan ultrasound techniques.

The main types of damage produced on these materials by drilling are thermal damages (burning of the matrix), pulling of fibres, defects of circularity, and delamination.

The damage that causes greater concern due to the difficulty that exists in the contouring is the occurrence of delamination which is marked by the detachment of adjacent layers of the laminate. This damage occurs when the force exerted on the material is superior to the interlaminar resistance of the laminate. The force that initiates the delamination is the so-called critical force, which may be formulated differently according to different authors [2][3][4][5].

Two types of delamination may occur and they are referred to in the literature as "peel-up" delamination when the damage occurs at the initial contact time of the drill with the material. In this situation, the material tends to be pealed up along the cutting edges of the drill due to the abrasion. The push down effect occurs when the drill approaches the last layer to be bored due to the compressive force, when the layer tends to deflect, causing the rupture of the connections between the consecutive layers [6][7].

After several tests carried out by different authors, they concluded that the peel-up delamination is easily avoided with the use of moderate cutting parameters, but the push-down delamination is not so easily eliminated. However, several methods have been developed to minimize the extent of this damage [8][9]. One of these implies the use of a force lower than the critical one, which may be analytically predicted and according to which the cutting parameters are adjusted [2][3][4][5].

Capello tested and verified the minimization of delamination by using a support base during drilling which prevents the deflection of the material during the drilling operation, thus minimizing the detachment of the lower layers of the laminate [10].

The damage caused by drilling is essentially quantified in two ways, which take into account different aspects. The first takes into account the quotient between the maximum diameter (D_{max}) reached after drilling and the nominal diameter (D_0) denominated by damage factor (F_d) [11], while the second takes into account the quotient between the damage area around the hole (D_{MAR}) and the average hole area (D_{AVG}), denominated by damage ratio (D_{RAT}).

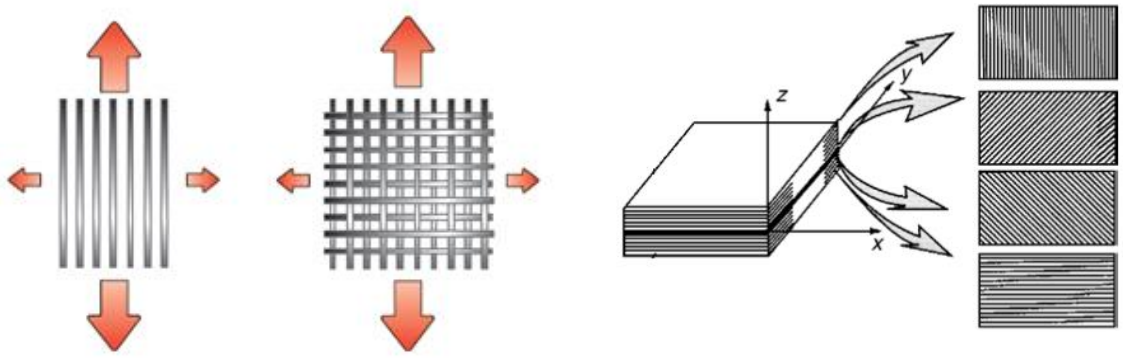
After several drilling tests of carbon fibre reinforced composite, Durão et al obtained a different factor that is related to the delamination of the carbon fibre reinforced composite,

called the variable delamination factor. For these composites, this factor was found to be between 1 e 2 [12].
Considering the mathematical formulation of the damage factor, Ramesh et al [13] studied the influence of the temperature during the drilling process on the quality of the obtained hole, by monitoring the damage factor and the average temperature reached in the drilling process without cooling system and with an internal and external cooling system. It was possible to conclude that a lower damage factor was associated to the use of cooling systems.
This work presents a preliminary characterization study of the affected regions during drilling operations. To this purpose an experimental setup was established in order enable the thermographic acquisition. Some images were further processed in order to obtain measures and calculate damage parameters.

2. MATERIALS AND METHODS

2.1. Composite laminates

The most commonly used long fibre reinforced composite materials may be unidirectional or woven constituting a fabric, as we may observe in Figure 1. Due to known reasons, the woven configuration presents a global improved performance, and a lower probability of occurrence of delamination [14].



(a) (b) (c)

Figure 1- (a) Unidirectional fibres; (b) Woven fibres [14]; (c) Layers' stacking [15]

However, the unidirectional layers are often considered in a stacking configuration as it is possible to observe in Figure 1-c). The orientation of these fibres can also be arranged aiming a better performance according to specific operating conditions, since their orientation influences the transformed reduced elastic coefficients of each layer, according to:

$$\begin{aligned}\bar{Q}_{11} &= Q_{11}m^4 + 2(Q_{12} + 2Q_{66})n^2m^2 + Q_{22}n^4 \\ \bar{Q}_{12} &= (Q_{11} + Q_{22} - 4Q_{66})n^2m^2 + Q_{12}(n^4 + m^4) \\ \bar{Q}_{13} &= Q_{13}m^2 + Q_{23}n^2 \\ \bar{Q}_{16} &= -Q_{11}mn^4 + Q_{11}m^3n - (Q_{12} - 2Q_{66})mn(m^2 - n^2)\end{aligned} \quad (1)$$

$$\bar{Q}_{22} = Q_{11}n^4 + 2(Q_{12} + 2Q_{66})n^2m^2 + Q_{22}m^4$$
$$\bar{Q}_{23} = Q_{13}n^2 + Q_{23}m^2$$
$$\bar{Q}_{26} = -Q_{22}m^3n + Q_{11}mn^3 + (Q_{12} + 2Q_{66})mn(m^2 - n^2)$$
$$\bar{Q}_{33} = Q_{33}$$
$$\bar{Q}_{36} = (Q_{13} - Q_{23})mn$$
$$\bar{Q}_{44} = Q_{44}m^2 + Q_{55}n^2$$
$$\bar{Q}_{45} = (Q_{55} - Q_{44})mn$$
$$\bar{Q}_{55} = Q_{55}m^2 + Q_{44}n^2$$
$$\bar{Q}_{66} = (Q_{11} + Q_{22} - 2Q_{12})m^2n^2 + Q_{66}(m^2 - n^2)^2$$

where m, n stand for $\cos(\theta), \sin(\theta)$, being θ the angle between the positive *x* and *1* directions of the laminated and the *k*-th layer in general terms [16]**.** Therefore, it is expected that layer stacking influences the effects of drilling operations.

2.2. Experimental method

As mentioned, thermography is one of the non-destructive techniques used in the analysis of damage in composite drilling. The present work aims to establish a methodology to enable mapping the distribution of temperatures reached along the surface of the material, so that damaged regions can be perceived, since temperatures in those regions will present a different pattern. The thermographic camera used was a TROTEC ec060 camera with the following characteristics: resolution of 160×120, measuring temperatures in the range of: $-2°C$ – $250°C$ and a laser wavelength: $635\ nm$ (red). A schematic representation of the setup used to drill the glass fibre specimens in the computer numerical command machine (CNC) is presented in Figure 2.

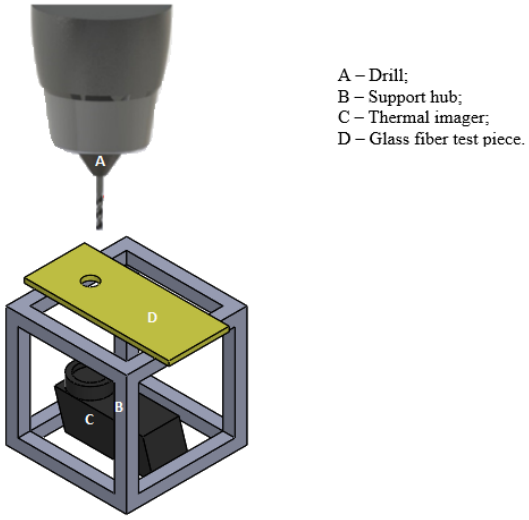


Figure 2- Schematic illustration of experimental setup used.

The specimen was fixed with hub clamps and inside the hub one has placed the thermographic

camera, in turn, the hub is fixed to the working table of the CNC machine by a mechanic table top.
The images acquired are processed in order to yield a configuration that may allow for the measurement of some metrics to characterize damage according to different authors. To achieve this, a threshold analysis is developed. The image histogram is used to identify the values to be used in the segmentation process, thus isolating the pixels presenting intensity within the range of values of interest.

2.3. Damage assessment methodologies

As the use of damage assessment parameters is intended, in the present sub-section a summarized description of methodologies available in the literature is presented. Singh et al [8] used x-ray images to analyse the damage produced by drilling operations on carbon fibre reinforced composites, using different drill geometries. The test pieces were previously immersed in a contrast medium for one hour so that the area of damage was visible on the radiography. Afterwards, they processed the images using a neural network that allowed segmenting the images. Figure 3 illustrates one of the images obtained by these authors. Using a threshold limit, they determined the damage factor as Chen [11]:

$$F_d = \frac{D_{max}}{D_0} \tag{2}$$

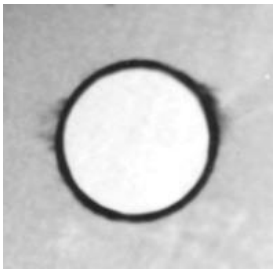


Figure 3 – Contrasting digital radiography of a perforated composite plate [8].

Durão et al used two techniques to develop a damage analysis, both C-Scan ultrasound and radiographic analysis. To characterize the damage these authors took into account not only the definition of Chen, as stated in Equation (2) [11], but also the Mehta et al definition [12]:

$$D_{RAT} = \frac{D_{MAR}}{D_{AVG}} \tag{3}$$

The C-Scan ultrasound analysis was performed using a probe that emits pulses at a certain frequency towards the specimen and a receiver that receives the signals reflected by that specimen, transforming them into electrical signals that are amplified. This allows obtaining an image in which colours scale or grayscale corresponds to different broadcast-receiving intervals. The ultrasound due to its wave propagation direction characteristics is capable of detecting delamination that occurs perpendicularly to the bore. The images were further

processed using an image processing and analysis platform that uses segmentation and threshold techniques. Figure 4 presents the original image obtained and the segmented one.

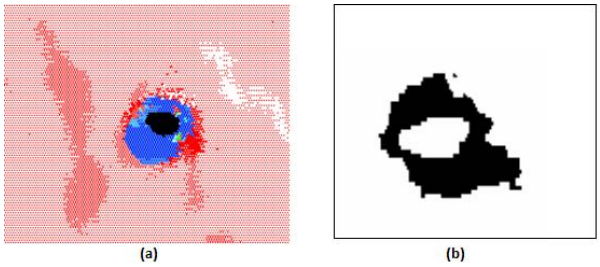


Figure 4 – (a) Image generated by ultrasound; (b) Image obtained after threshold [12].

Radiography was performed after the specimens had been immersed in diiodomethane for about one and half hour. The apparatus used was a Toshiba DG-073B with a focal length of 70mm and a time of exposure of 0.25 seconds. The maximum diameter measurement was carried out with the aid of an optical microscope. The processing of the radiographic images followed a three-stage process, in which the area of interest was selected in the first phase (Figure 5 (a)); in the second phase, a smoothing filter was applied to reduce abrupt variations in intensity and thus attenuate the area of interest (Figure 5 (b)) in order to, in the third phase, finally segment the areas of interest and subsequently measure the extent of the damage, using also the microscopic optics [12].

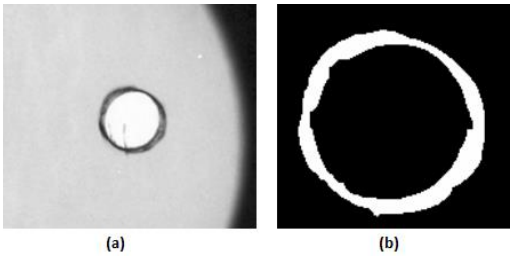


Figure 5 - (a) Image generated by x-ray; (b) Binary image [12].

Yang et al [17] verified that the use of thermography has a great potential and advantages in the detection of defects induced by drilling of composite materials. The authors performed thermographic analysis on drilling with and without the use of an external heat source, using various relative positions of the thermographic camera, among other aspects. The subsequent analysis process consisted in the evaluation of the temporal evolution of the surface temperature, thus allowing the detection of surface damage, since the presence of defects reduces the diffusion rate and the damaged zones have different temperatures when compared to the surrounding material. The major disadvantage of this method is that deep damage is observed with reduced temperature contrasts which make it difficult to detect.

Bhatnagar et al [18] used a less common technique. They sprayed the area of the hole with a contrast agent (Zyhlo 27-A) and allowed it to dry for about 30 minutes. The fluorescent dye penetrated the damage zone and gave a clear image of the drilling-induced damage around the hole.

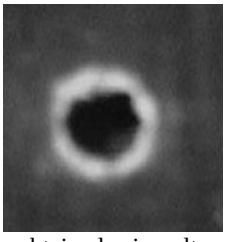


Figure 6 – Image obtained using ultra violet light [18].

The damaged area was visible and the image processing principles of segmentation and threshold were used. The resulting image is presented in Figure 6 [18].

3. RESULTS AND DISCUSSION

In order to validate the present study and preliminary results, a first study was carried out using two images (Figure 5 and Figure 6) from other authors. Two stages were considered: a first stage where the image was processed to achieve an appropriated format to obtain the required measurements, and a second stage where some damage assessments were made by calculating the damage parameters referred in sub-section 2.3.

3.1. Validation study

The validation study considered the images published in the papers of Durão et al [12] and Bhatnagar [18].

To enable obtaining the measurements needed for the damage factors, the images were first subjected to a processing procedure, comprising the following steps:

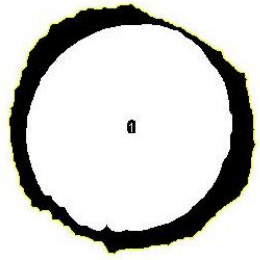
- Calibration considering a known distance in the image;
- Selection of the area of interest;
- Conversion of the original image to a binary image;
- Application of bandpass filter;
- Identification of the limits of the grey scale to be used
- Application of threshold limits;
- Calculation of the maximum diameter and area of damage.

The analysis data for Figure 5 (a) obtained by the authors using a freeware application, ImageJ [19], is presented in Table 1.

Figure 5						
	$D_{MAR}\ [mm^2]$	$D_{AVG}\ [mm^2]$	D_{RAT}	$D_0\ [mm]$	$D_{max}\ [mm]$	F_d
Durão et al [12]	-	-	1,394	6	-	1,188
Present study	38,217	28,274	1,352	6	9,517	1,586
Deviation[%]	-	-	3.01	-	-	33.5

Table 1 - Comparison of the damage factor and the damage ratio for image in Figure 5.

It is possible to conclude that the relative deviation is acceptable for the damage ratio, although very high for the damage factor. This may be justified by the fact that the image used in the present calculations is not the original image, thus not so accurate in terms of definition.
The image obtained by the manipulation of Figure 5 is shown in Figure 7, which can be compared with the binary image obtained by the reference authors in Figure 5 (b). From these images it is possible to conclude on a similar morphology of the damage, as it would be expected.



	Damage Area [mm^2]
1	38,217

Figure 7 – Binary image obtained and corresponding damage area.

The area shown in the Figure 7, refers to the area of damage generated by the composite drilling, calculated using ImageJ [19], which in this case is $38{,}217\text{mm}^2$.
The analysis data obtained for Figure 6 is presented in Table 2.

Figure 6			
	$D_{MAR}\ [mm^2]$	$D_{AVG}\ [mm^2]$	D_{RAT}
Bhatnagar et al [18]	-	-	2,4
Present study	95,730	50,265	1,904
Deviation [%]	-	-	20.67

Table 2 - Comparison of the damage ratio of Figure 6.

In this case, a significant relative deviation for the damage ratio was also obtained, which may be explained by the same reasons. The image obtained after manipulation of Figure 6 is shown in Figure 8.

	Damage Area [mm^2]
1	95,73

Figure 8 – (left) Binary image and (right) corresponding affected area.

The area shown in Figure 8 is the area affected by damage induced by drilling which, according to the ImageJ [19] application is 95.730mm^2.
As reference authors give no details concerning the image binarization and segmentation and further measurement methodologies used, the deviations may also be related to differences at these stages.

3.2. Preliminary experimental results

To carry out the experimental activity, a diamond-coated double-helix milling cutter was used with a diameter of 6mm and fibreglass composite test pieces with $200 \times 36 \times 2.6mm$ of epoxy matrix with glass woven sheet and two plates carbon fibre, one with $150 \times 100 \times 4.4mm$ of epoxy matrix with stacking [(+45/0/-45/90)4]S and other with $150 \times 100 \times 3.2mm$ of epoxy matrix with stacking [-45/+90/+45/0]S. The drilling parameters for the tests are presented in Table 3.

Test piece	Cutting speed [rpm]	Feed [mm/min]
Fibreglass	5500	350
Carbon fibre [(+45/0/-45/90)4]S	7000	550
Carbon fibre [(+45/0/-45/90)4]S	7000	750
Carbon fibre [-45/+90/+45/0]S.	7000	750

Table 3 - Drilling parameters.

The thermographic images obtained in these experiments are presented in Figure 9.

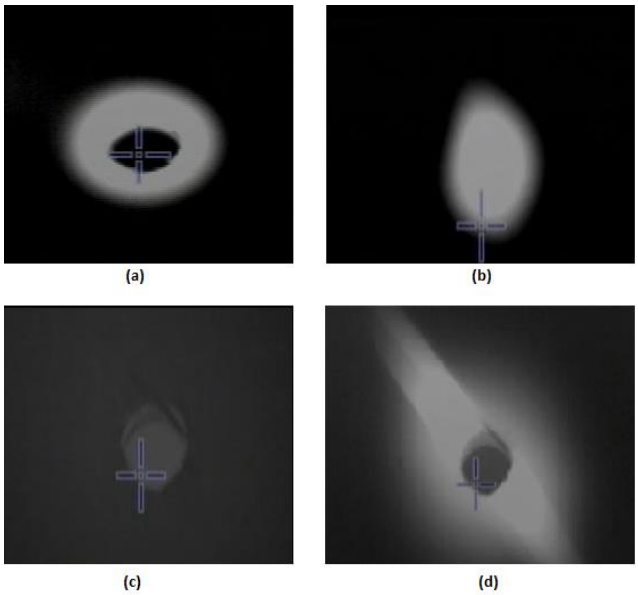


Figure 9 – Thermographic images: (a) Fiberglass plate drilling; (b) Carbon fiber plate drilling ($150 \times 100 \times 4.4mm$); (c) Carbon fibre plate drilling ($150 \times 100 \times 3.2mm$); (d) Carbon fiber plate drilling ($150 \times 100 \times 3.2mm$) just after the drill bit.

In the first image (Figure 9 (a)), it is possible to observe that a circularly shaped configuration was not obtained. This was due to a misalignment of the camera. In Figure 9 (b), the hole is not perceptible without using the threshold since at the time of image capture the drill was still in the hole. Figure 9 (d) is the same hole as in Figure 9 (c), the difference between these images is the time at which they were captured, Figure 9 (c) was captured after the specimen was cooled and Figure 9 (d) was captured immediately after the removal of the drill bit, where the temperature distribution is still visible along the fibres which during flexural drilling produce fibre heating visible in the image.

To process these images, the methodology previously mentioned through ImageJ [19] application was used. The first phase of the methodology consisted on the selection of the area of interest. Then the image was converted into a binary image and there were no evident differences in the binary image obtained. At this stage a bandpass filter was applied. This filter is used because it removes the high spatial frequencies (blurring the image) and low spatial frequencies (similar to subtracting a blurred image). It also supress horizontal or vertical stripes that were created by scanning an image by line. The bandpass filter uses a special algorithm to reduce edge artefacts (before the Fourier transform, the image is extended in size by attaching mirrored copies of image parts outside the original image, thus no jumps occur at the edges). The resulting image is presented in Figure 10.

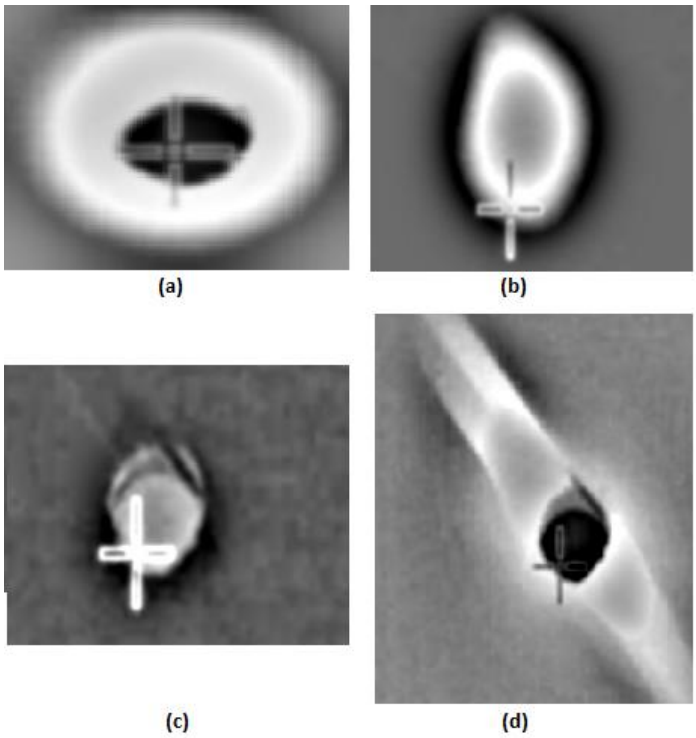


Figure 10 - Images with bandpass filter.

Finally the thresholding phase is performed in order to segment the image. The threshold tool

identifies the limits of the grey scale to be used, considering the histogram of the image obtained after applying the bandpass filter (Figure 10 a)) presented in the Figure 11. In this histogram the highest grey pixel intensity range is visible, the same corresponding to the damage zone. After segmenting the image to this interval, the affected area becomes clear and easy to determine. The same analysis was done for the remaining images.

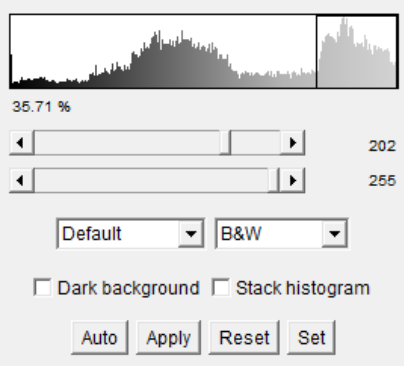
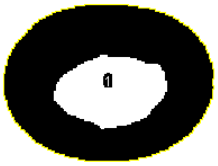


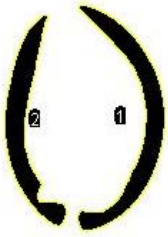
Figure 11 – Threshold.

The calculation of the affected area (black zone of Figure 12, 13, 14 and 15) can finally be obtained.



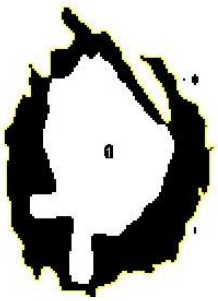
	Affected Area [mm^2]
1	259,811

Figure 12 – Affected area of the Figure 10 (a) obtained after applying the bandpass filter.



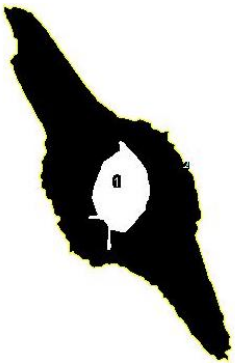
	Affected Area [mm^2]
1	37,314
2	31,477
Total	68,791

Figure 13 - Affected area of Figure 10 (b).



	Affected Area [mm^2]
1	88,66

Figure 14 - Affected area of Figure 10 (c).



	Affected Area [mm^2]
1	384,602

Figure 15 - Affected area of Figure 10 (d).

Considering that the thermally affected area coincides with the area of damage obtained (D_{MAR}) and assuming that the average area of the hole is equal to the area of the hole made with a $6mm$ diameter drill bit ($D_{AVG} = 28{,}274mm^2$), it follows that the calculation of the damage ratio formulated by Mehta et al (equation (3)) for each image obtained is presented in (Table 4).

Figure 10	$D_{MAR}\ [mm^2]$	$D_{AVG}[mm^2]$	D_{RAT}
(a)	259,811	28,274	9,189
(b)	68,791	28,274	2,433
(c)	88,66	28,274	3,136
(d)	384,602	28,274	13,603

Table 4 - Damage ratio of each image.

The high value ratio of Figure 10 (a) is due to the mentioned misalignment of the camera. In fact the temperatures profile developed during the drilling process follow a circular

distribution around the hole and consequently in the present case (ellipse) a greater damage area was obtained which influences the value of the damage ratio.
In Figure 10 (b) and (c) the damage ratio is lower and closer to the values of damage ratio obtained by Bhatnagar et al [18] for carbon fibre reinforced composites, although it is perceptible in these images that the presence of the thermographic camera focus indicator influences the determination of the area of damage.
Finally the damage ratio of Figure 10 (d) is much higher than the rest. This may be due to the fact that in this test delamination of the fibres occurred, as can be seen in the image, and so the affected area is larger in the direction of the delamination.

CONCLUSIONS

With the present work it is intended to develop an experimental methodology that enables the characterization of damaged and thermally affected areas by drilling operations on composite materials.
From the preliminary results obtained it is possible to verify that, by calculating the area affected by temperature, it is possible to evaluate on the extent of the damage, assuming the existence of a correspondence between these areas. The magnitude values of damage ratio and damage factors calculated are similar to the values obtained by different authors for similar types of materials.
However the differences found may also be due to differences that will exist between thermally affected area and damaged area, simultaneously to the images acquisition moments. This comparative study is to be considered soon on the ongoing project.
It is also possible to conclude that the applicability of thermography techniques in the characterization of the affected areas and on potentially induced damage by drilling operations is viable, although the first setup used was not the most suitable, since the thermographic camera was not aligned with the axis of drilling. This is an aspect that is also now corrected.
The presence of the focus indicator in some of the figures here presented is also an aspect that influences the image analysis. To overcome these types of undesirable external influences which may affect the results, a hierarchic segmentation process is now being carried out, in order to reach a more accurate identification of the affected areas.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the financial support of Project IPL, IDI&CA/COMPDRILL, and Project LAETA—UID/EMS/50022/2013.

REFERENCES

[1] D. J. S. Goncalves, “Avaliação de Ferramentas na Furação de Laminados Compósitos,” vol. 23, no. figura 1, pp. 109–113, 2011.
[2] I. Singh and N. Bhatnagar, “Drilling-induced damage in uni-directional glass fiber

reinforced plastic (UD-GFRP) composite laminates," *Int. J. Adv. Manuf. Technol.*, vol. 27, no. 9–10, pp. 877–882, 2006.
[3] H. Hocheng and C. C. Tsao, "Comprehensive analysis of delamination in drilling of composite materials with various drill bits," *J. Mater. Process. Technol.*, vol. 140, no. 1–3 SPEC., pp. 335–339, 2003.
[4] S. O. Ismail, S. O. Ojo, and H. N. Dhakal, "Thermo-mechanical modelling of FRP cross-ply composite laminates drilling: Delamination damage analysis," *Compos. Part B Eng.*, vol. 108, pp. 45–52, 2017.
[5] Z. Qi, K. Zhang, Y. Li, S. Liu, and H. Cheng, "Critical thrust force predicting modeling for delamination-free drilling of metal-FRP stacks," *Compos. Struct.*, vol. 107, pp. 604–609, 2014.
[6] A. A. Vieira, "COMPÓSITOS REFORÇADOS COM FIBRAS DE VIDRO E DE SISAL," 2011.
[7] O. Nicolau and G. Andrade, "Estudo de Delaminação em Compósitos de Matriz Polimérica," 2013.
[8] M. R. S. Tavares, P. Tecnol, E. Mec, R. Frias, and M. Comp, "Avaliação da delaminação após furação em compósitos laminados," 2010.
[9] A. Peer-reviewed and K. L. Makwana, "Engineering Science and Futuristic Technology Review : Potential applications of Liquid Desiccant technology in dehumidification and cooling the process air . A B S T R A C T : Keywords :," vol. 1, no. 1, pp. 1–8, 2015.
[10] E. Capello, "Workpiece damping and its effect on delamination damage in drilling thin composite laminates," *J. Mater. Process. Technol.*, vol. 148, no. 2, pp. 186–195, 2004.
[11] B. Işık and E. Ekici, "Experimental investigations of damage analysis in drilling of woven glass fiber-reinforced plastic composites," *Int. J. Adv. Manuf. Technol.*, vol. 49, no. 9–12, pp. 861–869, 2010.
[12] L. M. P. Durão, J. M. R. S. Tavares, A. T. Marques, M. Freitas, and G. Magalhães, "COMPÓSITAS," pp. 1–11, 2004.
[13] B. Ramesh, A. Elayaperumal, S. Satishkumar, and A. Kumar, "ScienceDirect Influence of cooling on the performance of the drilling process of glass fibre reinforced epoxy composites," *Arch. Civ. Mech. Eng.*, vol. 16, no. 1, pp. 135–146, 2015.
[14] Sandvik Coromant, "User's Guide: Machining carbon fibre materials," pp. 1–62, 2010.
[15] D. Gay, S. V Hoa, and S. W. Tsai, *COMPOSITE MATERIALS COMPOSITE MATERIALS*. 2003.
[16] J. Reddy, *Mechanics of Laminated Composite Plates and Shells: Theory and Analysis*, 2nd ed. 2004.
[17] R. Yang and Y. He, "Optically and non-optically excited thermography for composites: A review," *Infrared Phys. Technol.*, vol. 75, pp. 26–50, 2016.
[18] N. Bhatnagar, I. Singh, and D. Nayak, "Damage Investigation in Drilling of Glass Fiber Reinforced Plastic Composite Laminates," *Mater. Manuf. Process.*, vol. 19, no. 6, pp. 995–1007, 2004.
[19] "ImageJ." [Online]. Available: https://imagej.net/Welcome.

SYMCOMP 2017
Guimarães, 6-7 April 2017

MUSCL VS MOOD TECHNIQUES TO SOLVE THE SWE IN THE FRAMEWORK OF TSUNAMI EVENTS

Cláudia Reis[1]*, Jorge Figueiredo[1], Stéphane Clain[1], Rachid Omira[2], Maria Ana Baptista[3], Jorge Miranda[2]

1: Centre of Mathematics
School of Sciences
University of Minho
Campus de Gualtar, 4710 - 057 Braga
e-mail: claudiavdreis@sapo.pt, jmfiguei@math.uminho.pt, clain@math.uminho.pt
web: http://www.cmat.uminho.pt

2: Portuguese Institute for Sea and Atmosphere
Rua C do Aeroporto, 1749-077 Lisboa
e-mail: omirarachid10@yahoo.fr, miguel.miranda@ipma.pt

3: High Institute of Engineering of Lisbon
Lisbon Polytechnic Institute
R. Conselheiro Emídio Navarro 1, 1959-007 Lisboa
e-mail: mavbaptista@gmail.com

Keywords: Finite volume, MUSCL, MOOD, Japan 2011.

Abstract *The risk mitigation associated with tsunami events needs robust and accurate numerical tools to provide realistic solutions. We propose a comparative study between the efficiency of a finite volume numerical code, with second-order discretization in space and time, equipped with two different techniques to solve the non-conservative shallow-water equations: 1) the MUSCL (Monotonic Upstream-Centered Scheme for Conservation Laws) and, 2) the MOOD (Multi-dimensional Optimal Order Detection). A benchmarking process is carried out to validate the code. A one-dimensional simulation is performed to compare the MUSCL method equipped with the van Leer limiter, and the MOOD technique. We show that: 1) the quality of the solution may genuinely interfere with the scenario one wants to assess and, 2) the numerical tool equipped with the MOOD technique provides better solutions in comparison with the MUSCL results. At last, we apply and compare the two techniques to the real-case scenario Tohoku-Oki (Japan), 2011 tsunami.*

1. INTRODUCTION

Numerical tools, using non-linear hydrostatic shallow-water (NLHSW) models, are widely used to perform earthquake generated tsunami simulations in order to predict the tsunami behavior and build hazardous tsunami scenarios.

In this study, we present a finite volume numerical code, solving the NLHSW equations with second-order discretization in time and space and compare its performance considering two different techniques: Monotonic Upstream-Centered Scheme for Conservation Laws (MUSCL) [1] and Multi-dimensional Optimal Order Detection (MOOD) [2].

The MUSCL technique is based on *a priori* evaluation. The local linear reconstruction of the state variables is corrected using limiters to ensure that the Total Variation Diminishing (TVD) properties hold. The MOOD technique is based on *a posteriori* evaluation. A candidate solution is computed without any limitation using local polynomial reconstruction. If the solution fails to meet some stability criteria, the polynomial degree is decremented before recomputing the solution.

The performance of the numerical code was validated using the recommendations from the National Oceanic and Atmosphere Administration [3] and the studies of Synolakis et al. [4] and Tinti et al. [5]. The benchmarking process is fully described in Clain et al. [6]. In this paper, we show the code performance against a mathematical and a geophysical benchmark. The mathematical benchmark consists of a one-dimensional dam-break involving an irregular bathymetry with no dry zones, to test and validate the code capacities to deal with complex topographies. Numerical simulations are carried out with MUSCL and MOOD techniques, using different mesh sizes (*i.e.* the number of cells), to test the convergence of the solution.

The geophysical benchmark involves the comparison between the one-dimensional analytical formula proposed by Carrier and Greenspan [7] with the results (amplitude and velocity) obtained for the numerical simulation of a wave propagating on a uniformly sloping beach, the corresponding L^1- and L^∞-errors being evaluated for both techniques.

The satisfactory results on the benchmarking process encourage us to move towards a real-case, the Tohoku-Oki tsunami that struck the coast of Japan in 2011. We performed a one-dimensional simulation, along two different profiles extending from the source area to the target coast, aiming at reproducing the tsunami behavior, using the different second-order methods (MUSCL and MOOD). The quality of the numerical solutions along the two profiles is analyzed and compared.

We show: 1) the advantage of the MOOD technique with respect to MUSCL technique; and 2) that the quality of the numerical solution may genuinely interfere with the scenario one wants to assess.

2. MODELING AND NUMERICAL SCHEMES

The two-dimensional shallow-water system with varying bathymetry is given in Cartesian coordinates by equations (1.1) - (1.3):

$$\partial_t h + \partial_x(hu) + \partial_y(hv) = 0, \tag{1.1}$$

$$\partial_t(hu) + \partial_x\left(hu^2 + \frac{g}{2}h^2\right) + \partial_y(huv) = -gh\partial_x b, \tag{1.2}$$

$$\partial_t(hv) + \partial_x(huv) + \partial_y\left(hv^2 + \frac{g}{2}h^2\right) = -gh\partial_y b, \tag{1.3}$$

where h, u, v, b represent the water height, the velocity components following respectively the x and y axis, and the bathymetry, while $g = 9.81ms^{-2}$ is the gravitational acceleration constant. In the following, $\eta = b + h$ will stand for the total water height (or free surface). A one-dimensional version consists in considering just equations (1.1) and (1.2) with $v = 0$. Numerical schemes for solving the non-conservative shallow-water system have to respect some fundamental principles: mass conservation and preservation of some critical steady-states, such as the lake at rest, at the discrete level (well-balanced scheme). The finite volume method turns out to be a very efficient technique to numerically compute an approximation and we refer to Clain et al. [6] for a detailed description. Here, we just outline the method for the sake of consistency.

2.1. Numerical method

Since the one- or two-dimensional problem involves uniform grids, we only give a description of the method for axis x and denote by C_i a cell with center x_i and length Δx, while $x_{i-\frac{1}{2}} = x_i - \Delta x/2$, $x_{i+\frac{1}{2}} = x_i + \Delta x/2$ are the cell interfaces. For a generic function φ, we denote by $\varphi_i \approx \varphi(x_i)$ the numerical approximation of φ at x_i, while $\varphi_{i+\frac{1}{2},L}$ and $\varphi_{i+\frac{1}{2},R}$ stand for approximations on the left and right side of interface $x_{i+\frac{1}{2}}$, respectively. When the two values are equal, we use the notation $\varphi_{i+\frac{1}{2}}$ since no distinction between the left and right side of the interface is needed.

Assuming that all the state variables are known at time t^n, we compute their new values at x_i and time t^{n+1} using the following finite volume scheme:

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2}}^n + \varepsilon_{i+\frac{1}{2},L}^n - F_{i-\frac{1}{2}}^n - \varepsilon_{i-\frac{1}{2},R}^n\right) + \Delta t S_i^n, \tag{2}$$

where $F_{i+\frac{1}{2}}^n$ is the conservative flux across the interface $x_{i+\frac{1}{2}}$ while $U_i^n = (h_i^n, h_i^n u_i^n)$ is the vector of conservative variables. The source term S_i^n represents the discretization of the regular variation of the bathymetry, whereas $\varepsilon_{i+\frac{1}{2},L}^n$ and $\varepsilon_{i-\frac{1}{2},R}^n$ are the non-conservative flux contributions associated to the bathymetry discontinuity across the interface. Expressions for the source term and flux evaluations are given in Clain et al. [6] depending on the states computed on the left and right of the interface. Namely, for give values $h_{i+\frac{1}{2},L}, h_{i+\frac{1}{2},R}, u_{i+\frac{1}{2},L}, u_{i+\frac{1}{2},R}, b_{i+\frac{1}{2},L}, b_{i+\frac{1}{2},R}$ at time t^n, we compute the reconstructed hydrostatic states for interface $x_{i+\frac{1}{2}}$ in the following way. We denote by $b_{i+\frac{1}{2}}^* = \max(b_{i+\frac{1}{2},L}, b_{i+\frac{1}{2},R})$ the hydrostatic reconstruction bathymetry and set $h_{i+\frac{1}{2},L}^* =$

$\max(0, h_{i+\frac{1}{2},L} + b_{i+\frac{1}{2},L} - b^*_{i+\frac{1}{2}})$. We do the same for $h^*_{i+\frac{1}{2},R}$ which represents the effective water height in contact at the interface. We deduce $\eta^*_{i+\frac{1}{2},L} = h^*_{i+\frac{1}{2},L} + b^*_{i+\frac{1}{2}}$ (similarly for $\eta^*_{i+\frac{1}{2},R}$) the hydrostatic reconstruction free surface, while we denote $u^*_{i+\frac{1}{2},L} = u_{i+\frac{1}{2},L}$, $u^*_{i+\frac{1}{2},R} = u_{i+\frac{1}{2},R}$ for the sake of simplicity. We proceed in the same way for interface $x_{\mathrm{i}-\frac{1}{2}}$. We then compute the fluxes $F^n_{i+\frac{1}{2}}$ and $F^n_{i-\frac{1}{2}}$ as well as the source term S^n_i with the reconstructed variables, while the non-conservative fluxes $\varepsilon^n_{i+\frac{1}{2},L}$ and $\varepsilon^n_{i-\frac{1}{2},R}$ use the differences between the original variables and the hydrostatic reconstructed ones.

2.2. MUSCL vs MOOD

At this stage, we have to define the state of the left and right side of the interface, $\varphi_{i+\frac{1}{2},L}$ and $\varphi_{i-\frac{1}{2},R}$, from the states φ_i in the cells (we skip the time index for the sake of simplicity). The first-order scheme simply consists of setting $\varphi_{i-\frac{1}{2},R} = \varphi_{i+\frac{1}{2},L} = \varphi_i$ but the numerical method would suffer of a high amount of numerical diffusion. To improve the accuracy, the MUSCL method is usually employed due to its simplicity. Unfortunately, the resulting scheme may still suffer of numerical diffusion and instabilities may appear since the original method has been designed for conservative problems and not for the non-conservative equations. We propose an alternative based on the MOOD technique [8][9] which is more stable and accurate. We give hereafter a short overview of the two methods.

The MUSCL technique consists in a piecewise linear reconstruction controlled by a limiting procedure to ensure some stability properties such as positivity preserving and TVD [1]. To this end, we set:

$$\varphi_{i-\frac{1}{2},R} = \varphi_i - \frac{\Delta x}{2}\varphi'_i, \qquad \varphi_{i+\frac{1}{2},L} = \varphi_i + \frac{\Delta x}{2}\varphi'_i, \tag{3}$$

with φ'_i an approximation of the derivative of the function corrected by a limiting procedure to avoid the creation of spurious oscillations leading to non-physical approximations. Limiters such as minmod [10], van Albada [11] and van Leer [12] are often used. In the present work we use the van Leer limiter.

The MOOD technique is based on a complete different paradigm [8][9]. The principle is to compute a candidate solution for time t^{n+1} using the linear reconstruction without any limitation. Then, we check if the solution is admissible, in a sense we shall detail further on, and if there are problematic cells they are "cured" to provide an admissible approximation. Thus, the MUSCL technique is an *a priori* method since we perform the reconstruction limitation with the value at time t^n whereas the MOOD technique is an *a posteriori* method since the correction is carried out in function of the candidate solution at time t^{n+1}. The main advantage is that the MOOD method is less restrictive since we only perform the limitation for the cells which present some oscillations leading to a more accurate approximation. We now introduce the basic concepts to define the MOOD method.

Cell Polynomial Degree (CPD). For each cell, we associate a value, named the CPD, setting: 1 for linear reconstruction (the default) and 0 for constant value. We then determine the values on the left and right sides of the interfaces with:

- If CPD[i]=0, we set $\varphi_{i-\frac{1}{2},R} = \varphi_{i+\frac{1}{2},L} = \varphi_i$. (4.1)

- If CPD[i]=1, we set $\varphi_{i-\frac{1}{2},R} = \varphi_i - \frac{\Delta x}{2}\varphi_i', \qquad \varphi_{i+\frac{1}{2},L} = \varphi_i + \frac{\Delta x}{2}\varphi_i',$ (4.2)

  where we take: $\varphi_i' = \frac{\varphi_{i+1}-\varphi_{i-1}}{2\Delta x}$. (4.3)

The CPD map provides the information on the reconstructions we use to compute the candidate solution and we seek the best CPD map (higher degree) such that we have both accuracy and stability. To obtain a CPD map evaluation leading to an admissible solution at time t^{n+1} we proceed in the following way. For a given approximation (U_i^n) at time t^n and a prescribed CPD map, we compute a candidate solution $\left(U_i^{\#}\right)$ using the polynomial reconstruction given by the CPD map. Then, an iterative procedure involving a certain number of detectors (see below) is set to provide an admissible solution.

Detector. A detector is a small routine that checks a specific property of the candidate solution. We list hereafter three simple detectors we employ in the code:

Physical Admissible Detector (PAD). We say that approximation $U_i^{\#}$ on cell C_i is physically admissible if $h_i^{\#} \geq 0$. If the cell does not satisfy such a criterion, we set CPD[i] =0.

Extremum Detector (ED). We say that approximation $U_i^{\#}$ on cell C_i presents a local extremum if $h_i^{\#} > max(h_{i-1}^{\#}, h_{i+1}^{\#})$ or $h_i^{\#} < min(h_{i-1}^{\#}, h_{i+1}^{\#})$. If that is the case, we label cell C_i as problematic.

Smoothness detector (u2). When the candidate solution is considered problematic (*i.e.* presents an extremum on cell C_i), the local curvature χ_i is evaluated numerically as well as on the neighbor cells C_j.

- If the curvatures have opposite signs then the cell is not eligible and we set CPD[i]=0.
- If the absolute value of ratio $\frac{\min(|\chi_{i-1}|,|\chi_i|,|\chi_{i+1}|)}{\max(|\chi_{i-1}|,|\chi_i|,|\chi_{i+1}|)}$

  is too small with respect to a given threshold, the cell is not eligible and we set CPD[i]=0.
- Otherwise, we say that the candidate solution satisfies the u2 criterion and the solution for that cell is considered admissible (CPD[i]=1).

If the CPD map has been altered, we compute again the candidate solution for the cells that have been corrected and for their neighbor cells only. Otherwise, the candidate solution turns out to be the solution at time t^{n+1}. It has been shown that this procedure always provides an admissible solution [13] free from non-physical oscillations.

3. TEST CASES

In our code validation process, mass conservation, convergence, stability and the well-balanced property (or C-property) were considered and positively assessed [6]. Several benchmarks were carried out to perform the verification and validation of the numerical scheme, through comparisons of the code predictions with analytical solutions, laboratory experiments and field measurements [6].
This section is devoted to the synthetic benchmarks in order to assess the advantages of the MOOD technique with respect to the MUSCL one. The first test corresponds to a classical dam-break involving shocks and rarefactions (see [14] and references therein) to draw comparisons when discontinuities are involved. The second test case corresponds to an analytical benchmark [3][4] based on a one-dimensional smooth solution of an initial value problem for a single wave nonlinear propagation over a constant slope beach [7]. It belongs to the set of test cases proposed in the technical memorandum from the National Oceanic and Atmosphere Administration [3], based on the Long Wave Run-Up Models Workshops [15][16][17].

3.1. Dam-break

The first benchmark consists of a one-dimensional dam-break involving an irregular bathymetry with no dry zones. The simulation domain is $[0\text{m}, 1\text{m}]$ and the initial configuration is such that the water level is 1m for $x \in [0m, 0.5\text{m}]$ and 0.05m otherwise. The discontinuous bathymetry corresponds to a ramp located at $x \in [0.65m, 0.75\text{m}]$ with a slope coefficient of 0.4 as presented in Figure 1.

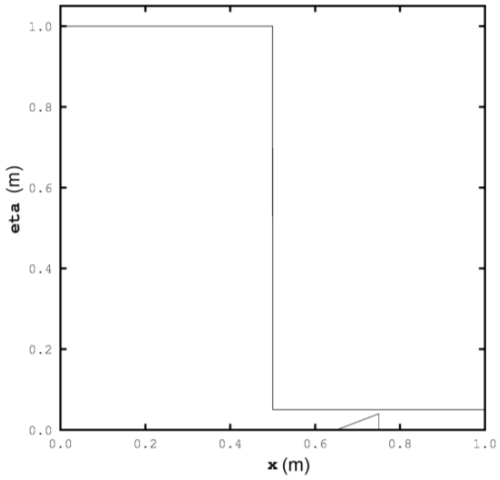


Figure 1. Dam-break: representation of bathymetry and initial configuration for the water free surface.

The system is initially at rest and the simulation is carried until a final simulation time $t_f = 0.125\text{s}$ is reached. Reflection boundary conditions are prescribed for both domain sides. Since no analytical solution is available for comparison, we consider a reference

solution obtained with the first-order scheme and a uniform mesh of 100 000 cells. The numerical simulations are carried out with the MUSCL and the MOOD methods for two uniform meshes of 100 and 200 cells. The conservative numerical flux is computed with the HLL law while the non-conservative flux derives from the hydrostatic reconstruction formalism presented in the previous section.
The water free surface and the velocity field obtained using the MUSCL and the MOOD techniques at the final time $t = \mathrm{t}_f$ are plotted in Figure 2 for the 100 and the 200 cells case together with the corresponding reference numerical solution.

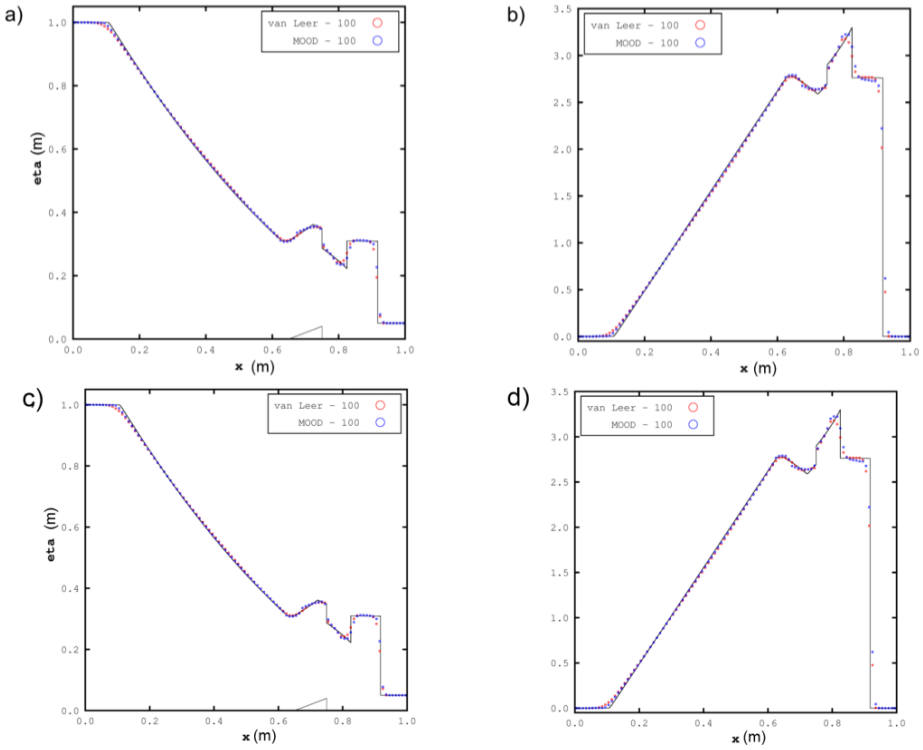


Figure 2. Dam-break: comparison between the reference numerical solution (solid line) and the numerical solutions obtained with the MUSCL method (red circles) and the MOOD method (blue circles). a) Water free surface at $t = \mathrm{t}_f$ for the 100 mesh. b) Water velocity at $t = \mathrm{t}_f$ for the 100 mesh. c) Water free surface at $t = \mathrm{t}_f$ for the 200 mesh. d) Water velocity at $t = \mathrm{t}_f$ for the 200 mesh.

We report that the MOOD method provides the best approximation since it better fits the reference curve. The two rarefactions feet situated at $\mathrm{x} = 0.15m$ and $\mathrm{x} = 0.65m$, corresponding to the beginning and the end of the rarefaction, are very sensitive to the numerical diffusion since they correspond to singular points involving a discontinuity of the derivative. The MUSCL curve presents a rather smeared shape whereas the MOOD curve really catches well the singular points. Such a good behavior of the MOOD technique is noticeable since the evaluation of the solution at these specific points represents a real

difficulty from the numerical point of view. On the other hand, an important challenge is to capture the shocks with very few points. The genuinely nonlinear shocks right after $x = 0.8m$ and $x = 0.9m$ are very well recovered by the two techniques but MOOD turns out to be better since the curve manages to reach the peak of the reference solution. At last, convergence in mesh has been checked and the 200 cells mesh provides a satisfactory improvement with respect to the 100 cells one, particularly for the rarefaction. Additionally, numerical experiences with finer grids (not presented here) have been carried out allowing to validate the convergence in mesh property.

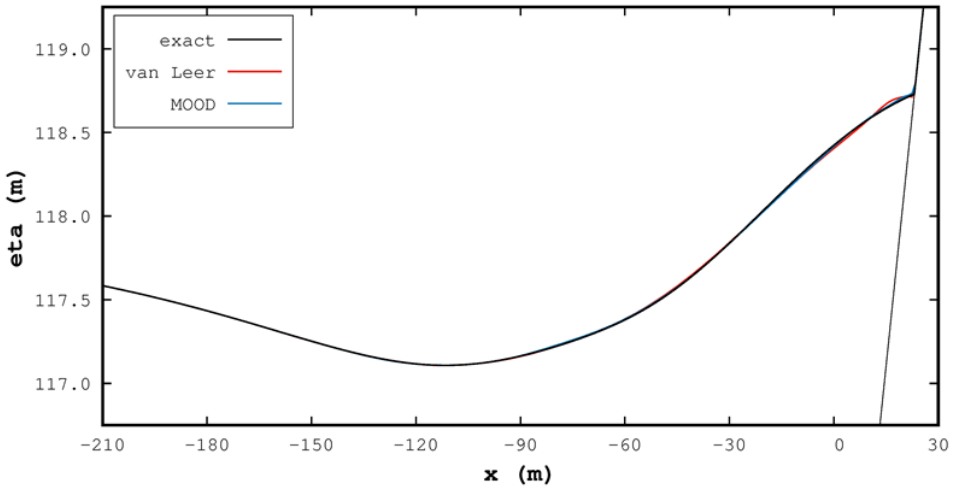
3.2. Single wave on a sloping beach

We perform a one-dimensional analytical benchmark where dry and wet zones are involved. A wave of given initial profile and initial null velocity propagates over a constant-slope beach. The leading-depression N wave travels across a 1/10 slope and we compare the numerical solution with the one obtained by the analytical integral formula given by Carrier and Greenspan [7] and Carrier et al. [18]. We consider the domain $[-570\text{m}, 30\text{m}]$, where the subdomain $[0\text{m}, 30\text{m}]$ corresponds to the region that is initially dry (i.e. $x = 0\text{m}$ corresponds to the shoreline position at $t = 0\text{s}$). The domain is discretized using a uniform mesh having 600 cells.

Figure 3 depicts, for $t = 13\text{s}$, the free surface using MUSCL and MOOD methods (global view and a zoom close to the dry/wet interface to assess the accuracy of the two second-order simulations).

We report that the solution obtained with the MOOD technique is more accurate in comparison with the solution obtained with the MUSCL technique.

a)



b)

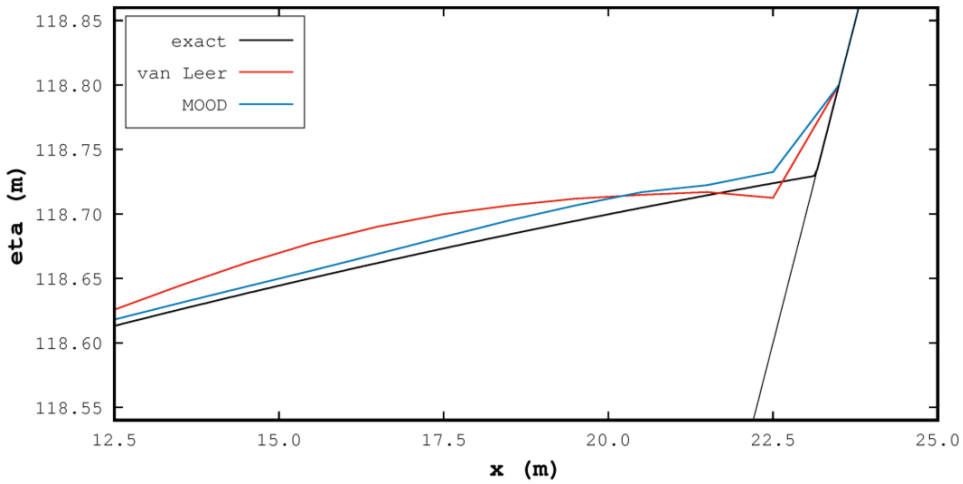


Figure 3. Sloping beach: comparison between the exact solution and the numerical simulations for the water free surface η at t = 13s. The exact solution is presented together with the numerical ones: a) Global view; b) Zoom close to the dry/wet interface.

Table 1 quantifies the accuracy by evaluating the maximum error ε_η^∞ and the L^1-error ε_η^1 on the free surface elevation (see [6], for the definition of the metrics). Evaluation of the free surface obtained with the MOOD technique provides a clear improvement over the MUSCL solution.

Scheme	Free surface		Velocity	
	ε_η^∞	ε_η^1	ε_u^∞	ε_u^1
MUSCL	2.82×10^{-2}	1.62×10^{-3}	3.04	7.62×10^{-3}
MOOD	1.27×10^{-2}	1.05×10^{-3}	1.96×10^{-1}	1.73×10^{-3}

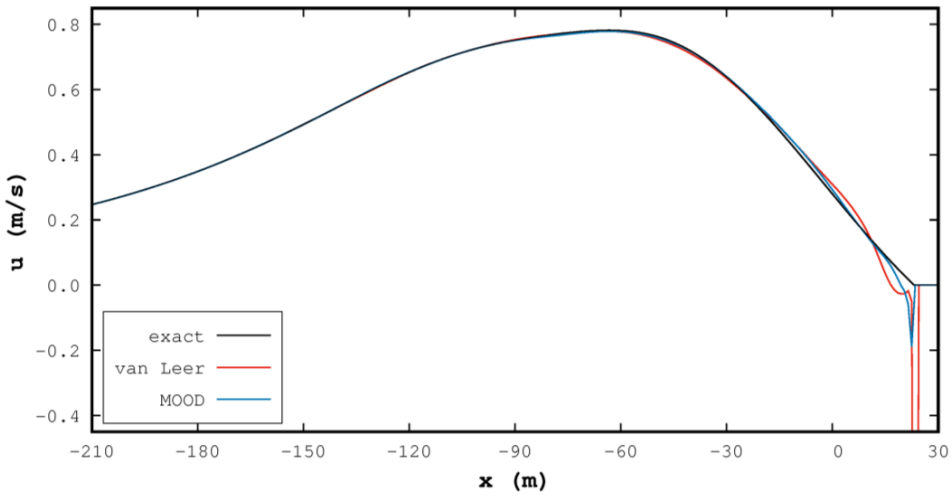
Table 1. Sloping beach: Errors for the free surface and velocity for t = 13s using two metrics.

Figure 4 depicts the exact analytical solution for the velocity and the simulated approximations at t = 13s.

The MOOD technique results are more accurate than the results obtained with the MUSCL technique. Table 1 provides a numerical quantification of the errors on the velocity for the two schemes. Notice that the approximation obtained with the van Leer scheme is particularly problematic. The reason is that we are dealing with a non-conservative problem and the computation of the velocity is very sensitive.

Evaluation of the velocity is one of the most difficult parts of the numerical scheme in presence of small water height. The L^1 metric shows that the MOOD method is better, whereas the MUSCL technique suffers of a large deviation.

a)



b)

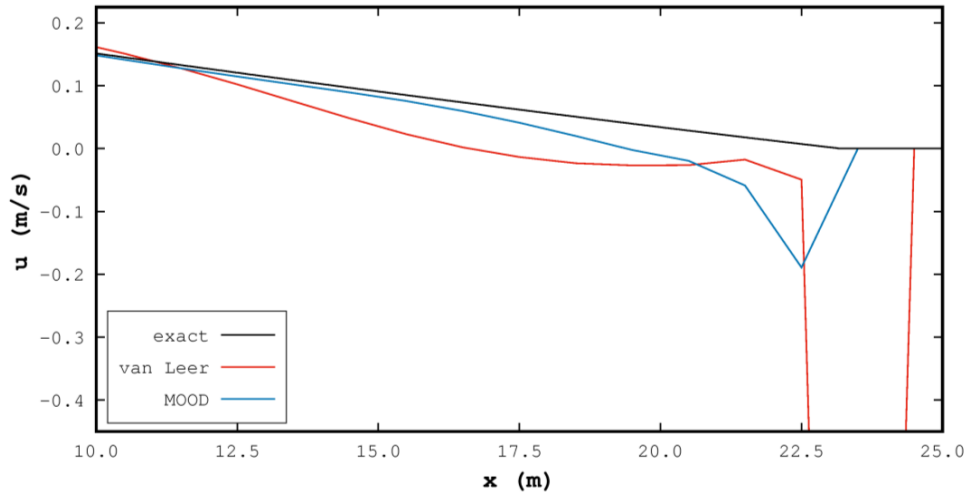


Figure 4. Sloping beach: comparison between the exact solution and the numerical approximations for the velocity u at t = 13s. The exact solution is presented together with the numerical ones: a) Global view; b) Zoom close to the dry/wet interface.

Both test cases (3.1 and 3.2) show that the MOOD technique provides a better approximation to the solution than the MUSCL approach, with sharper shock capture and less numerical diffusion.

4. REAL-CASE STUDY

At 14:46:18.1 local time (5:46:18.1 UTC), on March 11, 2011, a magnitude Mw9.0 tsunamigenic earthquake occurred in Japan: The earthquake epicenter (38.1035°N,

142.861°E) and depth (24 km) were estimated by the Japan Meteorological Agency (JMA) (see Figure 5). Multiple sea-level sensors recorded the event, namely the Deep-ocean Assessment and Reporting of Tsunamis (DART) and the Global Positioning (GPS) systems (Figure 5).

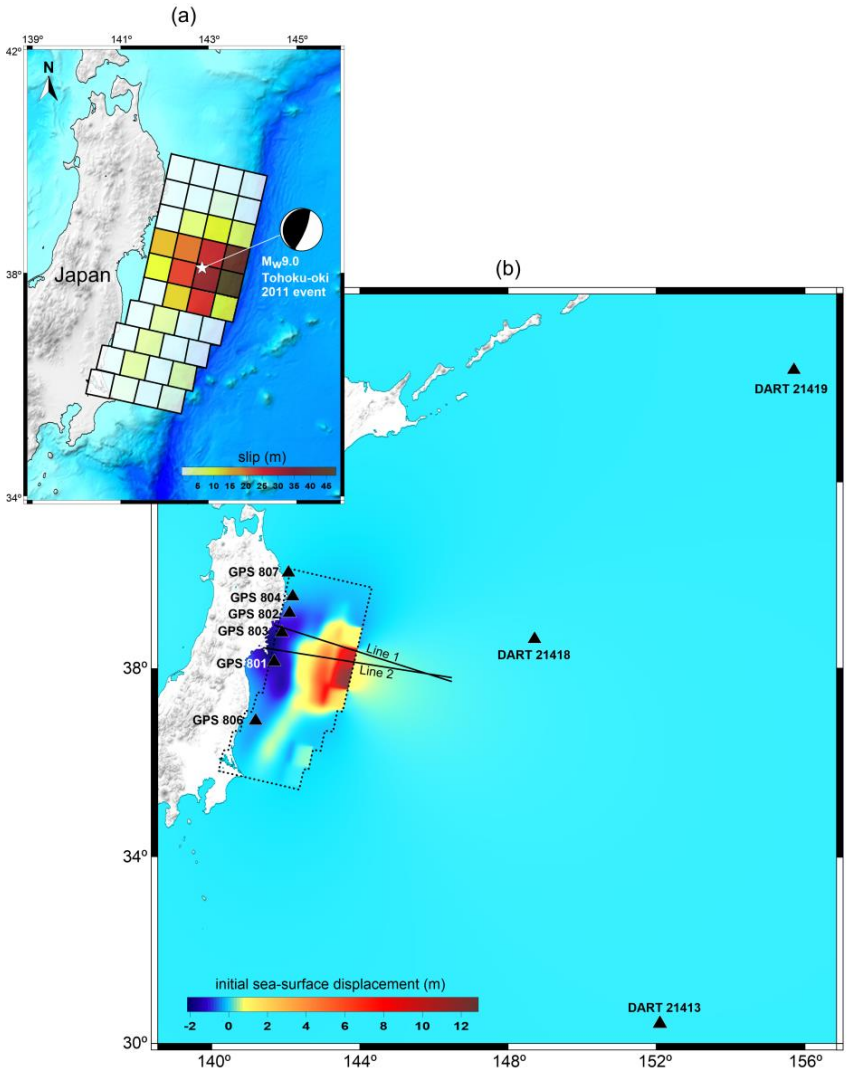


Figure 5. Tohoku-Oki, 2011, earthquake and tsunami model: a) Co-seismic slip distribution proposed by Fujii et al. [19], location of the epicenter (white star) and focal mechanism (beach ball) determined by JMA; b) The modeled sea-surface displacement caused by the Tohoku-Oki, 2011 earthquake. Black triangles represent the locations of the sea-level sensors. Lines 1 and 2 represent the profiles location used in the 1D simulation.

4.1. Tsunami generation

To simulate the tsunami wave creation caused by the 2011 Tohoku-Oki earthquake, we use

the source model proposed by Fujii et al. [19]. This finite-fault model considers heterogeneous slip distribution and constitute the base to define the tsunami initial conditions. Figure 5a depicts the heterogeneous slip distribution along the rupture area of the 2011 Tohoku-Oki earthquake fault as determined by Fujii et al. We calculate the co-seismic vertical displacements through the half space elastic theory [20] and transfer the vertical deformations corresponding to the sea-floor to the free ocean surface with the assumption that both deformations of sea bottom and ocean surface are equal [21]. The bathymetry model used for the tsunami simulation is from the General Bathymetric Chart of the Oceans (GEBCO) 30-arc sec gridded data (available at: http://www.gebco.net/). Figure 5b depicts the initial sea-surface displacement caused by the 2011 Tohoku-Oki earthquake.

4.2. One-dimensional simulation for code techniques comparison

The goal of this section is twofold: 1) show that a one-dimensional numerical model is appropriate to describe the wave propagation is some situations; and 2) show that the MOOD method is more efficient than the MUSCL method to provide a stable and accurate numerical approximation of the wave propagation.

4.2.1. Model reduction to one-dimensional problem

We extract two profile lines (labelled 1 and 2) from the two-dimensional map (see Figure 5b) joining the tsunami origin with the coastline such that the lines are essentially orthogonal to the direction of the waves' propagation. Such a model reduction is effective assuming that the waves preserve their unidimensional shape. Therefore, to ensure physical coherence between the two- and the one-dimensional models, the available potential energy of the one-dimensional initial water at rest configuration is changed by reducing the water height so that both two-dimensional and one-dimensional numerical models lead to the same maximum water elevation at approximately 10km from the initial configuration of the shoreline ($x = 0\text{m}$). The one-dimensional mesh of each line is uniform and inherited from the two-dimensional mesh, the cell length being approximately 364m for line 1 and 353m for line 2. Numerical simulations are carried out up to a simulation final time $t_f = 50\text{min}$.

To assess the quality of the correspondence between the two- and the one-dimensional numerical models when the MUSCL scheme is used we consider 10 reference points along each line, spanning approximately from $x = 90\text{km}$ to $x = 3\text{km}$. For each reference point, we look at the absolute difference between the results given by the two- and the one-dimensional simulations for the maximum free surface level and the corresponding time. The results obtained for line 1 (line 2 resp.) lead to deviations for the maximum free surface level with average 7.5% (2.5% resp.) and maximum 12.7% (3.4% resp.), while for the corresponding time we obtain differences with average 0.8% (4.7% resp.) and maximum 1.6% (11.4% resp.), being all deviations smaller close to the shoreline.

The one- and the two-dimensional models fit very well and confirm that the unidirectional model is relevant to analyze the run-up.

4.2.2. Comparison between the two high-order strategies

We have carried out several benchmarks to assess the quality of the numerical solutions considering two types of conditions for the wave-coast interaction (see Table 2):

- The run-up assumption, where the wave meets a sloping beach and spreads until it reaches a zero velocity;
- The cliff condition, where the wave meets a high cliff preventing the tsunami from entering in the dry area.

For the run-up and cliff conditions, we consider, for different points along lines 1 and 2 near the shoreline, the maximum water elevation η_{max} and the associated time t_{max} when the MUSCL and the MOOD numerical schemes are used. The maximum shoreline water elevation and the corresponding time, as well as the maximum inshore distance reached by the wave were also considered for the run-up case.
The results obtained show that the time corresponding to the maximum water elevation does not change significantly when different numerical schemes are considered for both line 1 and line 2.
We report that the MOOD scheme systematically leads to greater values of the maximum water elevation both in the wet zone and on the shoreline. In the run-up simulation, this difference amounts up to almost 1 and 2 meters, respectively at the shoreline and at $x = 0$m. This difference can reach almost 3 meters in the cliff condition scenario. From a practical point of view, a difference of 3 meters may result in rather catastrophic consequences. Indeed, a protection infrastructure such as a seawall designed on the basis of the MUSCL simulations will be underestimated while the MOOD method predicts that the water may overtop the infrastructure. For long-time simulation like the tsunami propagation, MUSCL technique is much diffusive and provides an underestimation of the wave height when reaching the coast. The MOOD method is less diffusive and gives more realistic estimates of the free-surface height.

5. CONCLUSIONS

We have developed a finite volume scheme and implemented the corresponding C++ code, with resolution up to second-order, hydrostatic reconstruction, well-balanced property and two high-order techniques: MUSCL and MOOD.
The solution quality achieved with the numerical tool is of utmost importance for the tsunami risk mitigation. In this study, the quality solutions were evaluated and compared in a benchmarking process and a real-case scenario, the Tohoku-Oki, 2011 event.
The analytical benchmarks used to validate the code show a better performance of the MOOD technique with respect to the MUSCL one.
The two-dimensional bathymetric and tsunami initial deformation models were used to calibrate the one-dimensional model for two profiles coinciding with two tide gauges. The unidimensional simulation was performed using the MUSCL and MOOD techniques to assess the accuracy. It brought out that the MUSCL method is too much diffusive for long-time simulation, providing underestimation of the water height whereas the MOOD method

manages to give a more realistic evaluation.
The results of the benchmarking process show that: 1) the first-order simulation presents the less accurate numerical solution, mostly due to diffusion; 2) the MUSCL limiters have influence on the numerical solution; and 3) MOOD provides the best numerical solution.

Run up condition									
Line 1					Line 2				
Point	MUSCL		MOOD		Point	MUSCL		MOOD	
x	η_{max}	t_{max}	η_{max}	t_{max}	x	η_{max}	t_{max}	η_{max}	t_{max}
-9.83	5.8	29.9	6.9	29.8	-10.66	9.8	33.9	10.9	33.8
-7.65	6.1	30.9	7.2	30.8	-7.42	10.7	35.2	11.8	35.2
-5.10	6.9	32.1	8.0	31.9	-4.95	12.1	36.9	13.2	36.7
-2.55	7.8	33.4	8.8	33.1	-2.47	13.5	37.7	14.5	37.9
0.0	11.1	35.4	12.9	35.0	0.0	19.0	38.4	20.8	39.5
shoreline	17.6	35.6	14.1	35.5	shoreline	18.4	40.8	19.2	40.4
x_{max}	728m		364m		x_{max}	353m		353m	
Cliff condition									
Line 1					Line 2				
Point	MUSCL		MOOD		Point	MUSCL		MOOD	
x	η_{max}	t_{max}	η_{max}	t_{max}	x	η_{max}	t_{max}	η_{max}	t_{max}
-9.83	5.8	29.9	6.9	29.8	-10.66	9.8	33.9	10.9	33.8
-7.65	6.1	30.9	7.2	30.8	-7.42	10.7	35.2	11.8	35.3
-5.10	7.2	32.1	8.2	31.9	-4.95	12.1	36.9	13.2	36.7
-2.55	7.7	33.3	8.9	33.1	-2.47	13.5	37.7	14.5	37.9
0.0	12.5	34.8	14.5	35.0	0.0	18.8	39.7	21.4	39.7

Table 2. 1D tsunami simulation: Maximum water elevation η_{max} (in meters) and corresponding time t_{max} (in minutes) for MUSCL and MOOD numerical schemes for lines 1 and 2. The different points are identified by the corresponding x-coordinate (in kilometers). For the run-up condition, the results for the shoreline are also presented.

REFERENCES

[1] LeVeque, R.J. "Finite volume methods for hyperbolic problems". *Cambridge university press*, vol. 31, 2002.

[2] Clain, S., Figueiredo, J. "The MOOD method for the non conservative shallow-water system". *Computers & Fluids*, 145, 99-128, 2017.

[3] Synolakis, C.E., Bernard, E.N., Titov, V.V., Kanoglu, U., Gonzalez, F.I. "Standards, criteria and procedures for NOAA evaluation of tsunami numerical models". *NOAA Technical Memorandum OAR PMEL-135*, 2007.

[4] Synolakis, C.E., Bernard, E.N., Titov, V.V., Kânoğlu, U., González, F.I. "Validation and verification of tsunami numerical models". *Pure and Applied Geophysics*, 165(11-12), 2197-2228, 2008.

[5] Tinti, S., Tonini, R. "The UBO-TSUFD tsunami inundation model: validation and application to a tsunami case study focused on the city of Catania, Italy". *Nat. Hazards Earth Syst. Sci.*, 13, 1795-1816, 2013.

[6] Clain, S., Reis, C., Costa, R., Figueiredo, J., Baptista, M. A., Miranda, J. M. "Second-order finite volume with hydrostatic reconstruction for tsunami simulation". *Journal of Advances in Modeling Earth Systems*, 8(4), 1691-1713, 2016.

[7] Carrier, G.F., Greenspan, H.P. "Water waves of finite amplitude on a sloping beach". *Journal of Fluid Mechanics*, 4(01), 97-109, 1958.

[8] Diot, S., Clain, S., Loubère, R. "Improved detection criteria for the Multi-dimensional Optimal Order Detection (MOOD) on unstructured meshes with very high-order polynomials". *Computers & Fluids*, 64, 43-63, 2012.

[9] Diot, S., Loubère, R., Clain, S. "The MOOD method in the three-dimensional case: Very-High-Order Finite Volume Method for Hyperbolic Systems". *International Journal for Numerical Methods in Fluids*, 73(4), 362-392, 2013.

[10] Roe, P.L. "Characteristic-based schemes for the Euler equations". *Ann. Rev. Fluid Mech.*, 18, p337, 1986.

[11] Van Albada, G.D., Van Leer, B., Roberts, W.W. "A comparative study of computational methods in cosmic gas dynamics". *Astron. Astrophysics*, 108, p76, 1982.

[12] Van Leer, B. "Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection". *Journal of computational physics*, 23(3), 276-299, 1977.

[13] Clain, S., Diot, S., Loubère, R. "A high-order polynomial finite volume method for hyperbolic system of conservation laws with Multi-dimensional Optimal Order Detection (MOOD)". *Journal of computational Physics*, 230, 4028-4050, 2011.

[14] Delestre, O., Lucas, C., Ksinant, P.A., Darboux, F., Laguerre, C., Vo, T.N.T., James, F., Cordier, S. "SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies". *Int. J. Numer. Meth. Fluids*, 74, 229-230, 2014.

[15] Liu, P.L.F., Synolakis, C.E., Yeh, H.H. "Report on the international workshop on long-wave run-up". *Journal of Fluid Mechanics*, 229, 675-688, 1991.

[16] Yeh, H., Liu, P., Synolakis, C.E. "Long-wave Runup Models: Friday Harbor, USA, 12-17 September 1995". *World Scientific*, 1996.

[17] Synolakis, C.E., Bernard, E.N. "Tsunami science before and beyond Boxing Day 2004". *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 364(1845), 2231-2265, 2006.
[18] Carrier, G.F., Wu, T.T., Yeh, H. "Tsunami run-up and draw-down on a plane beach". *Journal of Fluid Mechanics*, 475, 79-99, 2003.
[19] Fujii, Y., Satake, K., Sakai, S.I., Shinohara, M., Kanazawa, T. "Tsunami source of the 2011 off the Pacific coast of Tohoku Earthquake". *Earth, planets and space*, 63(7), 815-820, 2011.
[20] Okada, Y. "Surface deformation due to shear and tensile faults in a half-space". *Bulletin of the seismological society of America*, 75(4), 1135-1154, 1985.
[21] Kajiura, K. "Tsunami source, energy and the directivity of wave radiation". *Bull. Earthq. Res. Inst.*, 48, 835–869, 1970.

SYMCOMP 2017
Guimarães, 6-7 April 2017

JOINING OF SHEETS BY SHEET-BULK FORMING: A NUMERICAL AND EXPERIMENTAL STUDY

I.M.F. Bragança[1,2]*, M.A.R. Loja[1,2], C.M.A. Silva[2], L.M. Alves[2], P.A.F. Martins [2]

1: ADEM/GI-MOSM
ISEL, Instituto Superior de Engenharia de Lisboa,
Instituto Politécnico de Lisboa,
Rua Conselheiro Emídio Navarro, 1959-007 Lisboa, Portugal
e-mail: *ibraganca@dem.isel.pt, amelialoja@dem.isel.ipl.pt

2: IDMEC
IST, Instituto Superior Técnico,
Universidade de Lisboa,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mail: carlos.alves.silva@tecnico.ulisboa.pt, luisalves@ist.utl.pt,
pmartins@tecnico.ulisboa.pt

Keywords: Experimentation, Finite element analysis, Joining by plastic deformation, Sheet-bulk forming

Abstract The authors present a new mechanical joining process that allow to connect two sheets perpendicular to one another, a variant of the traditional 'mortise-and-tenon' joint.
This investigation is focused on joining similar and dissimilar sheets based on sheet-bulk forming and it is supported by experimental data and numerical simulation with an in-house finite element program, I-form.
Results show that is possible to join materials by sheet-bulk forming and is also presented the joinability window as a function of the main operating parameters. The divided flux technique could be a valid option to eliminate the observed small clearance related to the elastic recovery of polymers. Furthermore, this technique could also be applied whether a lower initial punch force is required or even a different head morphology is desired.
Destructive tensile tests were performed to determine the maximum force that the joints are capable to withstand without failure and to identify the failure mechanism.

1. INTRODUCTION

The growing demand for consumer goods has enhanced the development of new materials and products, leading to emerge innovative manufacturing techniques, such as new metal forming and joining processes. In fact, the need for processing new alloys and joining dissimilar materials is responsible for the establishment of new challenges such as bulk forming techniques and joining by plastic deformation processes.

Plastic deformation is generally used in forming processes to shape mechanical parts, however it can also be applied to join parts. Moreover, comparing with conventional welding techniques (fig. 1 a) and adhesive bonded joints that normally require surface preparation (fig. 1 b), these processes have the advantage of joining a wide range of materials (including dissimilar materials), showing less distortion, embrittlement, tensile residual stress, and higher process reliability among other advantages. Mechanical joint with fasteners or rivets is also widely used, but involves extra weight derived from the utilization of brackets, clips, stiffeners, washers, screws, nuts and rivets (fig. 1 c).

Under these circumstances, the authors present a new mechanical joining process, environmentally friendly, that allow to connect two sheets perpendicular to one another by sheet-bulk forming, a variant of the traditional 'mortise-and-tenon' joint (fig. 1 d). Sheet-bulk forming process allow to manufacture parts with local thinning thickening, or functional features such as teeth, ribs and may possibly be used to join parts [1,2].

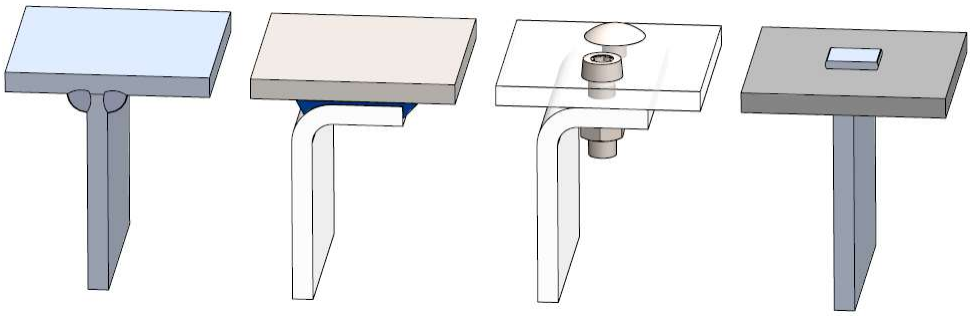


Figure 1. Types of joints to connect two sheets perpendicular to one another: (a) welded joint; (b) adhesive bonded joint; (c) mechanical fastened or riveted joint; (d) the proposed 'mortise-and-tenon' joint.

This investigation is focused on joining similar and dissimilar sheets based on sheet-bulk forming and it is supported by experimental data and numerical simulation with an in-house finite element program, I-form. The main purpose using this methodology that draws from material characterization to finite element modelling, is to increase process know-how, improve the technological practice during the process development and validate software use on this forming technology. Using the numerical simulation software, it was possible to predict the joint feasibility between the two sheets, the load-displacement evolution and the maximum force required from the hydraulic press.

The design of punches and the loading application procedure are also discussed. Special emphasis is placed on imposing divided plastic flow at the early stages of deformation in order to reduce the initial compression forces and reduce elastic recovery alloying to eliminate a possible clearance in the joint.

2 EXPERIMENTATION

2.1 Mechanical characterization: stress-strain curves

The stress-strain curve of the commercial aluminium alloy EN AW 5754 H111 sheets with 5mm thickness was performed by means of standard compression tests while polycarbonate was determined by means of compression and tensile tests since this polymer is pressure sensitive [3]. The cylinder test specimens were assembled by pilling up 3 circular discs machined out of the supplied aluminium and polycarbonate sheets by a hole-saw. To determine the tensile stress-strain curve, the test specimens were machined out of the supplied sheets in accordance to the ASTM D 638 standards. The compression tests were performed at room temperature on a hydraulic testing machine (Instron SATEC 1200 kN) and the resulting stress-strain curves are shown in Figure 2.

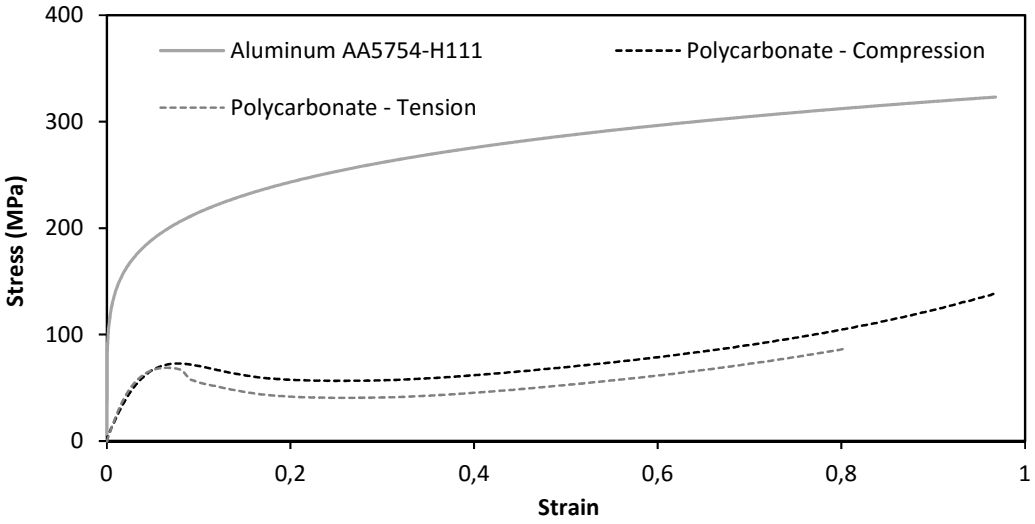


Figure 2. Aluminium AA5754-H111 and polycarbonate sheets stress-strain curves.

2.2 Work plan, Methods and Procedures

The laboratory work is performed combining finite element modelling and experimental investigation. To accomplish this, a laboratory tool was designed to test the specimens, identify the major process parameters and understand their influence on the overall joining feasibility. Special emphasis is placed on the utilization of V-shaped ('chisel') punches at the early stages of upsetting in order to reduce the initial compression forces, change head design and reduce the material elastic recovery.

The study was focused on the connection of polymer and metal sheets perpendicular to one another, joining different combinations.

The tests were performed in the tool developed, schematic represented in figure 3, highlighting the forming punches ('P1':flat and 'P2':V-shaped) and the blank holder that allow to clamp the lower sheets firmly. The tool installed in testing machine Instron SATEC1200 kN where carried out in displacement control at 5mm/min and room temperature.

The unit cell that fixes longitudinally in position two sheets perpendicular to one another, is characterized by a rectangular cavity, from now on called 'mortise', and a tenon longer than wider that passes completely through the other sheet. The mechanical locking is performed by a compression force perpendicular to the thickness direction that plastically deform the free length of the tenon along z-axis.

Firstly, the authors studied the upset compression of the polycarbonate and aluminium tenons, this allow to determine the process window without signs of plastic instability and risk of failure by buckling. Different tenons' length-to-width ratio were tested in both materials, setting the width twice the sheet thickness $t_0/w_0 = 0.5$ (figure 3).

The connection of two individual sheet specimens by means of the new proposed mechanical joining process, takes different parameters in consideration, the length-to-width ratio l_f/w_0, material and the geometry of the punch. The authors present two techniques, the first performed with just one-stage (without performing) and second with two-stages (with performing). The later involved the utilization of flat (P1) and V-shaped (P2) punches, figure 3, in order to evaluate the influence of promoting and facilitating the occurrence of divided flow in the plastically deformed tail of the tenon.

Destructive testing is carried out to characterize the overall performance of the new proposed 'mortise-and-tenon' joint and determining the maximum force to detach the two sheets.

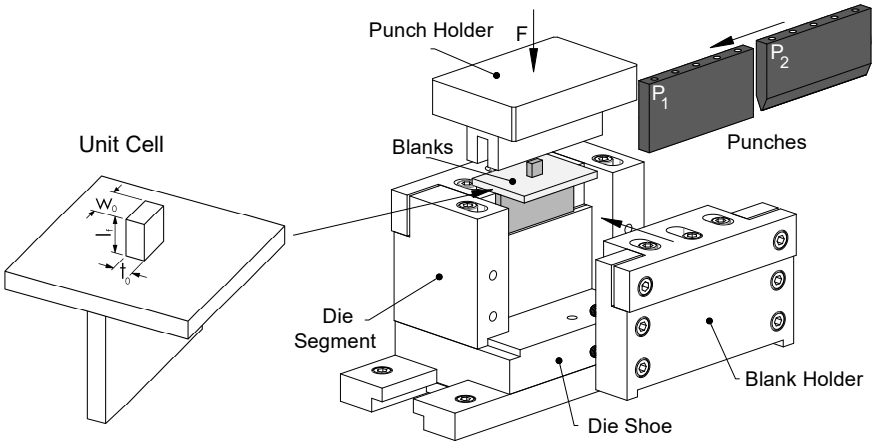


Figure 3. Laboratory tool schematic representation for joining a unit cell of two sheets perpendicular to one another.

3. NUMERICAL MODELLING

3.1 Extended finite element flow formulation to pressure sensitive polymers

The investigation on the joining of metal, polymer and polymer-metal sheets perpendicular to one another was supported by numerical simulations with the in-house finite element computer program I-form. I-form is based on the finite element flow formulation and includes an extension to pressure-sensitive polymers that allows modelling metal and polymer deformable objects simultaneously [4].

The extension of the finite element flow formulation to pressure-sensitive polymers utilizes the yield function $F(\sigma_{ij})$ proposed by Caddell et al. [5] for modelling cold plastic deformation of polymers with strength-differential effects resulting from the difference between tensile σ_T and compressive σ_C flow stresses,

$$F(\sigma_{ij}) = \bar{\sigma}^2 - \sigma_C \cdot \sigma_T + (\sigma_C - \sigma_T)\,\sigma_{kk} = 0 \tag{1}$$

where $\bar{\sigma} = \sqrt{\frac{3}{2}\sigma'_{ij}\sigma'_{ij}}$ is the effective stress and $\sigma_{kk} = \delta_{ij}\,\sigma_{ij}$.

However, and in contrast to what is commonly done in metals, the extension to pressure-sensitive polymers makes use of a non-associated flow rule in which the plastic potential Q is not related to the yield function $F(\sigma_{ij})$ but to the second invariant of the deviatoric stress tensor $Q = J_2$,

$$\dot{\varepsilon}_{ij}^{p} = \dot{\lambda}\frac{\partial Q(\sigma_{ij})}{\partial \sigma_{ij}} \tag{2}$$

where $\dot{\lambda}$ is a scalar factor of proportionality.
This methodology has been successfully applied by other authors to ensure incompressibility in the cold forming of pressure-sensitive polymers [6,7] and to perform numerical simulations with finite element computer programs [8].
The variational statement that gives support to the finite element flow formulation and allows metals and polymers to be treated simultaneously is written as,

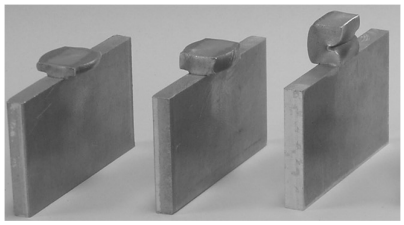
$$\Pi = \int_V \overline{\sigma}\,\dot{\bar{\varepsilon}}\,dV + \tfrac{1}{2}K\int_V \dot{\varepsilon}_v^2\,dV - \int_{S_T} T_i u_i\,dS + \int_{S_f}\left(\int_0^{|u_r|}\tau_f\,du_r\right)dS \tag{3}$$

where, $\dot{\bar{\varepsilon}}$ is the effective strain rate, $\dot{\varepsilon}_v$ is the volumetric strain rate, K is a large positive constant enforcing the incompressibility constraint of both metals and polymers and V is the control volume limited by the surfaces S_U and S_T, where velocity and traction are prescribed. Friction at the contact interfaces S_f is treated as a traction boundary condition and the additional power consumption term is modelled through the utilization of the law of constant friction $\tau_f = mk$ [9].
To investigate the tenon critical length length-to-width ratio l_f / w_0 to generate plastic instability and out-of-plane buckling it was employed finite element modelling, using three-dimensional and simplified two-dimensional models under plane strain deformation conditions and the sheets were modelled as deformable objects. When the simulations make use of three-dimensional models, the specimens were discretized by means of hexahedral elements and the blank holder and punch were discretized by means of contact-friction spatial linear triangular elements. As will be seen later, the simplified 2D models provide results near to the experimental observations and force evolutions and the complete simulations just take a few minutes. In figure 5 the images show the initial and final deformed meshes for two different test specimens which give rise to a symmetric plastic deformation and non-symmetrical upset deformation resulting from out-of-plane buckling.
Finite element analysis was also employed to simulate the joining of the individual sheet specimens of a ‘unit cell’ locked by sheet-bulk forming, made from aluminium, polycarbonate and combinations of both. The experienced gained in the upset compression of the tenons was special important, because allowed to better understand the process and, if considered viable, permitted to simplify the simulation modelling, building two-dimensional models, halving the two sheets lengthwise and discretizing the resulting cross section by means of quadrilateral elements under plane strain deformation conditions.

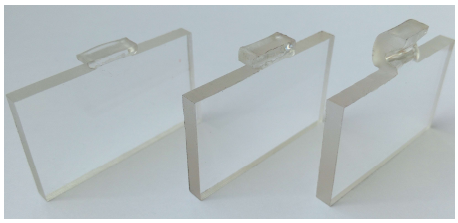
4. RESULTS AND DISCUSSION

4.1 Upset compression of the tenon

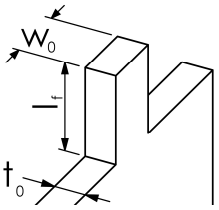
To characterize the process and identify the joinability window, the first task was to determine the critical length-to-width ratio l_f/w_0 that gives rise to plastic instability by upset compression of the tenon. To achieve that, different ratios where tested with aluminium, figure 4a, and polycarbonate, figure 4b. There are typically two deformation modes, symmetric and asymmetric, resulting from instability and out-of-plane buckling, characterized by two different trends of the force-displacement evolution.
Figure 4c summarizes the observed deformation modes and allows identifying the acceptable free tenon length to join sheets by sheet-bulk forming without instability.
In spite of the experimental conditions and geometries be exactly the some, the critical instability arise with different length-to-width ratio depending on the test specimen material. For polycarbonate, the critical instability arises when $1.0 < l_f/w_0 \leq 1.5$, but for aluminium alloy it arises when $1.5 < l_f/w_0 \leq 2.0$. The ratios before this instability give a rise to symmetric deformations and are suitable for the desired joints, instead of the non-symmetrical deformations because they are unsuitable to fix longitudinally in position two sheets perpendicular to one another.



a)



b)



		Material	
		Aluminium	**Polycarbonate**
l_f/w_0	**0.5**	Symmetric	Symmetric
	1.0	Symmetric	Symmetric
	1.5	Symmetric	Buckling
	2.0	Buckling	Buckling

c)

Figure 4. Specimens with different length-to-width ratios l_f/w_0 loaded in uniaxial compression perpendicular to the sheet thickness ($t_0/w_0 = 0.5$):

(a) Photographs of compressed aluminium specimens;
(b) Photographs of compressed polycarbonate specimens;
(c) Summary of the observed deformation modes

Figure 5 shows the experimental and finite element predicted evolutions of the force with the Z-axis displacement. For aluminium, until reaching the critical value, a steep rise is followed by a monotonic increase. The occurrence of out-of-plane buckling failure promotes a different force evolution, after the first step increase, depending on the free tenon length, the compression force may remain approximately constant or may even decrease. Afterward the strain hardening and the contact area with the punch grow, increasing the compression force.
Although the force evolution of the polymer specimens be similar, some differences should be emphasized. Independently of the deformation mode, the initial force increase up to a maximum value, then decrease slightly and starts to rise again until the end of the test. This behaviour is similar to the previously determined stress-strain curve of polycarbonate. The major difference when is reached the critical instability is related with the lower increase rate of the compression force.

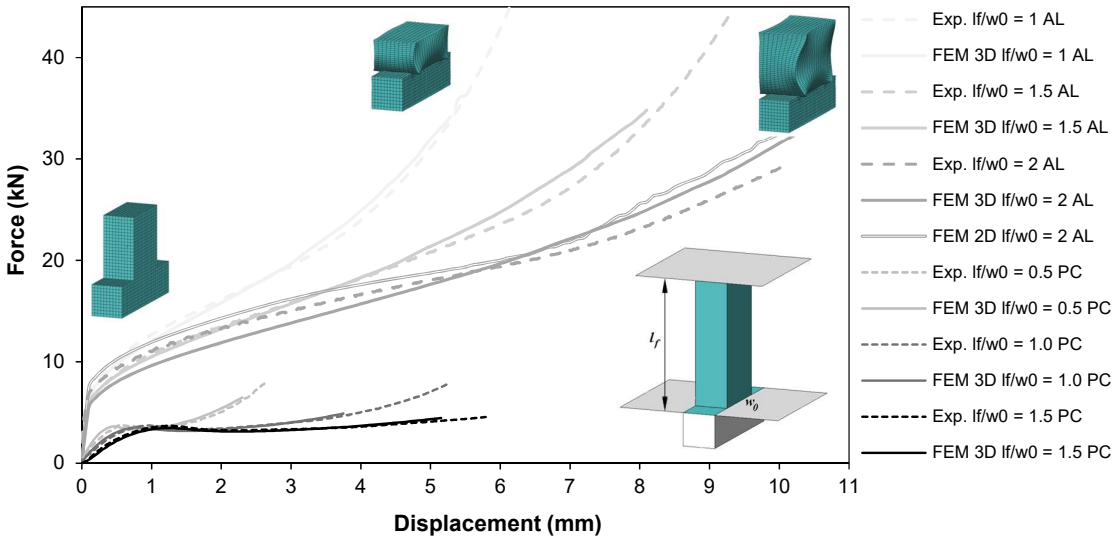


Figure 5. Experimental and finite element predicted evolution of the force with displacement during the upset compression of polycarbonate and aluminium tenons with different length-to-width ratios.

4.2 'Mortise and tenon' joints without preforming

This investigation is focused on joining similar and dissimilar sheets perpendicular to one another, based on sheet-bulk forming. Bearing in mind that, it was chosen a constant length-to-width l_f/w_0 ratio that ensures symmetric upset compression of tenons, equal to 0.5, since it promotes less material waste and lower compression forces. Thus, the aim now is to discuss the feasibility of the new mechanical joining process to connect sheets with different materials combinations, employing just one deformation stage with a flat punch (without preforming).

Four different combinations are presented, using two sheets of the same material, aluminium (Tenon_AL Mortise_AL) and Polycarbonate (Tenon_PC Mortise_PC), and two connections with arrangements of both materials (Tenon_PC Mortise_AL; Tenon_AL Mortise_PC). All connections tested were well performed and do not present defects like the upper sheet bent or defects involving buckling of the tenons [2], showing agreement with previous observations in section 4.1.

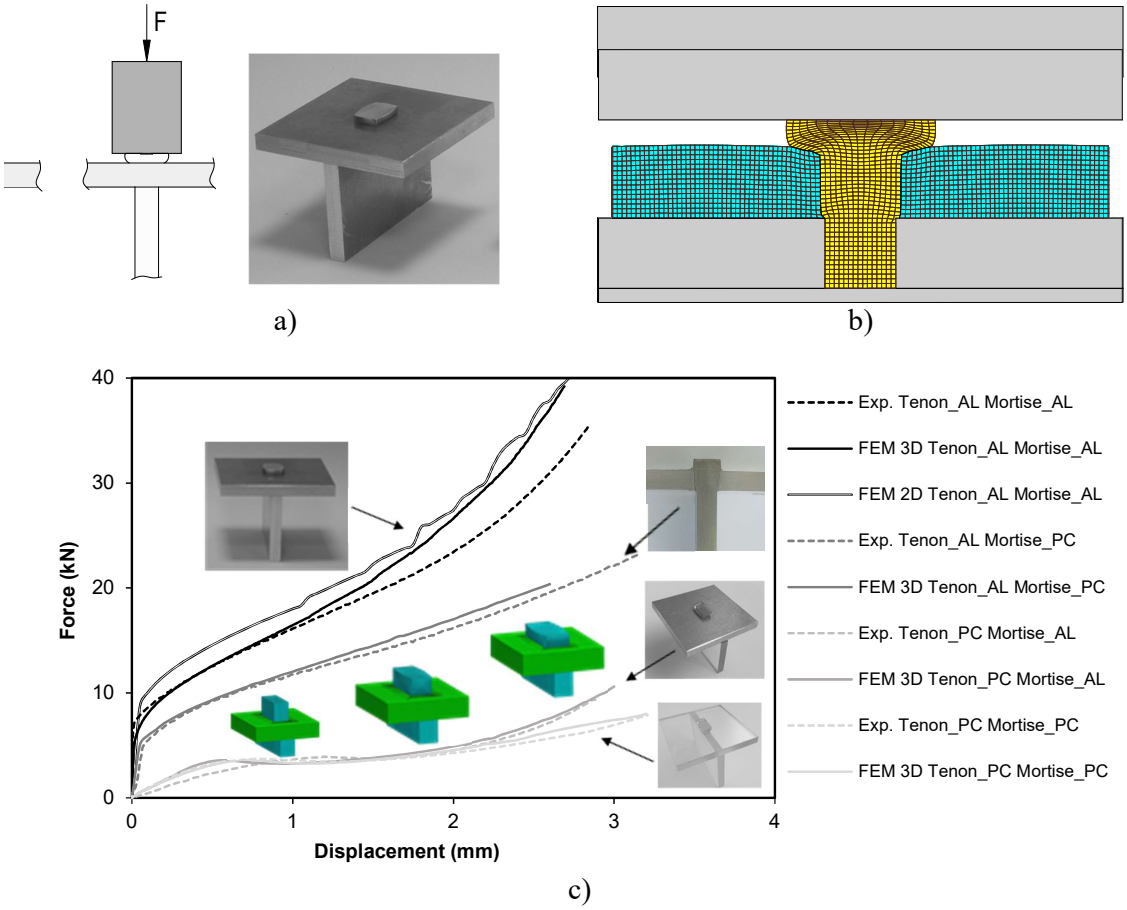


Figure 6. Joining sheets with different materials perpendicular to one another by means of a 'mortise-and-tenon' joint locked by sheet-bulk forming.

(a) Schematic representation of the joining process without preforming and photograph showing aluminium sheets joint $l_f / w_0 = 0.5$;

(b) Finite element predicted cross section of the joint in a);

(c) Experimental and finite element predicted evolution of the force with displacement. for various materials combination of sheets.

Figure 6 shows the experimental and finite element predicted evolutions of the force with displacement for different testing conditions. In all cases the maximum force is observed at the end of the unit cell joining, attaining the deformed flat surface with approximately some thickness. When the tenon is aluminium, the compression forces are much higher and with step increase in the early stage, on the other end for polycarbonate tenons, lower compression forces are required to produce the unit cell, despite the total displacement to produce the mechanical lock being identical.
The material of the mortise has also a major influence in the maximum compression force independently of tenons' materials, but far noticeable when testing the aluminium tenon. As observed in figure 6b, the mortise tends to deform when higher forces are applied, special if the material is less resistance (see cross section of Tenon_AL Mortise_PC in figure 6c). This observation is supported by the stress-strain curves of both materials, but could also be understood evaluating the project area of the tail, strain hardening and material flow. So, with aluminium tenons, the forces required to lock the sheets are higher and polycarbonate mortises promotes lower compression forces. The forces evolutions are also different and depend mainly from the material of the tenon, showing similar trends to the observed in section 4.1. The aluminium mortise tends to influence the final evolution of the connection showing a growing rate in the end, and when joining the aluminium tenons it is also observed a significant difference of the force along the connection.
A comparison of two-dimensional and three-dimensional finite element predicted evolutions of the force with displacement for test specimens with different materials allows concluding that two-dimensional models can also be successfully utilized for modelling the upset compression of tenons (figure 5) and joining sheets (figure 6) under plane strain deformation conditions.

4.3 'Mortise and tenon' joints with preforming

In order to understand the possibility to reduce the initial forces, the clearance between some connections and the risk of failure by buckling at the early stages of deformation, a two-stage technique involving preforming with a V-shaped punch was carried out before upsetting and locking the joint with the flat punch (figure 2). The V-shaped punch acts as a chisel and promotes divided material flow along its central edge to increase the width of the upper tail of the tenon. As a result, the final area of the mechanical lock is more symmetric than in case of a similar joint locked without preforming (Figure 7a, 7b).
Figure 7c shows the experimental and finite element predicted evolutions of the force with displacement using a V-shaped punch. The reduction in force derived from the utilization of a V-shape punch at the early stages of deformation, may be important for reducing the deformation inside the mortise, promoting less locking defects, or could eventually prevent the use of clamps to rigidly fix the tenons.

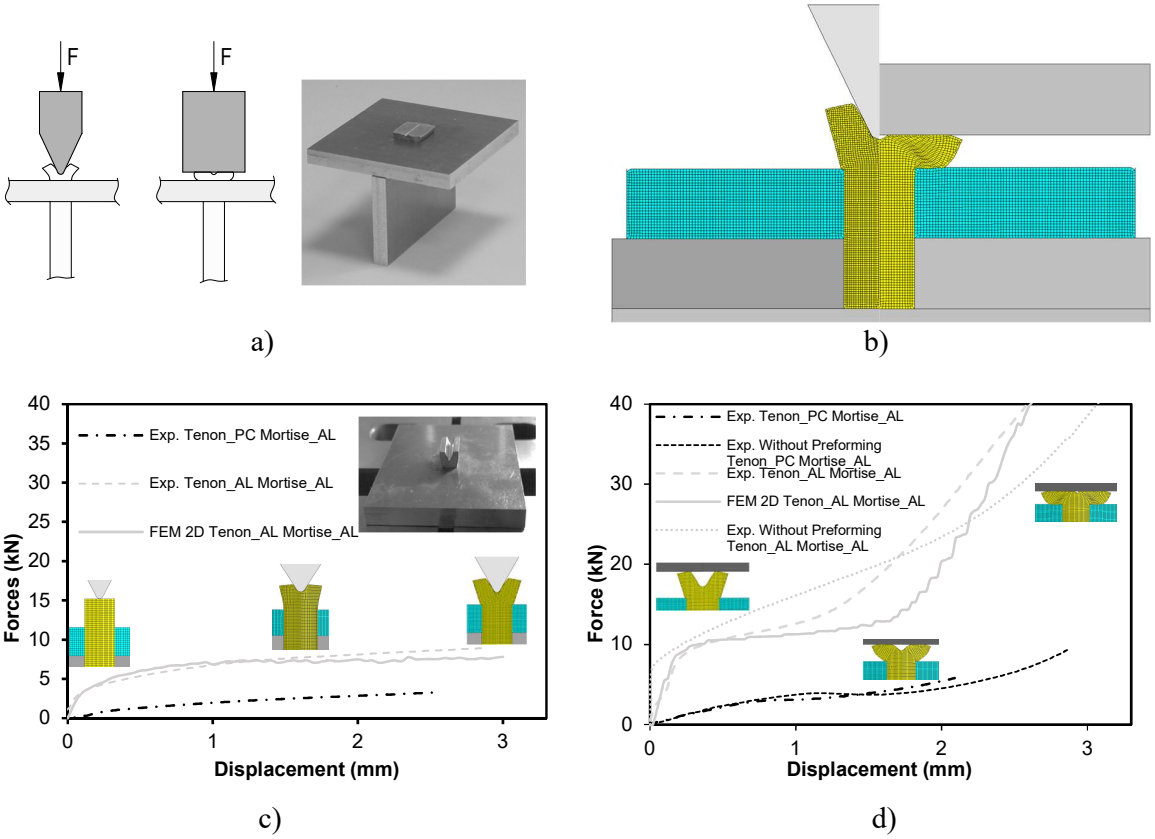


Figure 7. Joining sheets with different materials perpendicular to one another by means of a 'mortise-and-tenon' joint locked by sheet-bulk forming.

(a) Schematic representation of the joining process with a V-shaped preforming punch and Photograph showing aluminium sheets joint $l_f / w_0 = 0.5$;

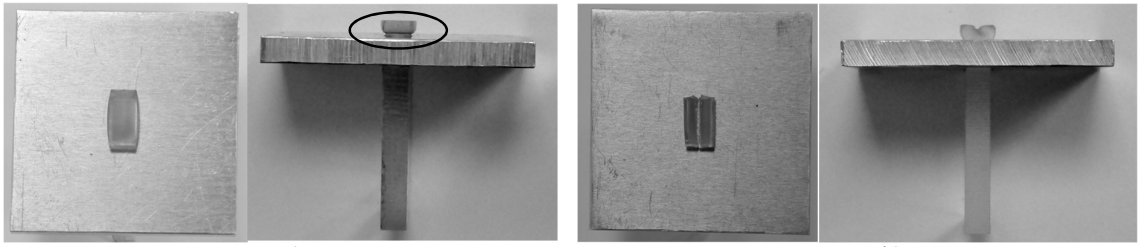
(b) Finite element predicted cross section of the joint in a);

(c) Experimental and finite element predict evolutions of the force with displacement for a unit cell with preforming by means of a V-shaped punch (P2). It is shown in one case both sheets of aluminium and other with a polycarbonate tenon and aluminium mortise;

(d) Experimental and finite element predict evolutions of the force with displacement for a unit cell with preforming with preforming by means of a Flat punch (P1) after first stage compression (P2). It is shown in one case both sheets of aluminium and other with a polycarbonate tenon and aluminium mortise. Joints produced without preforming are also presented for comparison.

Figure 7d) shows the experimental and finite element predict evolutions of the force with displacement for a unit cell by a flat punch after a performing by means of a V-shaped punch. Comparing with the one-stage specimen, the initial force at early stages is lower, but in the end of the locking, the force is similar. However, this two-stage technique is a good solution to eliminate the clearance observed in hybrid joints due to the different elastic recovery of the

materials. In fact, the elastic recovery of polycarbonate tenons may give rise to a small clearance between the deformed tenon and the aluminium mortise. This may not compromise the maximum tensile force that the joint is capable to withstand without failure, but could be unacceptable for applications in which the connection must be rigidly clamped.
Figure 8 shows the difference between the two methodologies studied to join the unit cell and helps to understand that the V-shaped punch that promotes de initial divided flux may also be used in materials with higher elastic recovery.

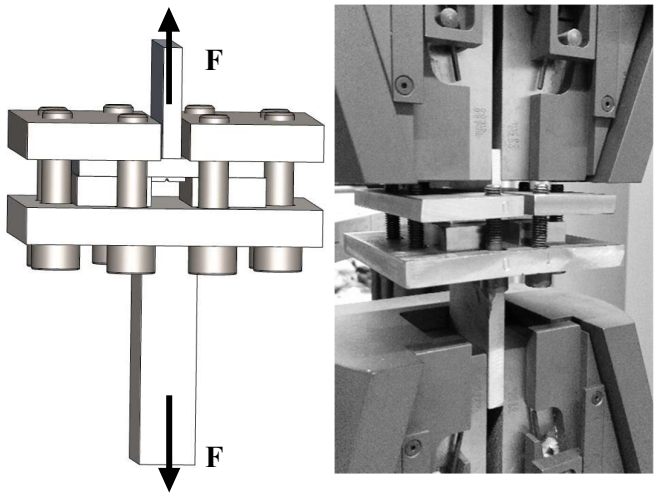


a) b)

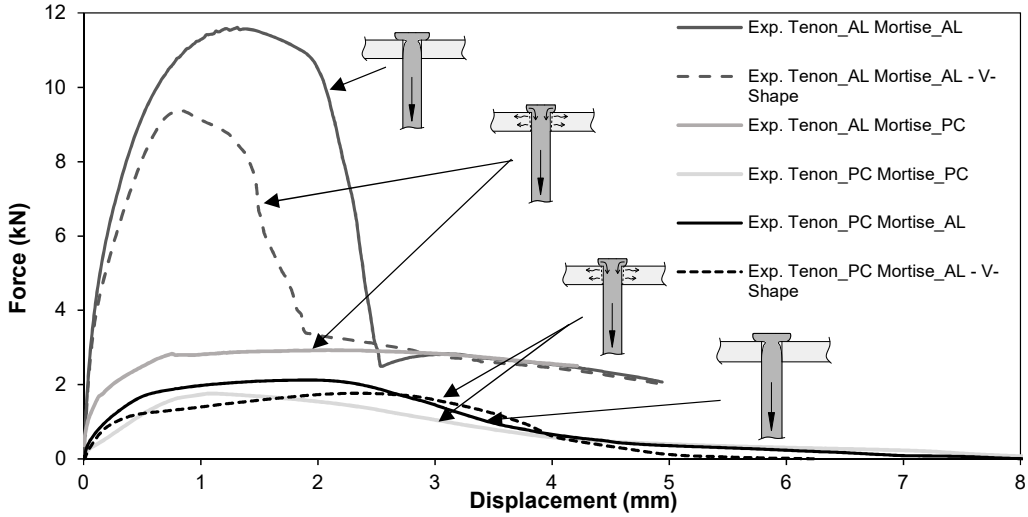
Figure 8. One-stage and two-stage variant of the joining process to eliminate the elastic recovery of polycarbonate tenons.
(a) Joint showing a clearance between the plastically deformed tenon and the mortise;
(b) Joint produced by the two-stage variant of the process without clearance.

4.4 Destructive testing of the 'mortise-and-tenon' joint

The maximum force that the joined unit cell can withstand was evaluated by destructive tensile (pull-out) tests. An experimental test setup was developed, as shown in figure 9a), which was designed to prevent the deformation of the sheets by bending during the tests.
Six different types of joints were considered (Fig.9b); (i) a monolithic aluminium-aluminium joint produced with a one-stage technique (flat punch only); (ii) a monolithic aluminium-aluminium joint produced with a two-stage technique (V-shaped punch plus flat punch); (iii) a monolithic polycarbonate-polycarbonate joint produced with a one-stage technique; (iv) a hybrid polycarbonate (tenon) - aluminium (mortise) joint produced with a one-stage technique; (v) a hybrid aluminium (tenon) - polycarbonate (mortise) joint produced with a one-stage technique; (vi) a hybrid polycarbonate (tenon) - aluminium (mortise) joint produced with a two-stage technique.



a)



b)

Figure 9 Destructive testing of the new proposed 'mortise-and-tenon' joint with and without preforming by means of a V-shaped punch.

(a) Schematic representation of the experimental setup for tensile testing;

(b) Experimental evolution of the force with displacement for different material combination and joining techniques.

The experimental evolution of the tensile force with displacement of the different joints is shown in Figure 9b. As seen, the tensile force required to destroy the joints depends significantly of the joint materials and joining technique. There are some differences in the overall performance of the joints that were produced with and without preforming. In fact, when a two-stage technic (V-shape plus flat plate) is used, the maximum force is ~15 to 20% lower

than the force supported by the sheets joined with one-stage technique.
As seen in Figure 9b, there are two different trends. In case of joints with aluminium mortise produced with a one-stage technique, the test allows concluding that failure occurs due to necking and subsequent cracking of the tenon. In contrast, in monolithic aluminium-aluminium joint produced by two-stage technique and in hybrid joints produced with a one-stage and two-stage techniques, failure is accomplished by drawing of the flat-shaped surface head of the tenon through the mortise.

5. CONCLUSIONS

The new proposed mechanical joining process is an alternative to the existing solutions that allows connecting sheets with similar and dissimilar materials perpendicular to one another. Initially upset compression tests were performed to identify the process window that is characterized by length-to-width ratio and material of the tenon. These variables influence the development of two modes of deformation, symmetric and asymmetric.
The joining feasibility of sheets perpendicular to one another with different materials were evaluated. The forces evolution are different, depending mainly from the tenons' material.
The utilization of preforming by means of a V-shaped punch diminishes the upset compression forces at the early stages of deformation. Preforming also leads to more symmetric areas of the final mechanical lock than those obtained in joints produced without preforming stage. The experimental evolution of the force with displacement during destructive testing (pull-out) reveal small differences between the 'mortise-and-tenon' joints produced with and without preforming (one-stage or two-stages techniques). The consequence of this is that, from a production point of view, the first choice should be placed on 'mortise-and-tenon' joints locked by sheet-bulk forming without preforming, unless a different head design is desired or some clearance between sheets is observed.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support provided by the Project IPL/2016/CompSBJ_ISEL, by Fundação para a Ciência e a Tecnologia of Portugal and IDMEC under LAETA - UID/EMS/50022/2013 and PDTC/EMS-TEC/0626/2014. The authors would also like to acknowledge the support provided by MCG – Mind for Metal, Carregado, Portugal.

REFERENCES

[1] Merklein M, Allwood JM, Behrens BA, Brosius A, Hagenah H,Kuzman K, Mori K, Tekkaya AE, Weckenmann A. "Bulk forming of sheet metal". *CIRP Annals-*

Manufucturing Technoligy 61:725–745, 2012

[2] Bragança, I. M. F., Silva C. M. A., Alves L. M., Martins P. A. F. "Joining sheets perpendicular to one other by sheet-bulk metal forming." *The International Journal of Advanced Manufacturing Technology*: 1-10, 2016

[3] Silva CMA, Silva MB, Alves LM, Martins PAF. "A new test for determining the mechanical and fracture behaviour of materials in sheet-bulk metal forming", Journal of Materials: Design and Applications, (in press), 2016

[4] Alves, L. M., Martins P. A.F. "Nosing of thin-walled PVC tubes into hollow spheres using a die." *The International Journal of Advanced Manufacturing Technology* 44.1 pp. 26-37, 2009.

[5] Caddell, R.M., Raghava, R.S. and Atkins, A.G. "Pressure dependent yield criteria for polymers." *Materials Science and Engineering*, 13(2), pp.113-120, 1974.

[6] Lee, J.H., Oung, J. "Yield functions and flow rules for porous pressure-dependent strain-hardening polymeric materials." *Transactions-American Society Of Mechanical Engineers Journal Of Applied Mechanics,* 67(2), pp.288-297, 2000.

[7] Sanomura, Y., Hayakawa, K. "Modification of Isotropic Hardening Model and Application of Kinematic Hardening Model to Constitutive Equation for Plastic Behavior of Hydrostatic-Pressure-Dependent Polymers." *Journal of the Society of Materials Science*, Japan, 53(2), pp.143-149, 2004.

[8] Zhu, Y.X., Liu, Y.L., Li, H.P., Yang, H., "Comparison between the effects of PVC mandrel and mandrel-cores die on the forming quality of bending rectangular H96 tube." *International Journal of Mechanical Sciences*, 76, pp.132-143, 2013.

[9] Bragança, I. M. F., Silva C. M. A., Alves L. M., Martins P. A. F. "Ligthweight joining of polymer and polymer-metal sheets by sheet-bulk forming." *Journal of Cleaner Production,* 2017.

SYMCOMP 2017
Guimarães, 6-7 April 2017

SOLVING INTEGRO-DIFFERENTIAL EQUATIONS WITH SPECTRAL METHODS

J.M.A. Matos[1*], M.J. Rodrigues[2] and J.C. Matos[3]

1: Instituto Superior de Engenharia do Porto
Centro Matemática da Universidade do Porto
Laboratório de Engenharia Matemática
Porto, Portugal
e-mail: jma@isep.ipp.pt

2: Faculdade of Ciências da Universidade do Porto
Centro Matemática da Universidade do Porto
Universidade of Porto
Porto, Portugal
e-mail: mjsrodri@fc.up.pt

3: Instituto Superior de Engenharia do Porto
Laboratório de Engenharia Matemática
Porto, Portugal
e-mail: jem@isep.ipp.pt

Keywords: Spectral methods, integro-differential equations

Abstract. *In this work we present a new approach for the implementation of operational Tau method for the solutions of linear integro-differential equations. Numerical examples are treated and compared with classic operational Tau method and collocation method.*

1 INTRODUCTION

Spectral methods are a scientific computing tool for solving differential and integral equations. These methods use matrices (called operational matrices) to represent linear operators defined in function spaces in a given orthogonal basis (see for instance [1–3]). In this paper we built operational matrices for all orthogonal polynomial basis for differential and integral operators using the tree therm recurrence relation associated to orthogonal polynomial basis. We also give a numerical example applied to an integral equation and compared it with the operational Tau method introduced in [4].

2 Operational Tau method

The key idea of the operational Tau method formulation, given in [4] and [5], is to represent in matrix form linear differential operators with polynomial coefficients. This matrix representation can be generalized to integral or integro-differential operators.

2.1 Matrix representation of linear operators in power basis

Let $\mathbb{P}[x]$ and $\mathbb{P}_n[x]$ denote the linear space of polynomials and the linear space of polynomials of degree at most n in one variable, x, respectively. Let $\nu = (\nu_0, \nu_1, \cdots)$ be a basis of $\mathbb{P}[x]$ and $p(x) \in \mathbb{P}[x]$ expanded on ν basis, $p(x) = a_0\nu_0(x) + \cdots + a_n\nu_n(x)$. We will represent $p(x)$ by the matricial product $\mathbf{p}_\nu = \mathbf{a}\,\boldsymbol{\nu}$, where $\mathbf{a} = [a_0, \cdots, a_n, 0, 0, \cdots]$ and $\boldsymbol{\nu} = [\nu_0, \nu_1, \cdots]^T$.

Let $\mathcal{L}$ be an linear differential operator with polynomial coefficients

$$\mathcal{L} = \sum_{i=0}^{m} p_i(x)\frac{\mathrm{d}^i}{\mathrm{d}x^i}, \quad p_i(x) = \sum_{j=0}^{n_i} p_{i,j}x^i, \in \mathbb{P}_{n_i}[x], \tag{1}$$

and let $y_n(x) \in \mathbb{P}_n[x]$, $y_n(x) = \sum_{i=0}^{n} a_i x^i$ written as $\mathbf{y}_n = \mathbf{a}\,\mathbf{x}$ in a matrix form. Then $\mathcal{L}\,[y_n(x)]$ has the following matrix representation, in the power basis,

$$\mathcal{L}\,[y_n(x)] = \mathbf{a}\,\mathbf{\Pi}\,\mathbf{x},$$

where the matrix $\mathbf{\Pi}$ is defined by $\mathbf{\Pi} = \sum_{i=0}^{m} \mathbf{H^i} p_i\,(\mathbf{M})$, with matrices $\mathbf{H}$ and $\mathbf{M}$ representing the linear differential and shift operator, respectively. That is, we have

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k}\,[y_n(x)] = \mathbf{a}\,\mathbf{H}^k\,\mathbf{x}$$
$$x^k\,y_n(x) = \mathbf{a}\,\mathbf{M}^k\,\mathbf{x}$$

with

$$\mathbf{H} = \begin{bmatrix} 0 & & \\ 1 & 0 & \\ & 2 & 0 \\ & & \cdots \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & 0 & 1 \\ & & & \cdots \end{bmatrix}.$$

We may also generalize this matrix representation of the operator (1) to integro and integro-diferentials linear operators, with polynomials coefficients, using the matrix

$$\mathbf{\Theta} = \begin{bmatrix} 0 & 1 & & \\ & 0 & \frac{1}{2} & \\ & & 0 & \frac{1}{3} \\ & & & \cdots \end{bmatrix}.$$

In fact, the primitive of polynomial y_n (with zero constant) can be given by

$$\int y_n(x)\,\mathrm{d}x = \mathbf{a}\,\mathbf{\Theta}\,\mathbf{x}.$$

2.2 Classic approach of operational Tau method

Now, let us consider a matrix $\boldsymbol{\nu} = [\nu_0, \nu_1, \nu_2, \cdots]^T$, where $\nu = (\nu_0, \nu_1, \nu_2, \cdots)$ is a basis of $\mathbb{P}[x]$ such that, for each non negative integer k, ν_k is a polynomial of degree k. Thus, the image of the polynomial y_n, expanded on basis ν, of the operator $\mathcal{L}$ is given by

$$\mathcal{L}[y_n] = \mathbf{a}\,\mathbf{\Pi}_{\boldsymbol{\nu}}\,\boldsymbol{\nu},$$

where, $\mathbf{\Pi}_{\boldsymbol{\nu}} = \mathbf{V}\,\mathbf{\Pi}\,\mathbf{V}^{-1}$, and $\mathbf{V}$ is the change of basis matrix that satisfies the relation $\boldsymbol{\nu}^T = \mathbf{V}\,\mathbf{x}$.

Consider a linear problem

$$\begin{aligned} \mathcal{L}u &= f, \quad x \in]a,b[&& (2)\\ g_j(u) &= \sigma_j, \quad j = 1, 2, \cdots, m && (3) \end{aligned}$$

where, g_j, $j = 1, 2, \cdots, m$ are linear functionals that represent the supplementary conditions and $f \in \mathbb{P}[x]$. A Tau solution of order n expanded on a basis ν is the solution $u_n \in \mathbb{P}_n[x]$ of the associated problem to the problem (4)

$$\begin{aligned} \mathcal{L}u &= f + h_n, \quad x \in]a,b[&& (4)\\ g_j(u) &= \sigma_j, \quad j = 1, 2, \cdots, m && (5) \end{aligned}$$

where h_n it is a perturbation polynomial. The coefficients, a_i, $i = 0, 1, \cdots, n$ of the tau solution $u_n(x) = \sum_{i=0}^{n} a_i \nu_i(x)$ are solutions of a system of a linear equations [4].

We note that this system includes the matrix, $\mathbf{\Pi}_{\boldsymbol{\nu}}$, that represent the linear operator $\mathcal{L}$, the supplementary conditions and the polynomial f. Another important remark is that we used infinity matrices to represent the linear operators. However, we need not to worry about the meaning of matrix multiplication. From the practical point of view we deal with polynomials. Thus all this products reduce to a finite number of non null parcels and the size of the finite matrices that we work depend on the number of supplementary conditions and on the hight of operator $\mathcal{L}$. For details see [4] and [5].

2.3 Matrix representation of linear operators in orthogonal basis

Consider a orthogonal polynomial basis $\nu = (\nu_0, \nu_1, \ldots)$ of $\mathbb{P}[x]$ defined by an inner product

$$< \nu_i, \nu_j >= \int_a^b \nu_i \ \nu_j \ w\mathrm{d}x = ||\nu_i||^2 \delta_{i,j}, \ i,j \in \mathbb{N}_0, \tag{6}$$

Where, as usual, w is a weight function and $||*||$ is the associated norm to the inner product $< *, * >$.

The Fourier coefficients f_i of a given function f are given by

$$f_i = \frac{1}{||\nu_i||^2} < \nu_i, f > \tag{7}$$

Assuming that all integrals $\int_a^b \nu_i \ \nu_j \ w\mathrm{d}x$ exist we will write, $f(x) = \sum_{i=0}^{\infty} f_i \nu_i(x)$, where the equality only holds when the infinite series converge to f. Then we have the following

Proposition 1. If $\mathcal{L}$ is a linear operator acting on $\mathbb{P}$ and $\mathbf{L}_\nu$ is the infinite matrix defined by

$$\mathbf{L}_\nu = [\ell_{i,j}]_{i,j\geq 0}, \quad \text{with} \quad \ell_{i,j} = \frac{1}{||\nu_i||^2} < \nu_i, \mathcal{L}[\nu_j] >, \tag{8}$$

then formally $\mathcal{L}\nu = \boldsymbol{\nu}\, \mathbf{L}_\nu$.

Proof: For each $j = 0, 1, \ldots$ we define the infinite unitary vector $\mathbf{e}_j = [\delta_{i,j}]_{i\geq 0}$. So that $\nu e_j = \nu_j$ and using (7) we get

$$\mathcal{L}[\boldsymbol{\nu}\mathbf{e}_j] = \mathcal{L}[\nu_j] = \sum_{i\geq 0} \frac{1}{||\nu_i||^2} < \nu_i, \mathcal{L}[\nu_j] > \nu_i = \sum_{i\geq 0} \ell_{i,j}\nu_i = \boldsymbol{\nu}\mathbf{L}_\nu \mathbf{e}_j, \ j = 0, 1, \ldots$$

and so $\mathcal{L}\nu = \boldsymbol{\nu}\, \mathbf{L}_\nu$, in the element wise sense. □

It is well known that a sequence of orthogonal polynomials, normalized with the condition $\nu_0 = 1$, satisfies a three term recurrence relation.

$$\begin{cases} x\nu_j = \alpha_j \nu_{j+1} + \beta_j \nu_j + \gamma_j \nu_{j-1}, \ j \geq 0 \\ \nu_{-1} = 0, \ \nu_0 = 1 \end{cases}, \tag{9}$$

This recurrence relation it is useful to find the matrices $\mathbf{M}_\nu$, $\mathbf{H}_\nu$ and $\boldsymbol{\Theta}_\nu$ that represent, respectively, the shift, differential and integral operators in ν basis [6].

For the shift operator we have,

Proposition 2. Let ν be a basis satisfying (9), defining

$$\mathbf{M}_\nu = [\mu_{i,j}]_{i,j\geq 0}, \quad \text{with} \quad \mu_{i,j} = \frac{1}{||\nu_i||^2} < \nu_i, x\nu_j >,$$

then

$$\begin{cases} \mu_{0,0} = \beta_0,\ \mu_{1,0} = \alpha_0 \\ \mu_{j-1,j} = \gamma_j,\ \mu_{j,j} = \beta_j,\ \mu_{j+1,j} = \alpha_j \quad ,\ j = 1, 2, \ldots \\ \mu_{i,j} = 0,\ |i-j| > 1 \end{cases} \tag{10}$$

and $x\nu = \boldsymbol{\nu}\mathbf{M}_\nu$.

Proof: From the definition of $\mu_{i,j}$ and (7) follows that

$$x\nu_j = \sum_{i=0}^{j+1} \mu_{i,j}\nu_i,\ j = 0, 1, \ldots$$

and using (9) we get (10). The fact that $x\nu = \boldsymbol{\nu}\mathbf{M}_\nu$, in the elementwise sense, is a consequence of proposition 1. □

For the differential operator stands the following

Proposition 3. Let ν be a basis satisfying (9), defining

$$\mathbf{H}_\nu = [\eta_{i,j}]_{i,j\geq 0}, \quad \text{with} \quad \eta_{i,j} = \frac{1}{||\nu_i||^2} < \nu_i, \frac{\mathrm{d}}{\mathrm{d}x}\nu_j >,$$

then for $j = 1, 2, \ldots$

$$\begin{cases} \eta_{i,j+1} = \frac{1}{\alpha_j}\big[\alpha_{i-1}\eta_{i-1,j} + (\beta_i - \beta_j)\eta_{i,j} \\ \qquad\qquad\qquad +\gamma_{i+1}\eta_{i+1,j} - \gamma_j\eta_{i,j-1}\big],\ i = 0, \ldots, j-1 \quad , \\ \eta_{j,j+1} = \frac{1}{\alpha_j}(\alpha_{j-1}\eta_{j-1,j} + 1) \end{cases} \tag{11}$$

and $\frac{\mathrm{d}}{dx}\nu = \boldsymbol{\nu}\mathbf{H}_\nu$.

Proof: Applying the operator $\frac{\mathrm{d}}{\mathrm{d}x}$ to both sides of (9) then

$$\nu_j + x\nu_j' = \alpha_j\nu_{j+1}' + \beta_j\nu_j' + \gamma_j\nu_{j-1}',\ j = 0, 1, \ldots$$

and, by definition of $\eta_{i,j}$

$$\nu_j + x\sum_{i=0}^{j-1}\eta_{i,j}\nu_i = \alpha_j\sum_{i=0}^{j}\eta_{i,j+1}P_i + \beta_j\sum_{i=0}^{j-1}\eta_{i,j}\nu_i + \gamma_j\sum_{i=0}^{j-2}\eta_{i,j-1}\nu_i,\ j = 0, 1, \ldots$$

and so

$$\begin{aligned} \alpha_j\sum_{i=0}^{j}\eta_{i,j+1}\nu_i \ &= \ \nu_j + \sum_{i=0}^{j-1}\eta_{i,j}(\alpha_i\nu_{i+1} + \beta_i\nu_i + \gamma_i\nu_{i-1}) \\ &\quad -\beta_j\sum_{i=0}^{j-1}\eta_{i,j}\nu_i - \gamma_j\sum_{i=0}^{j-2}\eta_{i,j-1}\nu_i \end{aligned}$$

rearranging indices and identifying similar coefficients we get (11). That $\frac{\mathrm{d}}{\mathrm{d}x}\nu = \boldsymbol{\nu}\mathbf{H}_\nu$ is a consequence of Proposition 1. □

To derive the matrix $\boldsymbol{\Theta}_\nu$ we have,

Proposition 4. Let ν be the basis satisfying (9), defining

$$\boldsymbol{\Theta}_\nu = [\theta_{i,j}]_{i,j\geq 0}, \quad \text{with} \quad \theta_{i,j} = \frac{1}{||\nu_i||^2} < \nu_i, \int \nu_j \,\mathrm{d}x >,$$

then for $j = 1, 2, \ldots$

$$\begin{cases} \theta_{j+1,j} = \dfrac{\alpha_j}{j+1} \\ \theta_{i+1,j} = -\dfrac{\alpha_i}{i+1} \displaystyle\sum_{k=i+2}^{j+1} \eta_{i,k}\theta_{k,j}, i = j-1, \ldots, 1, 0 \end{cases} \tag{12}$$

and $\int \nu \mathrm{d}x = \boldsymbol{\nu}\boldsymbol{\Theta}$.

Proof: By definition, considering that the primitive of ν_j is a polynomial of degree $j+1$ defined with an arbitrary constant term, we can write

$$\int \nu_j \mathrm{d}x = \sum_{i=1}^{j+1} \theta_{i,j} P_i, \;\; j = 0, 1, \ldots$$

Differentiating both sides and applying proposition 3

$$\nu_j = \sum_{i=1}^{j+1} \theta_{i,j} \nu_i' = \sum_{i=1}^{j+1} \theta_{i,j} \sum_{k=0}^{i-1} \eta_{k,i} \nu_k.$$

Rearranging indices and identifying similar coefficients,

$$\nu_j = \sum_{i=0}^{j} \left[\sum_{k=i+1}^{j+1} \eta_{i,k}\theta_{k,j} \right] \nu_i.$$

And so, for the coefficient of ν_j,

$$\eta_{j,j+1}\theta_{j+1,j} = 1$$

and, for the coefficients of $\nu_i, \;\; i = 0, \ldots, j-1$,

$$\sum_{k=i+1}^{j+1} \eta_{i,k}\theta_{k,j} = 0$$

The result is obtained solving for $\theta_{j+1,j}$ the first equation and for $\theta_{i+1,j}$ each one in the last set of equations.

□

3 Numerical results

As we saw above, it is possible to compute the matrix, $\mathbf{\Pi}_\nu$, that represent a linear operator using only the three-term recurrence relation of a orthogonal polynomial basis ν. Clearly, from a theoretical point of view, both matrices, $\mathbf{\Pi}$ and $\mathbf{\Pi}_\nu$ are equal. However the matrices V, used to compute $\mathbf{\Pi}$, are usually ill-conditioned which is a drawback from numerical point of view.

In this section, we will compare numerical results of Tau solutions obtained with these two approaches. We will say "classical Tau solution" when we use $\mathbf{\Pi}$ and we will say "new Tau solution" when we use $\mathbf{\Pi}_\nu$.

3.1 Example 1

Consider the Volterra integral equation

$$(x-a)^3 u(x) + \int_{-1}^{x} u(s)ds = -f(-1), \quad x \in [-1,1] \tag{13}$$

where $f(x) = \exp\left(\frac{1}{2(x-a)^2}\right)$ and a is a real parameter. The solution of (13) it is the function u defined by $u(x) = (a-x)^{-3} f(x)$.

Given $n \in \mathbb{N}$, we want to find the Chebyshev Tau solution u_n on the interval $[-1,1]$, where

$$u_n(x) = \sum_{i=0}^{n-1} c_i T_i(x).$$

We remark that the point a is a singularity of the solution u. This implies that the Tau method will converge slowly to the solution if a is close of the extreme points of $[-1,1]$ ($a \notin [-1,1]$). Thus if we want a good Tau approximation we need to choose a large enough value of n.

To compute the new Tau solution, u_n^{SD}, we use $\mathbf{\Pi}_T = (\mathbf{M}_T - a\mathbf{I})^3 + \mathbf{\Theta}_T$ while, to the classical Tau solution, u_n^D, we use $\mathbf{\Pi} = \mathbf{V}((\mathbf{M}-a\mathbf{I})^3 + \mathbf{\Theta})\mathbf{V}^{-1}$. Where the matrices $\mathbf{M}$ and $\boldsymbol{M}_T$ represent respectively, the shift operator defined above for polynomials in power basis and Chebyshev basis. The matrices $\mathbf{\Theta}$ and $\mathbf{\Theta}_T$ are similar to the matrices that represent the integral operator defined above, for polynomials in power basis and Chebyshev basis, but adapted to include the lower limit of integration.

To analyze how different $\mathbf{\Pi}$ and $\mathbf{\Pi}_T$ are, as n increases, we computed the infinity norm of $\mathbf{\Pi} - \mathbf{\Pi}_T$ for values of $n = 5, 6, \cdots, 61$, with the parameter $a = 1.25$. We show these results on Fig. 1. For small values of n the matrices $\mathbf{\Pi}$ and $\mathbf{\Pi}_T$ are almost equal. However, the infinity norms, $||\mathbf{\Pi} - \mathbf{\Pi}_T||_\infty$, increase at exponentially rate when the value of n increases. Thus we expect a different behavior of the numerical solutions u_n^D and u_n^{ST}, when we increase the value of n.

In fact, for small values of n the classic tau solutions are similar to the new ones. We illustrated this fact on left picture of Fig. 2 where we can see that the absolute

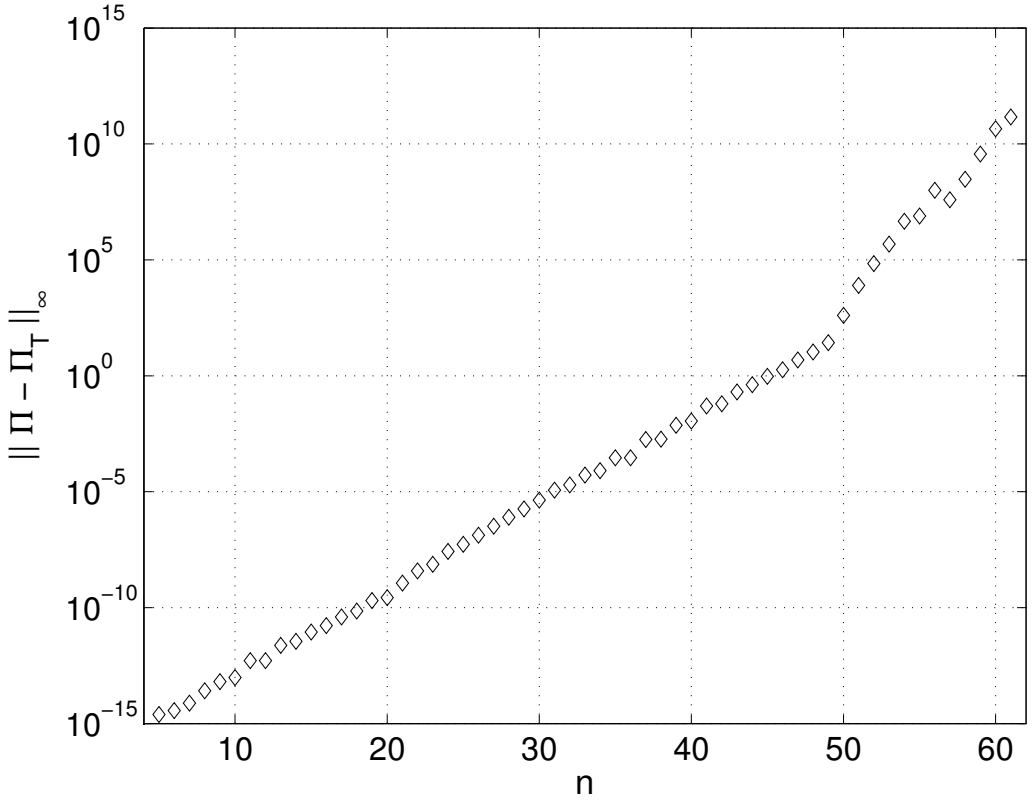


Figure 1: Infinity norm of the matrices $\mathbf{\Pi} - \mathbf{\Pi}_T$, for values of $n = 5, 6, \cdots, 61$ of the problem (13), with $a = 1.25$.

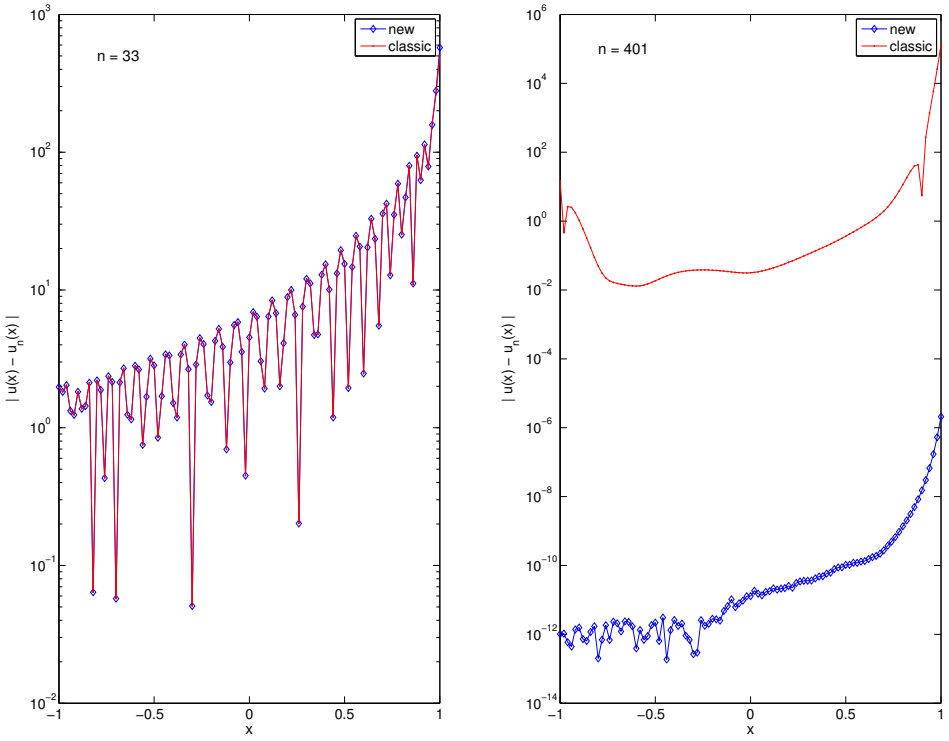


Figure 2: On left picture, we present the errors $|u_{33}(x) - u(x)|$ and on right picture, we present the errors $|u_{401}(x) - u(x)|$. On both pictures the classic Tau solution errors are represented by the red line and the new tau solution by the blue line.

errors of solutions u_{33}^{D} and u_{33}^{SD} are similar. However, the given approximations are clearly unsatisfactory. Thus to obtain better approximations we need higher order solutions. However, the classic method can not improve because the matrices $\mathbf{\Pi}$ are ill-conditioned. In contrast, the new method allows to get better approximations. To illustrate this fact we show on right picture of Fig. 2 the absolute errors of solutions u_{401}^{D} and u_{401}^{SD}.

4 Conclusions

Numerical results shows that the iterative process of built operational matrices applied to operational Tau method has advantages over the approach introduced in [4]. We note that our recursive method allows to build operational matrices in any orthogonal polynomial basis.

REFERENCES

[1] Canuto, C., Hussaini, M., Quarteroni, A, Zang, T. "Spectral Methods. Scientific Computation, fundamentals in single domains", Springer-Verlag, Berlin, 2006

[2] Funaro, D. "Polynomial Approximations of Differential Equations ", Springer-Verlag, 1992

[3] Gottlieb, D., Orszag, S. "Numerical Analysis of Spectral Methods: Theory and Applications ", SIAM-CBMS, Philadelphia, 1977

[4] Ortiz, E.L., Samara, H. "A new operational approach to the numerical solution of differential equations in terms of polynomials", *in Innovative Numerical Analysis for the Engineering Sciences* , The University Press of Virginia Vol. **27** , pp. 643-652, 1998

[5] Ortiz, E.L., Samara, H. "An operational approach to the tau method for the numerical solution of non-linear differential equations", *Computing*, Vol. **27(1)** , pp. 15-25, 1981

[6] Matos, J.M.A., Rodrigues, M.J. and Matos, J.C. " Explicit Formulae for Derivatives and Primitives of Orthogonal Polynomials", (In preparation).

INTERPRETATION OF MEDICAL IMAGES: END OF SUBJECTIVITY?

Ana Almeida[1]; LinaVieira[1,2,3] ;Sérgio Figueiredo[1]; José Alberto Rodrigues[2,4]

1: ESTeSL - Escola Superior de Tecnologia da Saúde de Lisboa, Av. D. João II, Lote 4.69.01, 1990-096 Lisboa
(ana.almeida@estesl.ipl.pt,lina.vieira@estesl.ipl.pt; sergio.figueiredo@estesl.ipl.pt)

2: GI-MOSM, ADEM, ISEL – Grupo de Investigação em Modelação e Optimização de Sistemas Multifuncionais, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa

3: IBEB - Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

4: ISEL - Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal.
(jrodrigues@adm.isel.pt)

Keywords: Medical images, Blood smears, renal scintigraphy

Abstract

In medicine, images are often used in order to obtain a diagnosis. Among the innumerable images those from optical microscopy, such as the blood smears, and the computational ones such as renal scintigraphy.

Blood smears allow identification and categorization of conditions that affect blood cells, as well as identification of microorganisms. An incorrect interpretation of images can have implications in the diagnosis and evaluation of therapeutic efficacy.

Renal scintigraphy is an imaging method of Nuclear Medicine that allows the study of renal dimension and location, evaluation of the renal cortex, and the presence of renal cortical scarring as a consequence of irreversible kidney damage after a urinary infection.

The importance in the analysis and interpretation of these studies is related with diagnosis and follow-up of renal pathologies, as well as to the therapeutic decision. However, it is known that there are currently some limitations in the detection of renal scars, which are caused by factors such as: physical, instrumental, and patient-related.

The interpretation of either blood smears or renal scintigraphy is associated with training and expertise. These are allied to the subjectivity of the observer, being a limitation in the quality and reproducibility of the results.

1. INTRODUCTION

Informatics has enabled the development of numerous applications in the area of Medicine. In medical environments, there are numerous software that allow a multiplicity of tasks such as: storage / retrieval of textual information of patients, processing and analysis of medical images [1].
The use of these systems has made it easier to access the information available in a medical environment. The physicians, for example, can use patient textual information along with information they can extract from medical examination images, whether for diagnosis, staging, follow-up, and / or assessment of response to therapy. Furthermore, medical imaging also plays a relevant role in both teaching and health research [2].
However, in order for the image to be really useful, it is necessary to understand what is represented in it. In computer science terms, it means understanding and describing the content of the image. In order to reach this level of understanding, it is necessary for each image modality to establish the relevant aspects that must be analyzed in each type of image.
Among the multitude of medical images, in this paper, the blood smears, and the renal scintigraphy are approached.
The aim of this study was to describe the importance of blood smears images, and the renal scintigraphy, in the clinic and consequently to describe the relevant aspects that must be analyzed in such images, taking into account the reduction of the subjectivity and variability in the interpretation of the findings.

2. BLOOD SMEARS

The microscopy analysis of blood smears is essential for the diagnosis and monitoring of patients suspect of haematological diseases and parasitological infections. Blood smears are made in diagnostic laboratories all over the world, from clinical haematology to clinical microbiology. Before microscopic examination, it is made a blood film on a glass microscope slide, which is allow drying, and it stains using Wright or Wright-Giemsa stain [3].
Microscopic examinations of blood smears remain the “gold standard” for laboratory confirmation in same pathologies, such as malaria [4], babesiosis [5], and an important tool in the diagnostic of other diseases as leukemia and anemia. Examination of fixed and stained blood smears in an essential tool in haematological investigation. Need of a correct diagnostic implies a careful observation of all elements figured from blood (red cells, leucocytes, and platelets).
The observation of blood smears is necessary even to confirm results obtain in automated blood count, for example to differentiate between pseudothrombocytopenia caused by platelet aggregation and a true thrombocytopenia [6]. While the first has no clinical significance, it can arise when blood is anticoagulated with EDTA (Ethylenediamine tetraacetic acid) and platelets surround /adhere to neutrophils, the later can be associated to diseases such as purpura, aplastic anaemia or lymphoma [3].

Automated analyzers perform a complete blood count, most of them based in three techniques: (1) Coulter principle – based on cell conductivity; (2) Light scatter and (3) flow cytometry. Cells can be analysed by one or more of these technologies or principles. For instance leucocytes can be classify using the Coulter principle, cell conductivity and light scatter in one equipment while others cell counters use multiple flow cytometry channels, or even others perform a combination of flow cytometry, cell conductivity, and light scattering [7]. Although some of these techniques allow us to distinguish normal cells from those with parasites, they cannot differentiate the species and stages of parasite in blood, this information is crucial, for diagnostic and monitoring diseases.

Good laboratory practices set that blood smears (thick and thin) should be observed as soon as possible, after blood collecting, and not be saved for being performed by personal more qualified, or waiting days and weeks to be confirm in reference laboratories[8]. In many cases, treatment has to be performed in a few hours; in order that patients infected are treated properly and in some cases, lives can be saved.

Optical microscopy is a time consuming activity and results obtained can have a wide inter – observer variability, with consequent errors in diagnosis and prognosis of the clinical situations[3]. Laboratories need a reduction of subjectivity and an augmentation of reproducibility in microscopic observation. For instance, in malaria diagnostic 200 fields of a blood thin smear have to be observed before a sample is considerer negative (figure 1).

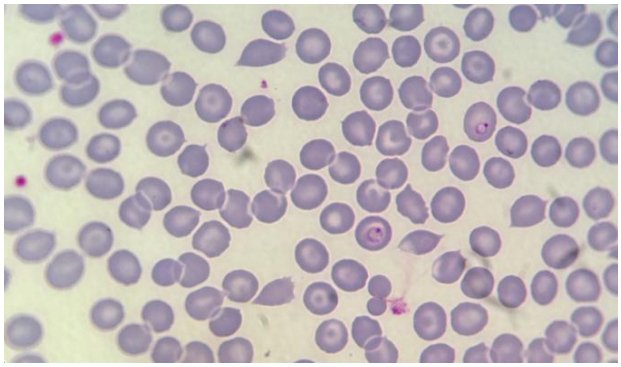


Figure1 *P. falciparum* 2 trophozoites in thin blood smear

3. RENAL SCINTIGRAPHY

Renal scintigraphy (RS) is a conventional Nuclear Medicine examination, which allows the study of kidney, referring to its size, morphology and location. Evaluation of the renal cortex mass, in particular the detection of focal cortical lesions in an infectious context (acute pyelonephritis or sequelae /scars, obstruction of renal blood supply and/or trauma) are common clinical indicators that potentially reflect the diagnostic impression based on image analysis [9].

For the RS, the most commonly used radiopharmaceutical is ^{99m}Tc-dimercaptosuccinic acid radiolabelled with technetium nighty nine metastable (^{99m}Tc-DMSA). This radiopharmaceutical is administered intravenously accumulating in the renal cells. The level of uptake of ^{99m}Tc-DMSA in such cells depends mainly on the renal blood flow, the rate of accumulation or glomerular uptake, the extraction and tubular fixation [10].
The pathophysiological mechanisms that explain the renal abnormalities observed in ^{99m}Tc-DMSA scintigraphy are multifactorial. The uptake of ^{99m}Tc-DMSA is determined by intra-renal blood flow and the function of transport of the proximal convoluted tubule within the cell membrane. Any pathological process that changes these parameters can result in regional alterations with decreased uptake [10]and proportional downgrading of the renal function.
The normal value of the renal differential function varies from 42.5% to 57.5% of the total function for each kidney [11].
^{99m}Tc-DMSA scintigraphy is nowadays recommended as the technique of choice for evaluation of renal sequelae, revealing a higher sensitivity than ultrasound and intravenous urography, in both acute and chronic pyelonephritis [12].

Hence, in this clinical context, planar-image ^{99m}Tc-DMSA is the gold standard for the diagnosis of acute pyelonephritis and renal scars, especially relevant within the pediatric population [13].

3.1. RS image acquisition

RS images are performed about two to three hours after the injection of the radiopharmaceutical, in the supine position to reduce attenuation differences, reduce respiratory movements and consequently improve spatial resolution [10].
The images can be acquired in dynamic mode or using tomographic techniques (Single Photon Emission Computed Tomography – SPECT), although in most Nuclear Medicine departments, the acquired images are static in incidences: Anterior (ANT), Posterior (POST), Right Posterior Oblique (RPO) and Left Posterior Oblique (LPO). Usually POST, RPO and LPO views are mandatory, while ANT additional image should be performed in case of horseshoe kidney, ectopic kidney and for calculating relative function [12].
The acquisition conditions in the static images are as follows: double detector gamma camera equipped with low energy high resolution collimators; 140 keV; 20% window; 256x256 image matrix and at least 300.000 counts per image should be collected. A zoom for acquisition is recommended for paediatric studies, varying between 1 and 2 as function of body size [12].

3.2 Processing techniques applied to planar ^{99m}Tc-DMSA scintigraphy

a) Estimation of Clinical Quantitative Parameters

The relative ^{99m}Tc-DMSA renal uptake may be used as a clinical quantitative index of the renal function and to evaluate the change in renal function at follow up, based on relative

renal function (RRF). Quantification can be performed using a couple of methods: 1) evaluation based on the posterior view only, with or without compensation for kidney depth; 2) evaluation based on both anterior and posterior views using their geometric mean to compensate for kidney depth [12][14].
Independently of the applied method, for the RRF assessment, different irregular regions if interest (ROI) are manually drawn on both anterior and posterior projections covering each kidney and its lower pole corresponding background, using highly contrasted images, as represented on Figure 1. [12][14].

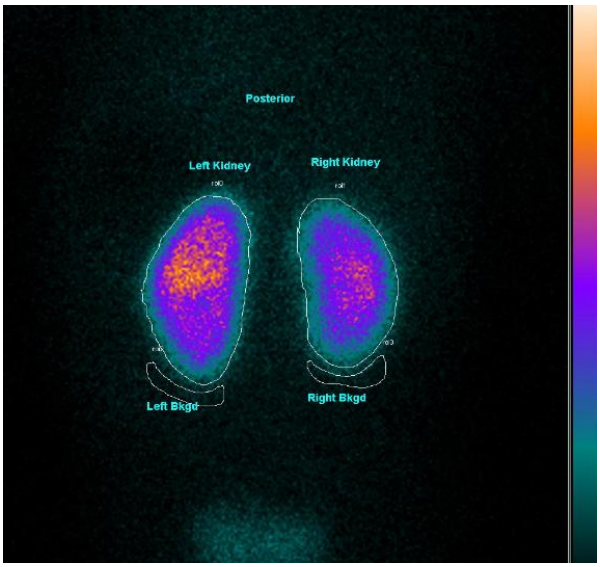


Figure 1. Image Renal scintigraphy in Posterior

The split renal uptake corrected for body background is then calculated. Normally, the split renal uptake varies from 50%/50% to 45%/55% [15]. However, it has limited value when there is bilateral impairment of function (in which case the relative function may be normal, *e.g.* 50% for each kidney) or in patients with single kidney (which, by definition, presents 100% of uptake) [10].
Thus, according to Ono et al. 2006, the computation of the percent RRF for the kidney *k* can be generally given by:

$$RRF_K = \frac{K}{K_R + K_L}.100 \quad (1)$$

where K represents the total counts detected in the specific kidney ROI (right or left), and K_R+K_L is the total counts measured for both right (K_R) and left kidney (K_L) ROIs. All kidney ROIs counts are corrected (CC) by subtracting the average value of corresponding background ROI (Bkg), when considering only posterior planar acquisition images.
When anterior and posterior images are available, *e.g.* for ectopic or malroated kidneys, it is

particulary important to apply attenuation correction factors due to the fact that kidneys may lie at different depths, affecting the image statistic [16].
Raynaud and Knipper methods can estimate kidney depth based on patient height, weight and age to correct gamma photon attenuation [10]. However, the geometric mean (GM) method is used more often and it is also usually considered as more valid to compensate kidney depth [14]. Therefore, based on the GM, the following formula can be applied for the split renal function calculation of the right kidney (RRF_{KR}):

$$RRF_{KR} = \frac{\sqrt{KRPost \times KRAnt}}{\sqrt{KRPost \times KRAnt} + \sqrt{KLPost \times KLAnt}} \times 100\% \quad (.2)$$

where Ant represents the acquisition in the anterior plan, and Post referes to the posterior plan acquisition. Evidently, based on equation 2 it is possible to obtain the relative renal function of the left kidney (RRF_{KL}).
The RRF renal function assessment performed by planar examination with GM as the kidney depth compensation is valid for common routine pediatric practice including cases with atypical renal localization [14], confirming that this additional quantitative clinical parameter can be used for the initial diagnostic evaluation and during follow-up of renal diseases [10].

b) Segmentation and quantification of structural renal damage

Recent academic image processing approaches have been investigated, particularly on kidney segmentation, extraction of morphometrics features and automatic renal damage detection.
Traditional segmentation methods generally stance difficulties due to the renal scintigraphy images, followed by noise. Usually, kidneys appear as bright regions arising from the background and its shape varies from one image to another. Kidneys appear as non-homogeneous regions in intensity, with more than one intensity maximum and diffuse edges[9].
Some mathematical methods based on morphology, such as watersheds, gradient based methods, and intensity threshold techniques tend to provide unsatisfactory segmentation results, as statement by Marcuzzo et al [9].
Recently Landgren et al [16] presented a complete automated system for detection and diagnosis of kidney lesions in scintigraphy planar ^{99m}Tc-DMSA images. They used active shape models that take kidney shape into account. The uptake of ^{99m}Tc-DMSA that allows the estimation of the RRF for each kidney, is compared to the typical uptake in a healthy kidney. This is done by creating a statistical map of normal uptake using several healthy examples where relative uptake is transformed to a common coordinate system using thin-plate splines. The pattern of potential lesions is obtained by localizing regions with lowered uptake of ^{99m}Tc-DMSA, based on some features (position, size, shape) in order to classify them as healthy or unhealthy. They have shown that linear discriminant analysis (LDA) classifier has the best performance with a mis-classification rate just over 14 % when the sensitivity is fixed to 96.5 % in order to lessen the risk of classifying an actual lesion as normal.
In a freshly new paper, Sampedro et al [13] proposed for the first time, for ^{99m}Tc-DMSA

scans, an automatic segmentation and quantification algorithm framework that seeks to provide clinically valuable indicators for the assessment of structural renal damage. In particular, they aimed to compute image-derived quantitative and subject-independent parameters designed to model accurately the underlying renal pathophysiology observed in ^{99m}Tc-DMSA scans. Such indicators were evaluated within three different contexts: automatic renal damage detection, indicators' correlation with non-imaging clinical data and early permanent renal lesion detection.

The fully automatic segmentation and quantification methodology was implemented as described in the follow pipeline (Figure **2)**

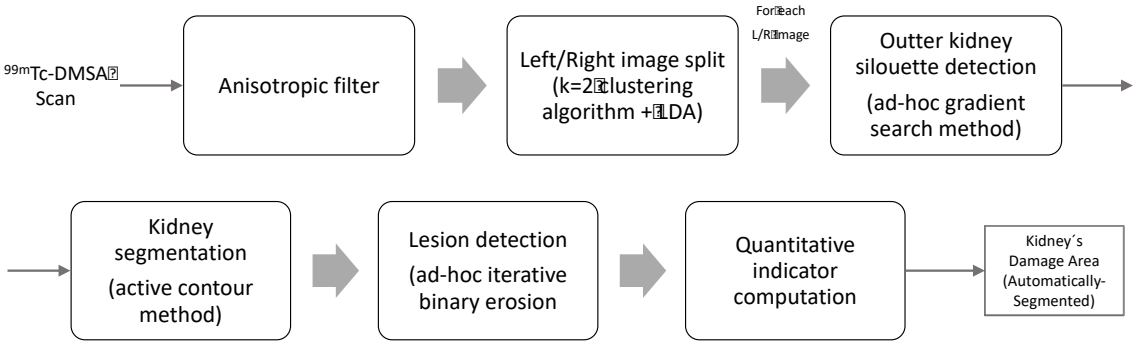


Figure 2 Adapted from [13] . Diagram of the automatic DMSA kidney lesion detection and segmentation system.

Sampedro et al [13] obtained a fully automatic lesion detection and segmentation system that was able to successfully classify ^{99m}Tc-DMSA-positive from negative scans, achieving significant diagnostic power using a Receiver operating characteristic (ROC) (Area Under the Curve, AUC=0.92, sensitivity=81% and specificity=94%).

Concerning this issues and the recent developments, the implementation of a computational framework algorithm for the quantification of structural renal damage from ^{99m}Tc-DMSA scans shows a promising potential to complement visual diagnosis and non-imaging indicators, that probability will need to be validated and cross-validated in larger cohort studies [13][16].

Sustaining by the goal of artificial intelligence, this decision support systems and computer assisted diagnosis algorithms contribute to alert the physician to cognitive biases and to reduce intra and interobserver variability, allowing faster interpretation rates and with a higher level of accuracy [17].

3.3. Final Considerations

The visual analysis of these images tend to be subjective, causing significant variability in the interpretation of the findings [9].

Quantitative ^{99m}Tc-DMSA image analysis has already been used in the field at the forefront in many papers [13]. Additionally, several authors have shown that quantification of the relative renal function (RRF) using ^{99m}Tc-DMSA is a viable proposal that provides additional

information about the functional status of the kidneys, with major clinical diagnosis importance [18][19]
As in most medical imaging scenarios, ^{99m}Tc-DMSA scan evaluation suffers from inter- and intra-observer variabilities, while it considers purely visual interpretation by physicians [19]. Moreover, its diagnostic product is descriptive and categorical, lacking a continuous modeling of the underlying renal damage [13]. Obvisiouly this variability indicates further need for the standardization of image criteria, findings, features and terminology.
With the recent computational advances, the trend to try to complement the visual diagnostic products with image-derived quantitative and observer-independent parameters is highly increasing in this research field [13].

CONCLUTIONS

The interpretation of either blood smears or renal scintigraphy is associated with training and expertise. These are allied to the subjectivity of the observer, being a limitation in the quality and reproducibility of the results.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the financial support of Project IPL, IDI&CA/ V2MIP.

REFERENCES

[1] J. S. Duncan and N. Ayache, “Medical image analysis: Progress over two decades and the challenges ahead,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 85–105, 2000.
[2] T. M. Lehmann, H. Müller, Q. Tian, N. P. Galatsanos, and D. Mlynek, “for Integrated Healthcare Solutions,” pp. 1–5.
[3] L. B. Maedel and K. Doig, “Examination of the Peripheral Blood Film and Correlation with the Complete Blood Count,” in *Hematology: clinical principles and applications*, 4ª., B. F. Rodak, G. A. Fritsma, and E. M. Keohane, Eds. 2012, pp. 192–209.
[4] A. R. Bharti, K. P. Patra, R. Chuquiyauri, M. Kosek, R. H. Gilman, A. Llanos-Cuentas, and J. M. Vinetz, “Short Report: Polymerase Chain Reaction Detection of Plasmodium vivax and Plasmodium falciparum DNA from Stored Serum Samples: Implications for Retrospective Diagnosis of Malaria.”
[5] A. E. Teal, A. Habura, J. Ennis, J. S. Keithly, and S. Madison-Antenucci, “A new real-time PCR assay for improved detection of the parasite Babesia microti.,” *J. Clin. Microbiol.*, vol. 50, no. 3, pp. 903–8, Mar. 2012.
[6] L. M. S. Dusse, L. M. Vieira, and M. das G. Carvalho, “Pseudotrombocitopenia,” *J. Bras. Patol. e Med. Lab.*, vol. 40, no. 5, pp. 321–324, Oct. 2004.
[7] J. C. Boyd and C. D. Hawker, “Automation in the Clinical Laboratory,” in *Tietz Textboook of Clinical Chemistry and Molecular Diagnostics*, C. A. Burtis, E. R. Ashwood, and D. E. Bruns, Eds. Elsevier Ltd, 2012, pp. 469–485.
[8] C. L. Lieseke and E. A. Zeibig, *Essentials of medical laboratory practice*. F.A. Davis

Company, 2012.
[9] M. Marcuzzo, P. R. Masiero, and J. Scharcanski, "Quantitative parameters for the assessment of renal scintigraphic images," *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, pp. 3438–3441, 2007.
[10] C. R. Ono, M. T. Sapienza, B. M. Machado, M. M. C. Pahl, L. Júnior, and C. A. de Paula, W., ... Buchpiguel, "Padronizacão do método para cálculo da captacão renal absoluta do 99mTc-DMSA em criancas.," *Radiol. bras*, vol. 39(1), no. 8, pp. 33–38, 2006.
[11] A. M. M. De Sousa, "Estudo qualitativo e quantitativo de Cintigrafias Dinâmicas vs Estáticas com DMSA," Universidade Católica do Porto, 2012.
[12] A. Piepsz, P. Colarinha, I. Gordon, K. Hahn, P. Olivier, I. Roca, and R. Sixt, "Revised Guidelines on 99m Tc-DMSA Scintigraphy in Children," *Eur J Nucl Med.*, vol. 28, no. 11, pp. 1–6, 2009.
[13] F. Sampedro, A. Domenech, S. Escalera, and I. Carrio, "Computing quantitative indicators of structural renal damage in pediatric DMSA scans," *Rev Esp Med Nucl Imagen Mol*, vol. 36, no. 2, pp. 72–77, 2017.
[14] D. Chroustová, J. Trnka, V. Širová, I. Urbanová, J. Langer, and J. Kubinyi, "Comparison of planar DMSA scan with an evaluation based on SPECT imaging in the split renal function assessment," *Nucl. Med. Rev.*, vol. 19, no. 1, pp. 12–17, 2016.
[15] S. T. Treves, W. E. Harmon, A. B. Packard, and A. Kuruc, "Kidneys," in *Pediatric Nuclear Medicine/PET*, S. T. (Ed. . Treves, Ed. Springer, 2007, pp. 239–285.
[16] M. Landgren, K. Sjöstrand, M. Ohlsson, D. Ståhl, K. Overgaard, Niels Christian Åström, R. Sixt, and L. Edenbrandt, "Automated System for the Detection and Diagnosis of Kidney Lesions in Children from Scintigraphy Images," in *17th Scandinavian Conference on Image Analysis (SCIA 2011)*, 2011.
[17] A. T. Taylor and E. V Garcia, "Computer assisted diagnosis in renal nuclear medicine: rationale, methodology and interpretative criteria for diuretic renography," *Semin Nucl Med.*, vol. 44, no. 2, pp. 146–158, 2014.
[18] R. Moorin, "Tc-DMSA Absolute Uptake : Normal Pediatric Values at 2 – 4 Hours," *J Nucl Med Technol*, vol. 29, pp. 16–23, 2001.
[19] M. Caglar, P. Ö. Kıratlı, and E. Karabulut, "Inter- and Intraobserver Variability of 99mTc-DMSA Renal Scintigraphy: Impact of Oblique Views," *J. Nucl. Med. Technol.*, vol. 35, no. 2, pp. 96–99, 2007.

SYMCOMP 2017
Guimarães, 6-7 April 2017

A TWO-STAGE ALGORITHM FOR SOLVING QUASI-BRITTLE FRACTURE PROBLEMS

P. Areias[1]* and J.I. Barbosa[2]

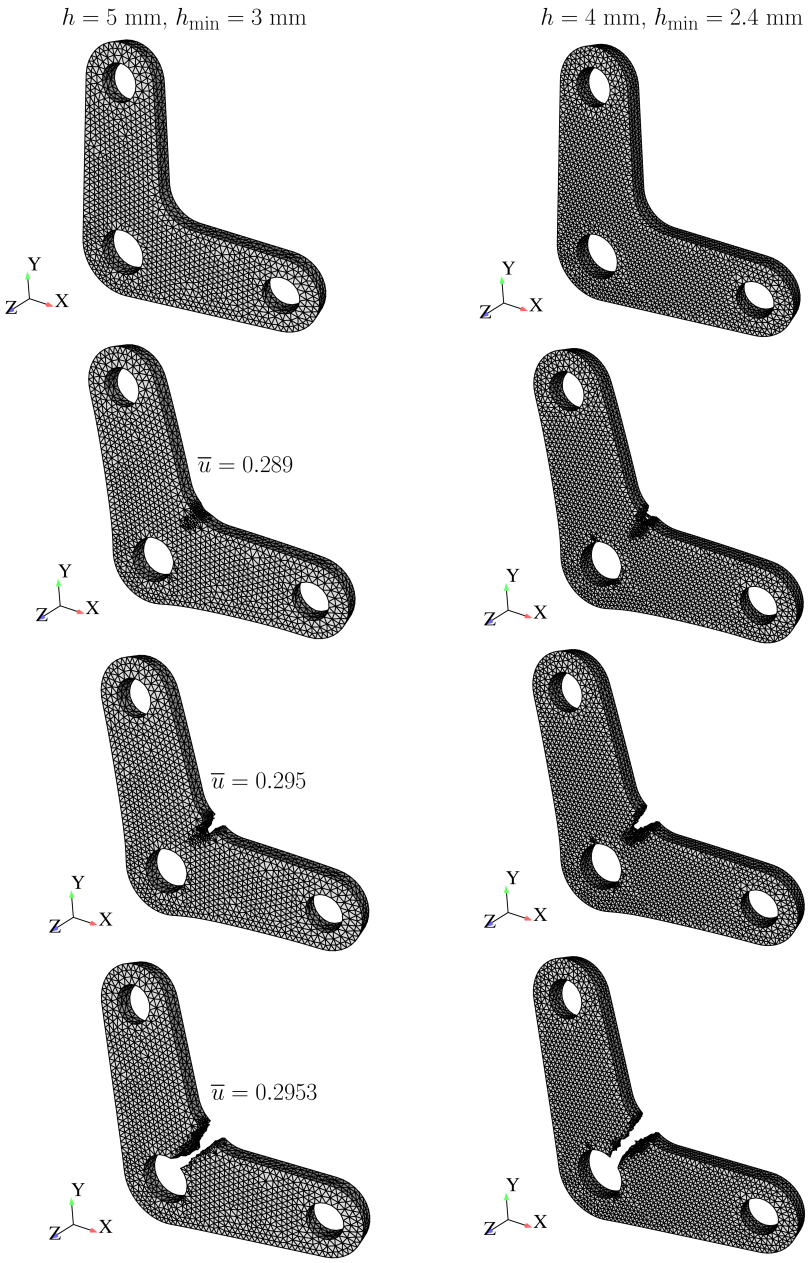
1: Department of Physics, University of Évora,
Colégio LuísAntónio Verney
Rua Romão Ramalho, 59, 7002-554 Évora

e-mail: {pmaa,jib}@uevora.pt

Keywords: Two-Stage 3D Fracture Algorithm, Mathematica, AceGen

Abstract *In this work a two-stage algorithm for solving quasi-brittle fracture problems, which includes an adaptive mesh refinement and a regularized continuum using a modified screened Poisson equation, is presented. This was made possible by extensive use of Mathematica with the AceGen add-on. A crack path criterion is not a prerequisite, as mesh refinement ultimately provides the crack path and the regularized continuum, within a smeared model, avoids the need of any type of cohesive law. Two solution parameters are included: a non-local length, appearing both in the screened-Poisson equation and the smeared model, and a mesh refinement length controlling the subdivision of elements. Initially a local approach to fracture is performed together with local remeshing and global node repositioning in order to establish the appropriate mesh for the regularized problem to be solved subsequently. This staggered approach is assessed through various quasi-brittle experiments. Both classical 2D benchmarks and 3D crack propagation examples, which predict slanting in plane stress, see the following Figure, are detailed.*

[1] P. Areias, T. Rabczuk, J. Cesar de Sa, A novel two-stage discrete crack method based on the screened Poisson equation and local mesh refinement, Comput. Mech. 58 (2016) 1003–1018.



(10× magnified displacement)

SYMCOMP 2017
Guimarães, 6-7 April 2017

FINITE ELEMENT TECHNIQUES FOR MEDICAL IMAGE PROCESSING

A. Almeida[13],J. I. Barbosa[12], A. Carvalho[12] , M. A. R. Loja[12], R. Portal[12] , J. A. Rodrigues[1] and L. Vieira[134]

1: GI-MOSM - Grupo de Investigação em Modelação e
Optimização de Sistemas Multifuncionais
ISEL - Instituto Superior de Engenharia de Lisboa,
Instituto Politécnico de Lisboa
Rua Conselheiro Emídio Navarro,
1959-007 Lisboa, Portugal

2: LAETA, IDMEC, Instituto Superior Técnico,
Universidade de Lisboa
Av. Rovisco Pais, 1,
1049-001 Lisboa, Portugal

3: ESTeSL - Escola Superior de Tecnologia da Saúde de Lisboa
Av. D. João II,
1990 - 096 Lisboa, Portugal

4: Instituto de Biofsica e Engenharia Biomdica
Faculdade de Cincias da Universidade de Lisboa
Lisboa, Portugal

Keywords: Variational method, Finite element method, Image processing

Abstract. *We consider some second-order variational model for solving medical image problems. The aim is to obtain as far as possible fine features of the initial image and identify medical pathologies. The approach consists of constructing a family of regularized functionals and to select locally and adaptively the regularization parameters. The parameters selection is performed at the discrete level in the framework of the finite element method. We present several numerical simulations to test the efficiency of the proposed approach.*

1 INTRODUCTION

Image processing and image analysis refers to some aspects of the process of computing with images. This process has been made possible by the advent of computers powerful enough to cope with the large dimensionality of image data and the complexity of the algorithms that operate on them.

The Image processing covers various aspects of data filtering, pattern recognition, feature extraction, computer aided inspection, and medical diagnosis. The above mentioned areas are treated in different scientific communities such as Imaging, Inverse Problems, Computer Vision, Signal and Image Processing, but all share the common thread of recovery of an object or one of its properties.

Currently , a core technology for solving imaging problems is regularization. The foundations of these approximation methods were laid by Tikhonov in 1963, when he generalized the classical definition of well-posedness (this generalization is now commonly referred to as conditional well-posedness). The aim of this technique is to specify a set of correctness on which it is known a priori that the considered problem has a unique solution. In 1963, Tikhonov [6] and [7] suggested what is nowadays commonly referred to as Tikhonov regularization. The abstract setting of regularization methods presented there already contains all of the variational methods that are popular in imaging.

Different techniques are applied to solve this problem, in particular, Partial Differential Equations (PDEs) are widely used and are proven to be efficient ([1] for example).

With this work we briefly introduce the variational model of RudinOsherFatemi [5] and derive the EulerLagrange equations correspondents for apply to two study cases of medical images obtained from microscopic observation with fixation and staining.

2 MATHEMATICAL MODEL

We begin by consider a 2D image of dimension $l \times m$ with given intensity $f : \Omega \subset \mathbb{R} \rightarrow [0, 1]$. At this stage we convert all colour images to grey levels. Here Ω represents the interior of the image. We would like to find a smoother image by minimizing the energy functional:

$$E(u) = \frac{1}{2} \int_{\Omega} \alpha \nabla u \cdot \nabla u d\Omega + \frac{\lambda}{2} \int_{\Omega} (u - f)^2 \, d\Omega \tag{1}$$

with boundary condition $\frac{\partial u}{\partial n} = 0$ on all four sides. The parameter $\alpha = \alpha(x, y) > 0$ is a "diffusion" coefficient whose magnitude controls how much smoothing occurs.

Restoring u from f is an inverse problem, here we use the Tikhonov regularization to ensure uniqueness:

$$\frac{1}{2} \int_{\Omega} \alpha(x, y) \nabla u \cdot \nabla u d\Omega \tag{2}$$

is a regularizer of u.

Using variational calculus, we obtain the strong form of the boundary value problem governing this problem [4] . The Euler-Lagrange equation for the energy functional E is

$$\begin{cases} \alpha\Delta u + \lambda\,(u - f) = 0 & \text{in} \quad \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{on} \quad \partial\Omega \end{cases} \tag{3}$$

By reasons of solution uniqueness we add the condition $u = 0$ to a non empty part of $\partial\Omega$.

The weak form for this problem is obtained by using the classical Green's identity, followed by the classical application of the Ritz method. We refer [3] for technical details and space approximation.

In this work, the PDE Problem (3) is numerically solved using FreeFem++ [2], a finite element free software, and we denotes u_h de numerical solution associated to the classical mesh parameter h.

3 HIDATIC LIQUID IMAGE

This first image 1 refers to a sample of a hidatic liquid, where we will find some infected pouch. This pouch derive from egg germination and can evolute to a hidatic cyst at the liver or the lung. Our goal is to automatic find infected pouch at the image.

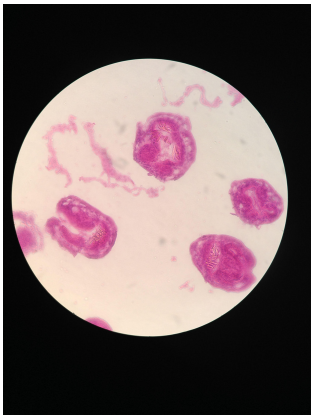


Figure 1: Hidatic liquid with infected pouch

The first step is to eliminate the black area from the image. For that, once the black area corresponds to the region where the numerical PDE solution vanish, we restrict this solution to the interior of a level line for a small value (0.05 in this case).

We observe that the infected areas corresponds to region where the numerical solution have a local great variation. This local comportment the numerical solution allow us to numerical identify the areas correspondents to infected pouch 3. As parameter control we use the level line length, assuming that its value is optimal. This assumption is based on the regularity on the geometrical shape of the infected areas.

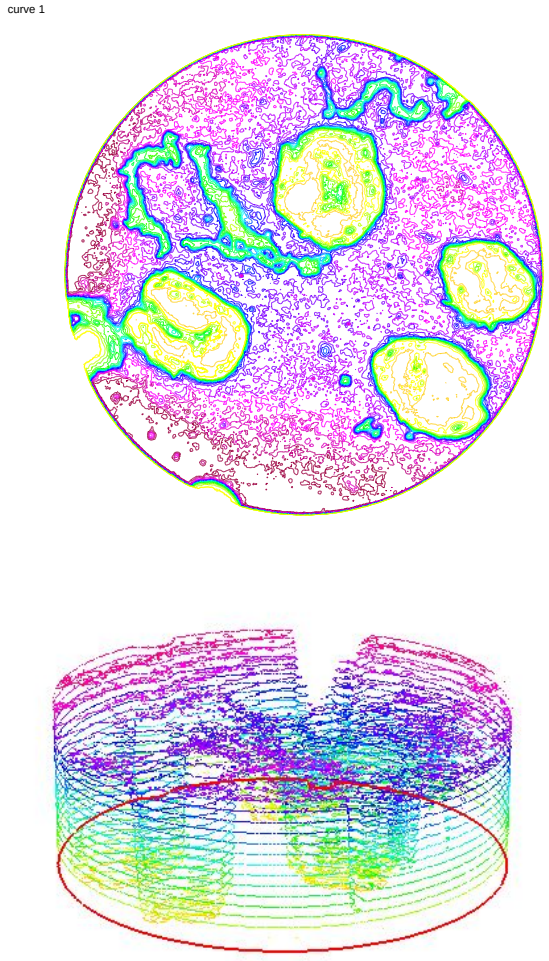


Figure 2: Numerical PDE solution

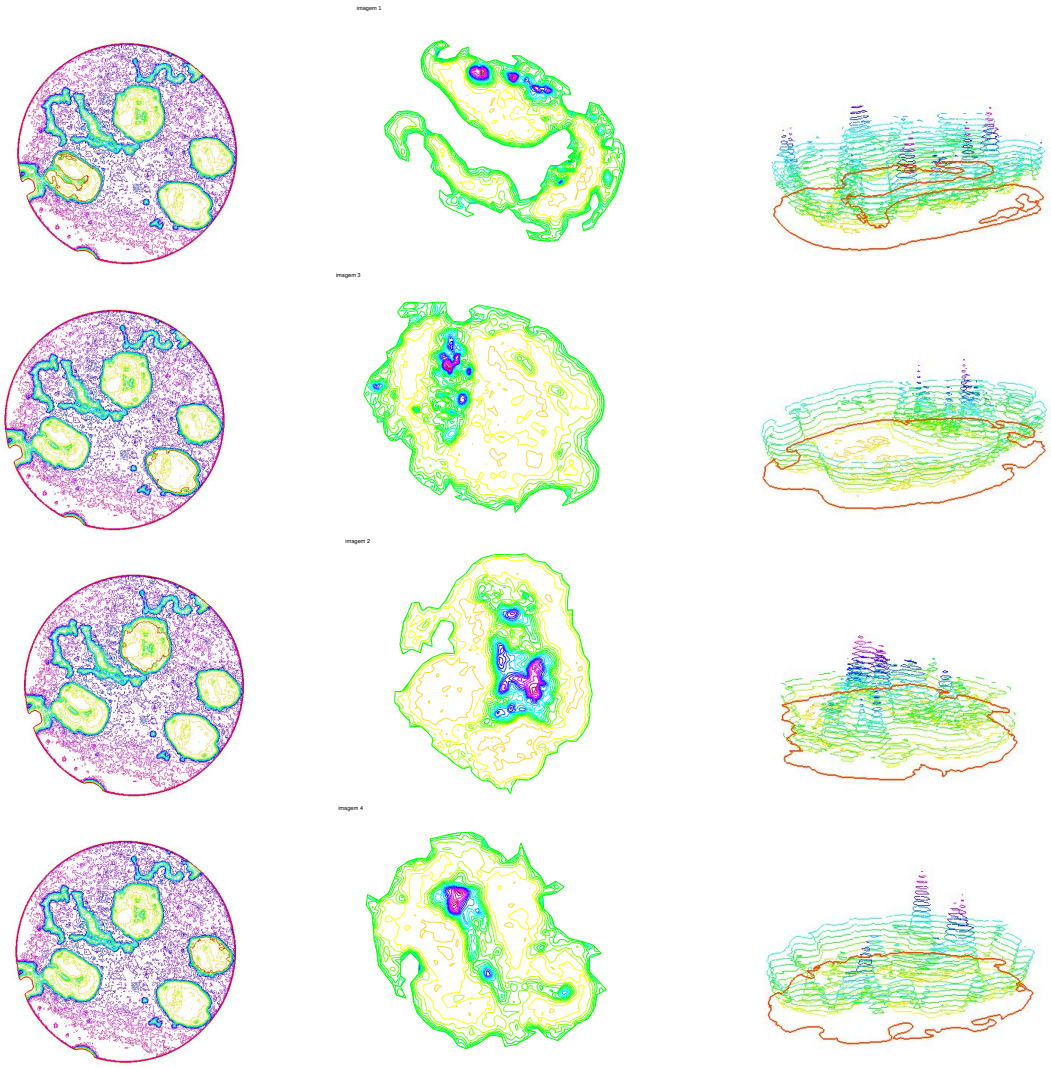


Figure 3: Local identification of the infected pouch

4 BLOOD SMEARS

A blood smears allow identification and categorization of conditions that affect blood cells, as well as identification of microorganisms. An incorrect interpretation of images can have implications in the diagnosis and evaluation of therapeutic efficacy. The second medical image 4 refers to blood smears, where we will find one parasitized cell between the normal blood cells. Like as in previous case, our goal is to automatic find parasitized cell at the image

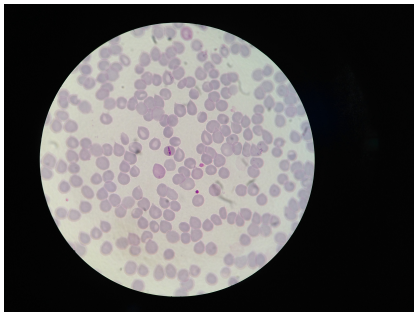


Figure 4: Blood smears with one parasitized cell

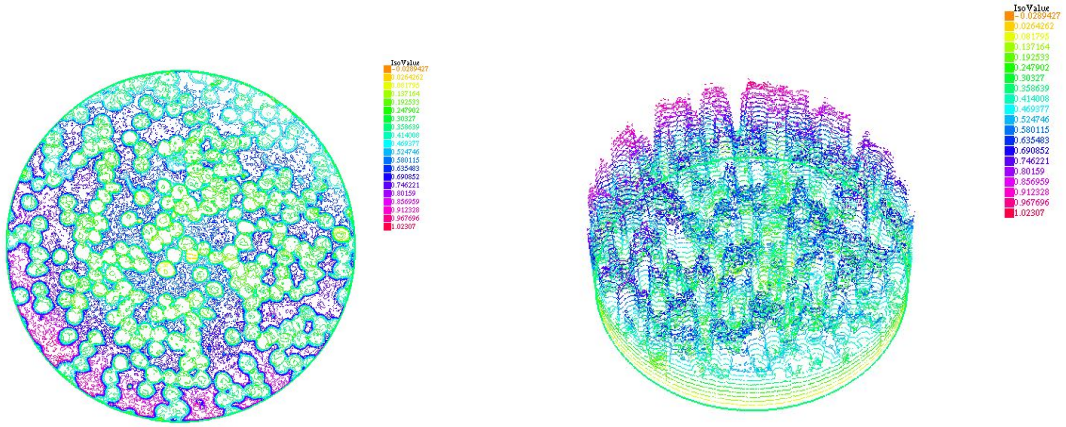


Figure 5: Numerical PDE solution

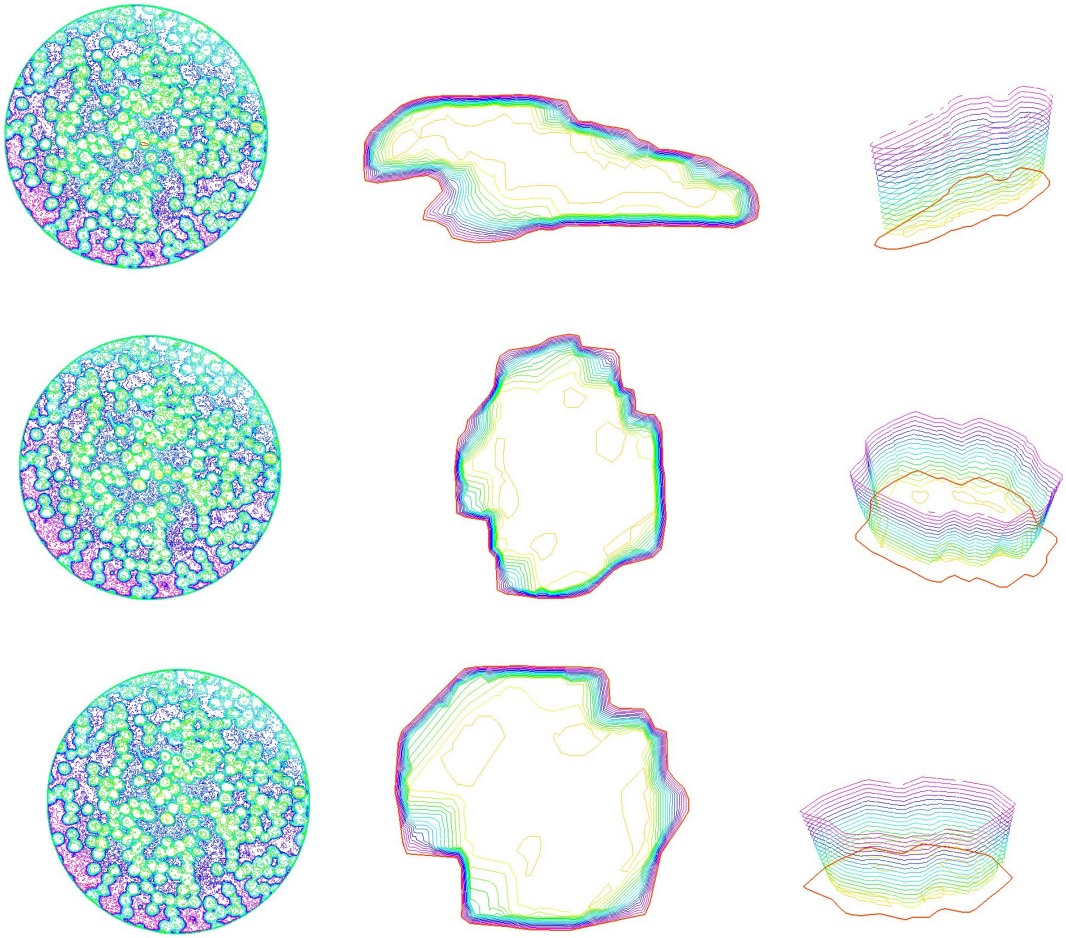


Figure 6: Local identification of the parasitized cell

From figures 6 we can conclude the criterion used at the above case do not works in this case. We observe that the infected area correspond to region where the numerical solution vanishes. By this reason we can use the integral criterion

$$\int_{\Omega_{\tau h}} \sqrt{u_h^2}\, d\Omega < \varepsilon \cdot m\left(\Omega_{\tau h}\right)$$

Where $\Omega_{\tau h}$ is the local region discretization with measure $m\left(\Omega_{\tau h}\right)$ and ε is a given parameter.

5 CONCLUSIONS

We have investigated an Finite Element Method approach for two kind of medical images based on the different parameters in the models. We started with the formulation of a linear variational model and proceed its numerical implementation based on software FreeFem++. Numerical experiments with two kinds of images were performed and showed the efficiency of the proposed method.

6 ACKNOWLEDGEMENTS

This research was supported by project V2MIP, IDI& CA-IPL 2017.

REFERENCES

[1] Esedoglu, S., Shen, J., "Digital image inpainting by Mumford-Shah-Euler model", *European Journal of Applied Mathematics*, Vol. **13**, pp. 353-370, 2002.

[2] Hecht, F., "New development in freefem++", *Journal of Numerical Mathematics*, Vol. **20**, pp. 251-265, 2002.

[3] Johnson, C., Numerical Solution of Partial Differential Equations by the Finite Element Method. CUP, 1990.

[4] Rodrigues, J. A., "Models on Variational Methods for image processing" *Proceedings of Symcomp2017* , Guimarães Portugal, 2017.

[5] Rudin, L., Osher, S., Fatemi, E., "Nonlinear total variation based noise removal algorithms", *Physica*, Vol. **60**, pp. 259-268, 1992.

[6] Tikhonov, A. N.,"Regularization of incorrectly posed problems", *Soviet Math. Dokl.*, **4**, pp. 1624-1627, 1963.

[7] Tikhonov, A. N.,"Solution of incorrectly formulated problems and the regularization methods", *Soviet Math. Dokl.*, **4**, pp. 1035-1038, 1963.

SYMCOMP 2017
Guimarães, 6-7 April 2017

AN APPLICATION OF SYMBOLIC COMPUTATION TO DERIVE A DOUBLE SCALE ASYMPTOTIC TECHNIQUE FOR LINEAR-BUCKLING OF PERIODIC MICROSTRUCTURES

Miguel Matos Neves

LAETA – IDMEC
Instituto Superior Técnico
Universidade de Lisboa,
IDMEC, IST, Av. Rovisco Pais, 1049-001 Lisboa, PORTUGAL.
e-mail: miguel.matos.neves@tecnico.ulisboa.pt, web:
https://fenix.tecnico.ulisboa.pt/homepage/ist13443

Keywords: Symbolic computation, asymptotic method, homogenization, eigenvalue buckling, cellular microstructures

Abstract *This manuscript presents a symbolic computation – developed in the MAPLE software – of interest for basic developments of asymptotic techniques like the classical homogenization by asymptotic methods. It allows to obtain a double scale asymptotic technique, for the linearized elastic stability problem of structures built of perfectly periodic microstructures which was presented in Neves, Sigmund and Bendsøe (2002) without any further detail. The use of symbolic computation here is very important as the complexity of this linear-buckling of periodic microstructures is considerably higher than the classic homogenization of material properties. The proposed asymptotic technique provides us with equations for the stability analysis at macroscopic and microscopic level. The obtained local stability condition for periodic modes corresponds to the problem of finding the minimum eigenvalue given by the quotient of two quadratic forms. This is similar to the concepts that Triantafyllidis and Bardenhagen (1996) proposed for periodic microstructure of infinite extent. From the first stationarity condition, we obtain the homogenized equations of the macroscopic elastostatic equilibrium, when $\varepsilon=d/D\rightarrow 0$ (where d is the characteristic length of the periodic cell and D the characteristic length of the structure). Collecting the lowest order ε terms, from the second stationarity condition we obtain a further set of equilibrium conditions. To avoid the difficulty of mode interactions in the two scales, it is introduced a simplification of scale separation also at the modes. To obtain an eigenvalue relation, the nonlinear terms of the derivatives of the perturbation are neglected in the elastic energy definition. These steps allows to obtain the homogenized stability problem. The commands of the symbolic computation developed for the problem are presented for each step of the development. They were kept at a basic level as the focus is only to show its application and to allow similar developments or to reproduce the results presented.*

1. INTRODUCTION

The present manuscript presents the symbolic computation implemented to formulate the proposed linearized elastic buckling problem for cellular microstructures using an asymptotic double-scale technique. The stability analysis at several scales is obtained from this asymptotic scale technique, from which appear naturally the elastic properties homogenization for the macroscopic level, requiring only certain periodicity at microscopic level. Its generalization to non-periodic case using for e.g. the Floquet-Bloch wave theory is possible but it is out of scope of this manuscript.
Periodic microstructures have been extensively used also in structural optimization, due the effectiveness of the homogenization theory to find the macroscopic properties of periodic microstructure solids [1]. Important contributions are due to [2,3] that shown evidence of connection between stability at microscopic - macroscopic levels using coercivity measures and Floquet-Bloch Theory.
Bendsϕe and Triantafyllidis present their study on the geometric scale effects using an analytical model of an infinite micro-structured medium [4]. As mentioned in the mentioned work, the main difficulty is due to the "internal buckling" instability in the microstructure scale where modes have wave-length that vary from base cell size to lengths much larger.
The use of perturbation procedure for the buckling and postbuckling analysis of elastic structures was presented in [5] which shown to be suited to be implemented as an automatic symbolic manipulation procedure. In that work the automatic procedure was used to generate the representation of the Fréchet operator for the strain field and to perform integration by parts. For the symbolic manipulation presented these authors used the program REDUCE.
This manuscript is organized as follows. In Section 2 is described the asymptotic technique proposed for the problem of linearized elastic stability of cellular microstructures. The symbolic computation methodology developed to generate the formulation is presented in section 3.

2. LINEARIZED ELASTIC BUCKLING OF A SOLID BUILT OF A CELLULAR MICROSTRUCTURE BY ASYMPTOTIC METHOD

A double-scale asymptotic technique is here applied to formulate a linearized buckling theory for a solid built from a cellular microstructure. This asymptotic technique assumes a perfect cellular microstructure and periodic displacements in the microstructure prior to buckling.
In certain circumstances the buckling instability presents periodicity and for that cases we can formally deduce the respective equations. Here the existence of "interior buckling" in the microstructure is considered, extending the previous work of Neves et al. [6]. It does not track the displacements after buckling, as after its base cell periodicity is often lost.
Let us assume the elastic response under applied load of a solid medium built of a perfectly periodic microstructure that is characterized by one base cell also named RVE, i.e. representative volume element.
The base cell is assumed to be in the domain $Y=]0,Y_1[\times]0,Y_2[\times]0,Y_3[$ with a hole (or more) in it, and the solid part of the cell is represented by ¥. The problem is to determine for which load

values the stiffness becomes non-positive definite.
At macroscale, the structure is fixed in the boundary Γ_u and a surface load is applied on boundary Γ_t . Increasing gradually the load factor P, the displacement remains unique as long as P is below the critical load factor P_{cr}, where the displacement solution is no more unique (bifurcation).
While P < Pcr, the displacement field is given by $u^{\varepsilon 0}$ where the superscript ε (ε= d/D is the microstructure size parameter) identifies the displacement dependence on the material microstructure. When achieved the load P= P_{cr} one can assume that are other possible equilibrium positions, and one can name them by the initial (classic linear elastic solution) displacement $u^{\varepsilon 0}$ and a secondary (post-buckling displacement), given by

$$\boldsymbol{u}^{\varepsilon} = \boldsymbol{u}^{\varepsilon 0} + \alpha\, \boldsymbol{u}^{\varepsilon 1} \tag{1}$$

where α is an infinitesimal real parameter (null before the bifurcation) and $\boldsymbol{u}^{\varepsilon 1}$ is the shift displacement from the initial to the secondary equilibrium position (see e.g. [7]). For simplicity no body forces f, and no tractions p, are considered inside the holes of the base cell. The total potential energy functional is

$$\Pi(\mathbf{u}^{\varepsilon}) = \mathrm{A}(\mathbf{u}^{\varepsilon}) - \mathrm{R}(\mathbf{u}^{\varepsilon}) \tag{2}$$

and its stationarity conditions characterize the equilibrium positions $u^{\varepsilon 0}$ and u^{ε} . $A(\boldsymbol{u}^{\varepsilon})$ is the total elastic strain energy of the body Ω given by

$$\mathrm{A}(\mathbf{u}^{\varepsilon}) = \frac{1}{2}\int_{\Omega^{\varepsilon}} \boldsymbol{\sigma}_{ij}(\mathbf{u}^{\varepsilon})\boldsymbol{e}_{ij}(\mathbf{u}^{\varepsilon})d\Omega \tag{3}$$

and $\mathrm{R}(\mathbf{u}^{\varepsilon})$ is the force potential given by

$$\mathrm{R}(\mathbf{u}^{\varepsilon}) = P\int_{\Gamma_t} \boldsymbol{t}_i \boldsymbol{u}_i^{\varepsilon} d\Gamma \tag{4}$$

and a linear elastic stress-strain relation is assumed

$$\sigma_{ij}(\mathbf{u}^{\varepsilon}) = E_{ijkl}\, e_{km}(\mathbf{u}^{\varepsilon}) \tag{5}$$

With the constitutive equation (5), the elastic energy density (3) takes the form

$$\mathrm{A}(\mathbf{u}^{\varepsilon}) = \frac{1}{2}\int_{\Omega^{\varepsilon}} E_{ijkl} e_{kl}(\mathbf{u}^{\varepsilon})\boldsymbol{e}_{ij}(\mathbf{u}^{\varepsilon})d\Omega \tag{6}$$

where the strain-displacement relation (the Green-Lagrangian strain tensor or Green – St-Venant strain tensor) is

$$e_{ij}(\mathbf{u}^{\varepsilon}) = \frac{1}{2}\left(\frac{\partial u_i^{\varepsilon}}{\partial x_j} + \frac{\partial u_j^{\varepsilon}}{\partial x_i}\right) + \frac{1}{2}\left(\frac{\partial u_k^{\varepsilon}}{\partial x_j}\frac{\partial u_k^{\varepsilon}}{\partial x_i}\right) \tag{7}$$

According to the homogenization theory, the displacement fields can be represented by asymptotic expansions in terms of the parameter ε [8],

$$\mathbf{u}^{\varepsilon 0}(\mathbf{x}) = \mathbf{u}^{00}(\mathbf{x}) + \varepsilon \mathbf{u}^{10}(\mathbf{x},\mathbf{y}) + \varepsilon^2 \mathbf{u}^{20}(\mathbf{x},\mathbf{y}) + ...,\mathbf{y} = \frac{\mathbf{x}}{\varepsilon} \tag{8}$$

and it is assume the same for the shift displacement [6]

$$\mathbf{u}^{\varepsilon 1}(\mathbf{x},\mathbf{y}) = \mathbf{u}^{10}(\mathbf{x},\mathbf{y}) + \varepsilon \mathbf{u}^{11}(\mathbf{x},\mathbf{y}) + \varepsilon^2 \mathbf{u}^{12}(\mathbf{x},\mathbf{y}) + ...,\mathbf{y} = \frac{\mathbf{x}}{\varepsilon} \tag{9}$$

In this formulation the macroscopic as well as the microscopic buckling modes are present, and so $u^{\varepsilon 1}(x,y)$ is also a function of y.

The expansion (9) is valid if $\mathbf{u}^{\varepsilon 1}$ presents some periodicity which can be of length greater than the base cell size. Introducing in (7) the expansions (8-9) of $\mathbf{u}^{\varepsilon}(\mathbf{x},\mathbf{y})$ according to (1), produces the following strain-displacement relation.

$$\begin{aligned} e_{ij}(\mathbf{u}^{\varepsilon}) &= \frac{1}{2}\left(\frac{\partial u_i^{0\varepsilon}}{\partial x_j} + \frac{\partial u_j^{0\varepsilon}}{\partial x_i}\right) + \frac{1}{2}\left(\frac{\partial u_k^{0\varepsilon}}{\partial x_j}\frac{\partial u_k^{0\varepsilon}}{\partial x_i}\right) \\ &+ \alpha\left\{\frac{1}{2}\left(\frac{\partial u_i^{1\varepsilon}}{\partial x_j} + \frac{\partial u_j^{1\varepsilon}}{\partial x_i}\right) + \frac{1}{2}\left(\frac{\partial u_k^{0\varepsilon}}{\partial x_j}\frac{\partial u_k^{1\varepsilon}}{\partial x_i}\right) + \frac{1}{2}\left(\frac{\partial u_k^{1\varepsilon}}{\partial x_j}\frac{\partial u_k^{0\varepsilon}}{\partial x_i}\right)\right\} \\ &+ \alpha^2\left\{\frac{1}{2}\left(\frac{\partial u_k^{1\varepsilon}}{\partial x_j}\frac{\partial u_k^{1\varepsilon}}{\partial x_i}\right)\right\} \end{aligned} \tag{10}$$

According to the linearized buckling (Euler buckling) assumption, this relation (10) can be simplified assuming that before the buckling the deformations are infinitesimal. In that case the term $\frac{1}{2}\left(\frac{\partial u_k^{0\varepsilon}}{\partial x_i}\frac{\partial u_k^{0\varepsilon}}{\partial x_j}\right)$ is small compared with the first order term $\frac{1}{2}\left(\frac{\partial u_k^{0\varepsilon}}{\partial x_i} + \frac{\partial u_k^{0\varepsilon}}{\partial x_j}\right)$.

Since there is a jump, i.e. a shift displacement from the initial to the secondary equilibrium position (the bifurcated position), the nonlinear crossed terms $\alpha\left\{\frac{1}{2}\left(\frac{\partial u_k^{0\varepsilon}}{\partial x_j}\frac{\partial u_k^{1\varepsilon}}{\partial x_i}\right) + \frac{1}{2}\left(\frac{\partial u_k^{1\varepsilon}}{\partial x_j}\frac{\partial u_k^{0\varepsilon}}{\partial x_i}\right)\right\}$ are also negligible.

From these simplifications, the strain-displacement relation is only given by

$$e_{ij}(\mathbf{u}^{\varepsilon}) = \frac{1}{2}\left(\frac{\partial u_i^{0\varepsilon}}{\partial x_j} + \frac{\partial u_j^{0\varepsilon}}{\partial x_i}\right) + \frac{\alpha}{2}\left(\frac{\partial u_i^{1\varepsilon}}{\partial x_j} + \frac{\partial u_j^{1\varepsilon}}{\partial x_i}\right) + \alpha^2\left\{\frac{1}{2}\left(\frac{\partial u_k^{1\varepsilon}}{\partial x_j}\frac{\partial u_k^{1\varepsilon}}{\partial x_i}\right)\right\} \tag{11}$$

To introduce the asymptotic expansions (8-9) in (11), remember that the differentiation involves double-scale Y-periodic functions, reason why one should consider the following form

$$\frac{\partial F(\mathbf{x},\mathbf{y})}{\partial x_j} = \frac{\partial F(\mathbf{x},\mathbf{y})}{\partial x_j} + \frac{1}{\varepsilon}\frac{\partial F(\mathbf{x},\mathbf{y})}{\partial y_j} \tag{12}$$

and the resulting strain-displacement relation (11) after be ordered according to the α powers is

$$e_{ij}(\mathbf{u}^{\varepsilon}) = e^{0}_{ij}(\mathbf{u}^{\varepsilon}) + \alpha\ e^{I}_{ij}(\mathbf{u}^{\varepsilon}) + \alpha^2 e^{II}_{ij}(\mathbf{u}^{\varepsilon}) \tag{13}$$

where

$$e^{0}_{ij}(\mathbf{u}^{\varepsilon}) = \frac{1}{\varepsilon}\left\{\frac{1}{2}\left(\frac{\partial u_i^{00}}{\partial y_j} + \frac{\partial u_j^{00}}{\partial y_i}\right)\right\} + \frac{1}{2}\left(\frac{\partial u_i^{00}}{\partial x_j} + \frac{\partial u_j^{00}}{\partial x_i}\right) + \frac{1}{2}\left(\frac{\partial u_i^{01}}{\partial y_j} + \frac{\partial u_j^{01}}{\partial y_i}\right) + \varepsilon\left\{\frac{1}{2}\left(\frac{\partial u_i^{01}}{\partial x_j} + \frac{\partial u_j^{01}}{\partial x_i}\right)\right\} \tag{14.1}$$

$$e^{I}_{ij}(\mathbf{u}^{\varepsilon}) = \frac{1}{\varepsilon}\left\{\frac{1}{2}\left(\frac{\partial u_i^{10}}{\partial y_j} + \frac{\partial u_j^{10}}{\partial y_i}\right)\right\} + \frac{1}{2}\left(\frac{\partial u_i^{10}}{\partial x_j} + \frac{\partial u_j^{10}}{\partial x_i}\right) + \frac{1}{2}\left(\frac{\partial u_i^{11}}{\partial y_j} + \frac{\partial u_j^{11}}{\partial y_i}\right) + \varepsilon\left\{\frac{1}{2}\left(\frac{\partial u_i^{11}}{\partial x_j} + \frac{\partial u_j^{11}}{\partial x_i}\right)\right\} \tag{14.2}$$

$$\begin{aligned}
e^{II}_{ij}(\mathbf{u}^{\varepsilon}) = {} & \frac{1}{\varepsilon^2}\left\{\frac{1}{2}\left(\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial u_j^{10}}{\partial y_i}\right)\right\} + \\
& \frac{1}{\varepsilon}\left\{\frac{1}{2}\left(\frac{\partial u_k^{10}}{\partial y_i}\frac{\partial u_k^{10}}{\partial y_j}\right) + \frac{1}{2}\left(\frac{\partial u_k^{10}}{\partial y_i}\frac{\partial u_k^{10}}{\partial x_j}\right) + \left(\frac{\partial u_k^{10}}{\partial y_i}\frac{\partial u_k^{11}}{\partial y_j}\right)\right\} + \\
& \left\{\frac{1}{2}\left(\frac{\partial u_k^{10}}{\partial x_i}\frac{\partial u_k^{10}}{\partial x_j}\right) + \frac{1}{2}\left(\frac{\partial u_k^{10}}{\partial x_i}\frac{\partial u_k^{11}}{\partial y_j}\right) + \frac{1}{2}\left(\frac{\partial u_k^{11}}{\partial y_i}\frac{\partial u_k^{10}}{\partial x_j}\right) + \frac{1}{2}\left(\frac{\partial u_k^{10}}{\partial y_i}\frac{\partial u_k^{11}}{\partial x_j}\right) + \frac{1}{2}\left(\frac{\partial u_k^{11}}{\partial x_i}\frac{\partial u_k^{10}}{\partial y_j}\right) + \left(\frac{\partial u_k^{11}}{\partial y_i}\frac{\partial u_k^{11}}{\partial y_j}\right)\right\} + \\
& \varepsilon\left\{\frac{1}{2}\left(\frac{\partial u_k^{11}}{\partial y_i}\frac{\partial u_k^{11}}{\partial x_j}\right) + \frac{1}{2}\left(\frac{\partial u_k^{11}}{\partial x_i}\frac{\partial u_k^{11}}{\partial y_j}\right) + \frac{1}{2}\left(\frac{\partial u_k^{11}}{\partial x_i}\frac{\partial u_k^{10}}{\partial x_j}\right) + \frac{1}{2}\left(\frac{\partial u_k^{10}}{\partial x_i}\frac{\partial u_k^{11}}{\partial x_j}\right)\right\} + \\
& \varepsilon^2\left\{\frac{1}{2}\left(\frac{\partial u_k^{11}}{\partial x_i}\frac{\partial u_k^{11}}{\partial x_j}\right)\right\}
\end{aligned} \tag{14.3}$$

The main objective is to find an approximation for $\mathbf{u}^{\varepsilon 0}$ and $\mathbf{u}^{\varepsilon 1}$ such that

$$\delta\Pi(\mathbf{u}^{\varepsilon}) = \delta A(\mathbf{u}^{\varepsilon}) - \delta R(\mathbf{u}^{\varepsilon}) = 0 \tag{15}$$

Combining (1) with (8-9) to introduce $\mathbf{u}^{\varepsilon}(\mathbf{x})$ in the total potential energy (2), that the stationarity condition (15) is obtained for a variation on $\delta\mathbf{u}^{\varepsilon} = \alpha\ (\mathbf{v}^{10}(\mathbf{x},\mathbf{y}) + \varepsilon\ \mathbf{v}^{11}(\mathbf{x},\mathbf{y}) + \varepsilon^2(...)\ )$ where $\mathbf{v}^{11} \in V_{\Omega\times Y} = \left\{\mathbf{v}(\mathbf{x},\mathbf{y}) : \mathbf{v}\big|_{\Gamma_u} = \mathbf{0}, \mathbf{v} \text{ is smooth enough and Periodic in } \mathbf{y}\right\}$. This stationarity condition for the total potential energy equilibrium has two terms in α. Equating to zero each of these α power terms, one obtains the equilibrium equations (16-17). Notice that for α^0 there is no variation (the equation vanish), but for α and α^2 we get respectively

$$\alpha\int_{\Omega^\varepsilon} E_{ijkm}\left(e^0_{ij}(\mathbf{u}^\varepsilon)\,e^I_{km}(\mathbf{v}^\varepsilon)+e^0_{ij}(\mathbf{v}^\varepsilon)\,e^I_{km}(\mathbf{u}^\varepsilon)\right)d\Omega-\alpha P\int_{\Gamma_t} t_i\,v_i\,d\Gamma=0,\forall\mathbf{v}\in V_{\Omega\times Y} \tag{16}$$

$$\alpha^2\int_{\Omega^\varepsilon} E_{ijkm}\left\{e^0_{ij}(\mathbf{u}^\varepsilon)e^{II}_{km}(\mathbf{v}^\varepsilon)+e^0_{ij}(\mathbf{v}^\varepsilon)e^{II}_{km}(\mathbf{u}^\varepsilon)+\left(e^I_{ck}(\mathbf{u}^\varepsilon)e^I_{cm}(\mathbf{v}^\varepsilon)+e^I_{ck}(\mathbf{v}^\varepsilon)e^I_{cm}(\mathbf{u}^\varepsilon)\right)+\left(e^0_{ij}(\mathbf{u}^\varepsilon)e^I_{ck}(\mathbf{u}^\varepsilon)e^I_{cm}(\mathbf{v}^\varepsilon)+e^0_{ij}(\mathbf{u}^\varepsilon)e^I_{ck}(\mathbf{v}^\varepsilon)e^I_{cm}(\mathbf{u}^\varepsilon)+e^0_{ij}(\mathbf{v}^\varepsilon)e^I_{ck}(\mathbf{u}^\varepsilon)e^I_{cm}(\mathbf{u}^\varepsilon)\right)\right\}d\Omega=0,\quad \forall\mathbf{v}\in V_{\Omega\times Y} \tag{17}$$

Making the terms with the same power of ε in (16) equal to zero, one obtains the linear elastic equations of a homogenized elastic property of a periodic solid.

$$\frac{1}{\varepsilon^2}\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{00}}{\partial y_j}\frac{\partial v_k^{10}}{\partial y_m}\right)d\Omega=0,\ \forall\mathbf{v}^{10}\in V_{\Omega\times Y} \tag{18}$$

From (18) one concludes that the first term of the expansion of u^0 is is not a function of **y**

$$u^{00}(x,y)=u^{00}(x) \tag{19}$$

The equation (20) presents two arbitrary variations, respectively v^{10} and v^{11}.

$$\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left[\left(\frac{\partial u_i^{00}}{\partial y_j}\frac{\partial v_k^{10}}{\partial x_m}\right)+\left(\frac{\partial u_i^{00}}{\partial y_j}\frac{\partial v_k^{11}}{\partial x_m}\right)+\left(\frac{\partial u_i^{00}}{\partial x_j}\frac{\partial v_k^{10}}{\partial y_m}\right)+\left(\frac{\partial u_i^{01}}{\partial y_j}\frac{\partial v_k^{10}}{\partial y_m}\right)\right]d\Omega=0, \tag{20}$$

$$\forall\mathbf{v}^{10}\in V_\Omega,\mathbf{v}^{11}\in V_{\Omega\times Y}$$

As these variations are arbitrary, one may choose v^{10} =**0** and it returns again the result (19). But, choosing v^{11} =**0** in (20), and introducing (19), results in the following equation.

$$\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left[\left(\frac{\partial u_i^{00}}{\partial x_j}\frac{\partial v_k^{10}}{\partial y_m}\right)+\left(\frac{\partial u_i^{01}}{\partial y_j}\frac{\partial v_k^{10}}{\partial y_m}\right)\right]d\Omega=0,\ \forall\mathbf{v}^{10}\in V_{\Omega\times Y} \tag{21}$$

The material heterogeneities have a characteristic dimension d much smaller than the global dimension of the structure D and the equation (21) can be rewritten as

$$\int_{\Omega^\varepsilon}\frac{1}{|Y|}\int_Y E_{ijkm}\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\frac{\partial v_k^{10}}{\partial y_m}dY\,d\Omega=0,\ \forall\mathbf{v}^{10}\in V_{\Omega\times Y} \tag{22}$$

Because $u^{00}(x,y)=u^{00}(x)$, it can be shown from (22) that

$$u_i^{01}(x,y)=-\chi_i^{km}(y)\frac{\partial u_k^{00}(x)}{\partial x_m}+C^{te}(x) \tag{23}$$

where C^{te} is an arbitrary constant and χ_p^{km} is the solution of the following elastostatic problem

$$\int_Y E_{ijpq}\frac{\partial \chi_p^{km}}{\partial y_q}\frac{\partial v_i}{\partial y_j}dY=\int_Y E_{ijkm}\frac{\partial v_i}{\partial y_j}dY,\ \forall \mathbf{v}\in V_Y \tag{24}$$

with $V_Y=\{\mathbf{v}\ \text{is smooth enough and}\ Y-\text{Periodic}\}$. This equation obtains the characteristic displacements χ_p^{km} by solving a linear system of equation over one base cell.

Now, considering the equation in ε^0 one gets

$$\varepsilon^0\int_{\Omega^\varepsilon}E_{ijkm}\left[\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\frac{\partial v_k^{10}}{\partial x_m}+\left(\frac{\partial u_i^{01}}{\partial x_j}\frac{\partial v_k^{10}}{\partial y_m}\right)\right]d\Omega=\varepsilon^0\int_{\Gamma_u}t_i v_i^{10}d\Omega,\ \forall \mathbf{v}^{10}\in V_{\Omega\times Y} \tag{25}$$

By choosing $v^{10}=\mathbf{0}$ in the equation (25), it returns again the relation (22). On the other hand, choosing $v^{11}=\mathbf{0}$ in the equation (25) results in the following equation for the static equilibrium at the macroscopic level **x**.

$$\varepsilon^0\int_{\Omega^\varepsilon}E_{ijkm}\left[\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\frac{\partial v_k^{10}}{\partial x_m}+\left(\frac{\partial u_i^{01}}{\partial x_j}\frac{\partial v_k^{10}}{\partial y_m}\right)\right]d\Omega=\varepsilon^0\int_{\Gamma_u}t_i v_i^{10}d\Omega,\ \forall \mathbf{v}^{10}\in V_{\Omega\times Y} \tag{26}$$

One can observe that $u^{01}(x,y)$ must be a function of only y, $u^{01}(x,y)=u^{01}(y)$, or the variation v^{10} cannot be a function of y in order to obtain the static equilibrium equation. The first assumption is correct and it is derived from another stationarity condition. At this point one can use the rule for the integration of a Y-periodic function when $\varepsilon\to 0$ to obtain the following equilibrium equation

$$\int_{\Omega^\varepsilon}\frac{1}{|Y|}\int_Y\left(E_{ijkm}-E_{ijpq}\frac{\partial}{\partial y_q}\chi_p^{km}\right)dY\frac{\partial u_k^{00}}{\partial x_m}\frac{\partial v_i^{10}}{\partial x_j}d\Omega=\int_{\Gamma_u}t_i v_i^{10}d\Omega,\ \forall \mathbf{v}^{10}\in V_{\Omega\times Y} \tag{27}$$

where the term within parenthesis corresponds to

$$E_{ijkm}^H=\frac{1}{|Y|}\int_Y\left(E_{ijkm}-E_{ijpq}\frac{\partial}{\partial y_q}\chi_p^{km}\right)dY \tag{28}$$

the homogeneized elastic properties of the microstructured material. The equation (27) can be rewritten in the following form:

$$\int_{\Omega^\varepsilon}E_{ijkm}^H\frac{\partial u_k^{00}}{\partial x_m}\frac{\partial v_i^{10}}{\partial x_j}d\Omega=\int_{\Gamma_u}t_i v_i^{10}d\Omega,\ \forall \mathbf{v}^{10}\in V_{\Omega\times Y} \tag{29}$$

From (29) it is clear the macroscopic displacements u^{00} can be calculated directly at macroscopic level, once the homogeneized elastic properties of material (evaluated with only

one base cell) were obtained. The microscopic and macroscopic problems are not explicitly coupled and do not require explicitly any value of ε other than be enough small (i.e. D >>d). Finally, considering the equation in ε one gets

$$\varepsilon\int_{\Omega^\varepsilon} E_{ijkm}\left[\left(\frac{\partial u_i^{01}}{\partial x_j}\frac{\partial v_k^{10}}{\partial x_m}\right)+\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\frac{\partial v_k^{11}}{\partial x_m}+\left(\frac{\partial u_i^{01}}{\partial x_j}\frac{\partial v_k^{11}}{\partial y_m}\right)\right]d\Omega=\varepsilon\int_{\Gamma_u} t_i v_i^{11} d\Gamma,$$

$$\forall \mathbf{v}^{10}\in V_\Omega, \mathbf{v}^{11}\in V_{\Omega\times Y} \qquad (30)$$

Assuming $v^{10}=\mathbf{0}$ in (30) this returns again a given relation but now w.r.t. all v^{11} variations. On the other hand, for $v^{11}=0$ we obtain

$$\varepsilon\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{01}}{\partial x_j}\frac{\partial v_k^{10}}{\partial x_m}d\Omega=0 \quad \forall \mathbf{v}^{10}\in V_{\Omega\times Y} \qquad (31)$$

From this equation, one concludes that u^{01} is is not a function of x, i.e.

$$u^{01}(x,y)=u^{01}(y) \qquad (32)$$

The equations obtained up to this point are enough to evaluate the periodic microscopic displacements (also named characteristic displacements), the homogeneized elastic properties $\mathbf{E}^H$ of the microstructure material and the macroscopic displacements u^{00}.

By equating to zero the terms with the same power of ε in the equation (17), one obtains the linearized elastic buckling equations of a microstructure solid. Anyway, notice that the equations are based on the asymptotic expansion of $u^{\varepsilon 0}$ and $u^{\varepsilon 1}$, which requires periodicity assumption on these displacements. It is clear that $u^{\varepsilon 0}$ is Y-Periodic (i.e. cell periodicity). The same cannot in general be assured for the first terms of the shift displacement expansion of $u^{\varepsilon 1}$, although one can find particular cases where it happens. For example when the "global" instability (macroscopic mode) appear before of the "local" instability (microscopic mode) then $u^{10}(x,y)=u^{10}(x)$. Also when the instability in the microstructure present certain periodicity, $u^{10}(\mathbf{x},\mathbf{y}+\delta Y)=u^{10}(\mathbf{x},\mathbf{y})$, where is not necessary a periodicity of length equal to the base cell size. The development of the asymptotic method is done in these conditions or particular cases.

Now, considering the term in ε^{-3} at equation (17) one has that

$$\frac{1}{\varepsilon^3}\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{00}}{\partial y_j}\left(\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{10}}{\partial y_m}\right)d\Omega = 0,\ \forall \mathbf{v}^{10} \in V_{\Omega\times Y} \tag{33}$$

This equation confirm that the first term of the expansion of $u^{0\varepsilon}$ is such that $u^{00}(x,y)=u^{00}(x)$.

Continuing by considering the equation in ε^{-2} one gets

$$\frac{1}{\varepsilon^2}\int_{\Omega^\varepsilon} E_{ijkm}\left\{\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{10}}{\partial y_m}+\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{10}}{\partial y_m}\right\}d\Omega = 0,\ \forall \mathbf{v}^{10} \in V_{\Omega\times Y} \tag{34}$$

As before the bifurcation the displacement field is linear elastic and Y-periodic, and assuming that u^{10} is #Y-periodic (# is a factor number) then the equation (34) can be rewritten as

$$\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{10}}{\partial y_m}d\Omega + \int_{\Omega^\varepsilon}\left(E_{ijkm}-E_{ijpq}\frac{\partial \chi_p^{km}}{\partial x_q}\right)\frac{\partial u_i^{00}}{\partial x_j}\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{10}}{\partial y_m} = 0,\ \forall \mathbf{v}^{10} \in V_{\Omega\times(\#Y)} \tag{35}$$

or simply

$$\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{10}}{\partial y_m}d\Omega + \int_{\Omega^\varepsilon}\sigma_{km}^0\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{10}}{\partial y_m} = 0,\ \forall \mathbf{v}^{10} \in V_{\Omega\times(\#Y)} \tag{36}$$

See that this equation corresponds to consider the problem at the microstructure having highly heterogeneous structure with boundary conditions. If the displacement u^{10} is not a function of y, e.g. since is there a "global" instability, then the equation (36) is trivial.

Observe that $\boldsymbol{\sigma}^0$ denotes the first approximation of the stress field in the microstructure. The higher order terms are not required by the equation (36) that represents an eigenvalue (and eigenmode) problem over all the microstructure. Homogenization was used to obtain $\mathbf{u}^{00}$ that is associated with $\boldsymbol{\sigma}^0$.

When a certain eigenmode periodicity exists in the microstructure, say a #Y-periodicity ('#' can be k=1,2,…, n or a non-integer real that multiply the Y-period length) , then also $\boldsymbol{E}^H$ and $\boldsymbol{\sigma}^0$ maintain its #Y-periodicity. So assuming a periodicity, and only in that particular case, it is possible to simplify the problem (36) to obtain a much simple problem. That is of interest to be used with Floquet-Bloch theory which is out of scope of this manuscript.

$$\int_{\#Y} E_{ijkm}\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{10}}{\partial y_m}dY + \int_{\#Y}\sigma_{km}^0\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{10}}{\partial y_m}dY = 0,\ \forall \mathbf{v}^{10} \in V_{\Omega\times(\#Y)} \tag{37}$$

Consider now the term in ε^{-1} it gives the same equation (36) considered with the variation $v^{10}=\mathbf{0}$.

$$\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left\{\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{11}}{\partial y_m}+\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{11}}{\partial y_m}\right\}d\Omega=0,\ \forall \mathbf{v}^{11}\in V_{\Omega\times Y} \tag{38}$$

Now, as v^{1i} are arbitrary functions we choose $\mathbf{v}^{11}=0$ in the equation in ε^{-1} to get the following.

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{11}}{\partial y_j}\frac{\partial v_k^{10}}{\partial y_m}\right)d\Omega+\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\left(\frac{\partial u_c^{11}}{\partial y_k}\frac{\partial v_c^{10}}{\partial y_m}\right)d\Omega+\\
&\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{10}}{\partial x_m}+\frac{\partial u_i^{10}}{\partial x_j}\frac{\partial v_k^{10}}{\partial y_m}\right)d\Omega+\\
&\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{00}}{\partial x_j}\left(\frac{\partial u_c^{10}}{\partial x_k}\frac{\partial v_c^{10}}{\partial y_m}+\frac{\partial u_c^{10}}{\partial y_m}\frac{\partial v_c^{10}}{\partial x_k}\right)d\Omega+\\
&\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{01}}{\partial x_j}\left(\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{10}}{\partial y_m}\right)d\Omega+\\
&\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{01}}{\partial y_j}\left(\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{10}}{\partial x_m}+\frac{\partial u_c^{10}}{\partial x_k}\frac{\partial v_c^{10}}{\partial y_m}\right)d\Omega=0, \qquad \forall \mathbf{v}^{11}\in V_{\Omega\times Y}
\end{aligned} \tag{39}$$

The two first terms correspond to (34) which equal to zero these terms. The equation (40) gives the connection between the simultaneously "global" and "local" instabilities.

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{10}}{\partial x_m}\right)d\Omega+\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\left(\frac{\partial u_c^{10}}{\partial y_m}\frac{\partial v_c^{10}}{\partial x_k}\right)d\Omega+\\
&\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{10}}{\partial x_j}\frac{\partial v_k^{10}}{\partial y_m}\right)d\Omega+\frac{1}{\varepsilon}\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\left(\frac{\partial u_c^{10}}{\partial x_k}\frac{\partial v_c^{10}}{\partial y_m}\right)d\Omega=0,\ \forall \mathbf{v}^{10}\in V_{\Omega\times Y}
\end{aligned} \tag{40}$$

Considering now the equation in ε^0 with the variation $v^{10}=\mathbf{0}$ it returns the following equation.

$$\begin{aligned}
&\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{10}}{\partial x_j}\frac{\partial v_k^{11}}{\partial y_m}+\frac{\partial u_i^{11}}{\partial y_j}\frac{\partial v_k^{11}}{\partial y_m}+\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{11}}{\partial x_m}\right)d\Omega+\\
&\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{00}}{\partial x_j}\left(\frac{\partial u_c^{10}}{\partial x_k}\frac{\partial v_c^{11}}{\partial y_m}+\frac{\partial u_c^{11}}{\partial y_k}\frac{\partial v_c^{11}}{\partial y_m}+\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{11}}{\partial x_m}\right)d\Omega+\\
&\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{01}}{\partial x_j}\left(\frac{\partial u_c^{10}}{\partial x_k}\frac{\partial v_c^{11}}{\partial y_m}\right)d\Omega+\\
&\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{01}}{\partial y_j}\left(\frac{\partial u_c^{10}}{\partial x_k}\frac{\partial v_c^{11}}{\partial y_m}+\frac{\partial u_c^{11}}{\partial y_k}\frac{\partial v_c^{11}}{\partial x_m}+\frac{\partial u_c^{11}}{\partial y_k}\frac{\partial v_c^{11}}{\partial y_m}\right)d\Omega=0,\ \forall \mathbf{v}^{11}\in V_{\Omega\times Y}
\end{aligned} \tag{41}$$

By rearranging the terms of (41) one obtains the following relation. The next equation reveal in which circumstances one can use this formulation for the linearized buckling problem of the homogeneized properties microstructured solid. For example, it is the case when there is no other instability mode than the macroscopic ("global") mode, i.e. u10=u10(x).

$$\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{10}}{\partial x_j}+\frac{\partial u_i^{11}}{\partial y_j}\right)\frac{\partial v_k^{11}}{\partial y_m}d\Omega+$$
$$\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{00}}{\partial x_j}+\frac{\partial u_i^{01}}{\partial y_j}\right)\left(\frac{\partial u_c^{10}}{\partial x_k}+\frac{\partial u_c^{11}}{\partial y_k}\right)\frac{\partial v_c^{11}}{\partial y_m}d\Omega+$$
$$\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{10}}{\partial y_j}\frac{\partial v_k^{11}}{\partial x_m}\right)d\Omega+$$
$$\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{00}}{\partial x_j}\left(\frac{\partial u_c^{10}}{\partial y_k}\frac{\partial v_c^{11}}{\partial x_m}\right)d\Omega+\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\frac{\partial u_i^{01}}{\partial y_j}\left(\frac{\partial u_c^{11}}{\partial y_k}\frac{\partial v_c^{11}}{\partial x_m}\right)d\Omega=0,\ \forall \mathbf{v}^{11}\in V_{\Omega\times Y} \quad (42)$$

Considering also that nonlinear terms involving derivatives of u^{11} are negligible, the equation (42) can be simplified to the following (43). This simplification means neglecting the contribution of nonlinear terms involving derivatives of the deformation u^{11} to the elastic strain energy density.

$$\varepsilon^0\int_{\Omega^\varepsilon} E_{ijkm}\left(\frac{\partial u_i^{10}}{\partial x_j}+\frac{\partial u_i^{11}}{\partial y_j}\right)\frac{\partial v_k^{11}}{\partial y_m}d\Omega=0,\ \forall \mathbf{v}^{11}\in V_{\Omega\times Y} \quad (43)$$

Again it is assumed here that the material heterogeneities have a characteristic dimension (d) much smaller than the global dimension of the structure (D). In these conditions and due to the periodic characteristic of the heterogeneities the above equation (43) can rewritten as

$$\int_{\Omega^\varepsilon}\frac{1}{|Y|}\int_Y E_{ijkm}\left(\frac{\partial u_i^{10}}{\partial x_j}+\frac{\partial u_i^{11}}{\partial y_j}\right)\frac{\partial v_k^{11}}{\partial y_m}dY\,d\Omega=0,\ \forall \mathbf{v}^{11}\in V_{\Omega\times Y} \quad (44)$$

Since $u^{10}=u^{10}(x)$, it can be shown from (44) that

$$u_i^{11}(x,y)=-\chi_i^{km}(y)\frac{\partial u_k^{10}(x)}{\partial x_m}+C^{te}(x) \quad (45)$$

Considering now $v^{11}=0$ for the equation in ε^0 one obtains the following.

$$\varepsilon^0 \int_{\Omega^\varepsilon} E_{ijkm} \left\{ \left(\frac{\partial u_i^{10}}{\partial x_j} + \frac{\partial u_i^{11}}{\partial y_j} \right) \frac{\partial v_k^{10}}{\partial x_m} + \frac{\partial u_i^{11}}{\partial x_j} \frac{\partial v_k^{10}}{\partial y_m} \right\} d\Omega +$$
$$\varepsilon^0 \int_{\Omega^\varepsilon} E_{ijkm} \frac{\partial u_i^{00}}{\partial x_j} \left(\frac{\partial u_c^{10}}{\partial x_k} \frac{\partial v_c^{10}}{\partial x_m} + \frac{\partial u_c^{11}}{\partial y_k} \frac{\partial v_c^{10}}{\partial x_m} + \frac{\partial u_c^{11}}{\partial x_k} \frac{\partial v_c^{10}}{\partial y_m} \right) d\Omega +$$
$$\varepsilon^0 \int_{\Omega^\varepsilon} E_{ijkm} \frac{\partial u_i^{01}}{\partial x_j} \left(\frac{\partial u_c^{10}}{\partial x_k} \frac{\partial v_c^{10}}{\partial y_m} + \frac{\partial u_c^{11}}{\partial y_k} \frac{\partial v_c^{10}}{\partial y_m} + \frac{\partial u_c^{10}}{\partial y_k} \frac{\partial v_c^{10}}{\partial x_m} \right) d\Omega +$$
$$\varepsilon^0 \int_{\Omega^\varepsilon} E_{ijkm} \frac{\partial u_i^{01}}{\partial y_j} \left(\frac{\partial u_c^{10}}{\partial x_k} \frac{\partial v_c^{10}}{\partial x_m} + \frac{\partial u_c^{11}}{\partial y_k} \frac{\partial v_c^{10}}{\partial x_m} + \frac{\partial u_c^{11}}{\partial x_k} \frac{\partial v_c^{10}}{\partial y_m} \right) d\Omega = 0, \ \forall \mathbf{v}^{10} \in V_{\Omega \times Y} \quad (46)$$

Considering here $u^{10}=u^{10}(x)$ [and $v^{10}=v^{10}(x)$] one obtains the following.

$$\int_{\Omega^\varepsilon} E_{ijkm} \left(\frac{\partial u_i^{10}}{\partial x_j} + \frac{\partial u_i^{11}}{\partial y_j} \right) \frac{\partial v_k^{10}}{\partial x_m} d\Omega + \int_{\Omega^\varepsilon} E_{ijkm} \left(\frac{\partial u_i^{00}}{\partial x_j} + \frac{\partial u_i^{01}}{\partial y_j} \right) \left(\frac{\partial u_c^{10}}{\partial x_k} \frac{\partial v_c^{10}}{\partial x_m} + \frac{\partial u_c^{11}}{\partial y_k} \frac{\partial v_c^{10}}{\partial x_m} \right) d\Omega = 0, \quad \forall \mathbf{v}^{10} \in V_{\Omega \times Y} \quad (47)$$

Finally, remembering that nonlinear terms involving derivatives of u^{11} are negligible, the equation (47) can be successively simplified to the following.

$$\int_{\Omega^\varepsilon} \frac{1}{|Y|} \int_Y \left(E_{ijkm} - E_{ijpq} \frac{\partial}{\partial y_q} \chi_p^{km} \right) dY \frac{\partial u_i^{10}}{\partial x_j} \frac{\partial v_k^{10}}{\partial x_m} d\Omega$$
$$+ \int_{\Omega^\varepsilon} \frac{1}{|Y|} \int_Y \left(E_{ijkm} - E_{ijpq} \frac{\partial}{\partial y_q} \chi_p^{km} \right) dY \frac{\partial u_i^{00}}{\partial x_j} \frac{\partial u_c^{10}}{\partial x_k} \frac{\partial v_c^{10}}{\partial x_m} d\Omega = 0, \ \forall \mathbf{v}^{10} \in V_{\Omega \times Y} \quad (48.1)$$

$$\int_{\Omega^\varepsilon} E_{ijkm}^H \frac{\partial u_i^{10}}{\partial x_j} \frac{\partial v_k^{10}}{\partial x_m} d\Omega + \int_{\Omega^\varepsilon} E_{ijkm}^H \frac{\partial u_i^{00}}{\partial x_j} \frac{\partial u_c^{10}}{\partial x_k} \frac{\partial v_c^{10}}{\partial x_m} d\Omega = 0 \ \forall \mathbf{v}^{10} \in V_{\Omega \times Y} \quad (48.2)$$

$$\int_{\Omega^\varepsilon} E_{ijkm}^H \frac{\partial u_i^{10}}{\partial x_j} \frac{\partial v_k^{10}}{\partial x_m} d\Omega + \int_{\Omega^\varepsilon} \sigma_{km}^0 \frac{\partial u_c^{10}}{\partial x_k} \frac{\partial v_c^{10}}{\partial x_m} d\Omega = 0 \ \forall \mathbf{v}^{10} \in V_{\Omega \times Y} \quad (48.3)$$

The equations obtained allow to evaluate the microscopic buckling as well as the macroscopic buckling.

3. SYMBOLIC COMPUTATION METHODOLOGY

A double-scale asymptotic technique is here applied to formulate a linearized buckling theory. In the following boxes is presented the application of the symbolic computation developed for the formulation presented in the previous section. The commands can be copied from

here to the command window of the symbolic tool. It should be mentioned that the script is kept at a basic level and certainly can be improved. To ease the presentation of the command lines, these steps are delimited in this text by a box.

In this developments was used the software Maple [9] but others can be equally used. From equation (7) to (10) the following commands were used.

```
##
## Linear Buckling Equations *** File 01 **eqs 7-10
##
#
# Buckling requires 2nd order term of strain
#
eij:=1/2*( diff(u(x,y,i,j),x)+diff(u(x,y,j,i),x))+1/2*( diff(u(x,y,k,i),x)*diff(u(x,y,k,j),x));
#
# The bifurcation is characterized by u=u0+alfa*u1
eij2:=subs(u(x,y,i,j)=u0(x,y,i,j)+alfa*u1(x,y,i,j),
    u(x,y,j,i)=u0(x,y,j,i)+alfa*u1(x,y,j,i),
    u(x,y,k,i)=u0(x,y,k,i)+alfa*u1(x,y,k,i),
    u(x,y,k,j)=u0(x,y,k,j)+alfa*u1(x,y,k,j),eij);
eij3:=expand(eij2):
eij4:=collect(eij3,alfa):
eija:=sort(eij4,alfa);
#
eijs:=sort((1/2*( diff(u0(x,y,k,i),x)*diff(u0(x,y,k,j),x))
       +alfa/2*( diff(u1(x,y,k,i),x)*diff(u1(x,y,k,j),x))
       +alfa^2/2*(diff(u1(x,y,k,i),x)*diff(u1(x,y,k,j),x))),alfa);
```

From equation (11) to (14) the following commands were used. Notice that before restart the computation it was used a text editor with the command find and replace on eijs to convert each diff(u1(x,y,k,i),x) to (diff(u1(x,y,k,i),x)+1/epson*diff(u1(x,y,k,i),y)).

```
## Linear Buckling Equations *** File 02 **eqs 11-14
#
# Df/Dx=df/dx+1/epson*df/dy
#
#
eijs:=1/2*((diff(u1(x,y,k,i),x)+1/epson*diff(u1(x,y,k,i),y))
        *(diff(u1(x,y,k,j),x)+1/epson*diff(u1(x,y,k,j),y)))*alfa^2
   +1/2*((diff(u1(x,y,i,j),x)+1/epson*diff(u1(x,y,i,j),y))
      +(diff(u1(x,y,j,i),x)+1/epson*diff(u1(x,y,j,i),y)))*alfa
   +1/2*((diff(u0(x,y,i,j),x)+1/epson*diff(u0(x,y,i,j),y))
      +(diff(u0(x,y,j,i),x)+1/epson*diff(u0(x,y,j,i),y)));
```

```
# The asymptotic behavior is characterized by
#                       u0=u00+epson*u01+epson**2*u02+...
#                       u1=u10+epson*u11+epson**2*u12+...
print(`Only u0=u00+epson*u01 and u1=u10+epson*u11 since epson is small`);
eijsB:=collect(subs(
    u0(x,y,i,j)=u00(x,y,i,j)+epson*u01(x,y,i,j),    u0(x,y,j,i)=u00(x,y,j,i)+epson*u01(x,y,j,i),
    u0(x,y,k,i)=u00(x,y,k,i)+epson*u01(x,y,k,i),    u0(x,y,k,j)=u00(x,y,k,j)+epson*u01(x,y,k,j),
    u1(x,y,i,j)=u10(x,y,i,j)+epson*u11(x,y,i,j),    u1(x,y,j,i)=u10(x,y,j,i)+epson*u11(x,y,j,i),
    u1(x,y,k,i)=u10(x,y,k,i)+epson*u11(x,y,k,i),    u1(x,y,k,j)=u10(x,y,k,j)+epson*u11(x,y,k,j),
    eijs),alfa):
# Obtain
eijsB0:=sort(expand(coeff(eijsB,alfa,0)),epson);
eijsB1:=sort(expand(coeff(eijsB,alfa,1)),epson);
eijsB2:=sort(expand(coeff(eijsB,alfa,2)),epson);
```

Next, equations (16-32) were obtained with considerable result interpretation.

```
#
#  Only B0 i.e. terms with power zero over alfa
#
##------------------------------------------------------
eijsB0_sum:=2*(
1/2*diff(u01(x,y,i,j),x)*epson+
1/2*diff(u00(x,y,i,j),y)/epson+
1/2*diff(u00(x,y,i,j),x)+
1/2*diff(u01(x,y,i,j),y));
eijsB0_sumV:=0:
##------------------------------------------------------
ekmsB0_sum:=2*(
1/2*diff(u01(x,y,k,m),x)*epson+
1/2*diff(u00(x,y,k,m),y)/epson+
1/2*diff(u00(x,y,k,m),x)+
1/2*diff(u01(x,y,k,m),y)):
ekmsB0_sumV:=0:
##------------------------------------------------------
eijsB1_sum:=2*(
1/2*diff(u11(x,y,i,j),x)*epson+
1/2*diff(u10(x,y,i,j),y)/epson+
1/2*diff(u10(x,y,i,j),x)+
1/2*diff(u11(x,y,i,j),y)):
eijsB1_sumV:=2*(
1/2*diff(v11(x,y,i,j),x)*epson+
1/2*diff(v10(x,y,i,j),y)/epson+
1/2*diff(v10(x,y,i,j),x)+
1/2*diff(v11(x,y,i,j),y)):
ekmsB1_sum:=2*(
1/2*diff(u11(x,y,k,m),x)*epson+
1/2*diff(u10(x,y,k,m),y)/epson+
1/2*diff(u10(x,y,k,m),x)+
1/2*diff(u11(x,y,k,m),y)):
ekmsB1_sumV:=2*(
```

```
1/2*diff(v11(x,y,k,m),x)*epson+
1/2*diff(v10(x,y,k,m),y)/epson+
1/2*diff(v10(x,y,k,m),x)+
1/2*diff(v11(x,y,k,m),y)):
##------------------------------------------------------ (16-17)
integrFxy:=sort(    expand(Eijkm*    (eijsB0_sum*ekmsB1_sumV+eijsB0_sumV*ekmsB1_sum)    )
,epslon):
##------------------------------------------------------
ioption:=1;
if (ioption = -2) then
##------------------------------------------------------ (18)
    Em2:=int(coeff(integrFxy,epson,-2),x);
elif (ioption = -1) then
##------------------------------------------------------ (20-24)
    Em1:=int(coeff(integrFxy,epson,-1),x):
    Em1v11:=expand(subs(v10(x,y,k,m)=0,v10(x,y,i,j)=0,Em1));
    Em1v10:=expand(subs(v11(x,y,k,m)=0,v11(x,y,i,j)=0,Em1));
elif (ioption = 0) then
##------------------------------------------------------ (25-29)
    Em0:=int(coeff(integrFxy,epson, 0),x):
    Em0v11:=expand(subs(v10(x,y,k,m)=0,v10(x,y,i,j)=0,Em0));
    Em0v10:=expand(subs(v11(x,y,k,m)=0,v11(x,y,i,j)=0,Em0));
elif (ioption = +1) then
##------------------------------------------------------ (30-32)
    Ep1:=int(coeff(integrFxy,epson,+1),x);
    Ep1v11:=expand(subs(v10(x,y,k,m)=0,v10(x,y,i,j)=0,Ep1));
    Ep1v10:=expand(subs(v11(x,y,k,m)=0,v11(x,y,i,j)=0,Ep1));
else
print(`ioption not valid`);
fi;
```

Finally, equations (33-48) were obtained also with considerable result interpretation effort.

```
## Linear Buckling Homogenization Equations *** File 04
#  Only  terms with power two on alfa
#     (alfap0*alfap2 + alfap1*alfap1)
#
##------------------------------------------------------
eijsB0_sum:=2*(
1/2*diff(u01(x,y,i,j),x)*epson+1/2*diff(u00(x,y,i,j),x)+1/2*diff(u01(x,y,i,j),y)):
eijsB0_sumV:=0:
##------------------------------------------------------
ekmsB0_sum:=2*(
1/2*diff(u01(x,y,k,m),x)*epson+1/2*diff(u00(x,y,k,m),x)+1/2*diff(u01(x,y,k,m),y)):
ekmsB0_sumV:=0:
##------------------------------------------------------
eijsB1_sum:=2*(
1/2*diff(u11(x,y,i,j),x)*epson+1/2*diff(u10(x,y,i,j),y)/epson+
1/2*diff(u10(x,y,i,j),x)+1/2*diff(u11(x,y,i,j),y)):
eijsB1_sumV:=2*(
1/2*diff(v11(x,y,i,j),x)*epson+1/2*diff(v10(x,y,i,j),y)/epson+
```

```
1/2*diff(v10(x,y,i,j),x)+1/2*diff(v11(x,y,i,j),y)):
ekmsB1_sum:=2*(
1/2*diff(u11(x,y,k,m),x)*epson+1/2*diff(u10(x,y,k,m),y)/epson+
1/2*diff(u10(x,y,k,m),x)+1/2*diff(u11(x,y,k,m),y)):
ekmsB1_sumV:=2*(
1/2*diff(v11(x,y,k,m),x)*epson+1/2*diff(v10(x,y,k,m),y)/epson+
1/2*diff(v10(x,y,k,m),x)+1/2*diff(v11(x,y,k,m),y)):
##------------------------------------------------------
eijsB2:=
1/2*diff(u11(x,y,c,i),x)*diff(u11(x,y,c,j),x)*epson^2+
1/2*diff(u11(x,y,c,i),y)*diff(u11(x,y,c,j),x)*epson+1/2*diff(u11(x,y,c,i),x)*diff(u11(x,y,c,j),y)*epson+
1/2*diff(u11(x,y,c,i),x)*diff(u10(x,y,c,j),x)*epson+1/2*diff(u10(x,y,c,i),x)*diff(u11(x,y,c,j),x)*epson+
1/2*diff(u10(x,y,c,i),x)*diff(u10(x,y,c,j),y)/epson+1/2*diff(u11(x,y,c,i),y)*diff(u10(x,y,c,j),y)/epson+
1/2*diff(u10(x,y,c,i),y)*diff(u11(x,y,c,j),y)/epson+1/2*diff(u10(x,y,c,i),y)*diff(u10(x,y,c,j),x)/epson+
1/2*diff(u10(x,y,c,i),y)*diff(u10(x,y,c,j),y)/epson^2+
1/2*diff(u10(x,y,c,i),y)*diff(u11(x,y,c,j),x)+1/2*diff(u10(x,y,c,i),x)*diff(u11(x,y,c,j),y)+
1/2*diff(u11(x,y,c,i),y)*diff(u10(x,y,c,j),x)+1/2*diff(u10(x,y,c,i),x)*diff(u10(x,y,c,j),x)+
1/2*diff(u11(x,y,c,i),x)*diff(u10(x,y,c,j),y)+1/2*diff(u11(x,y,c,i),y)*diff(u11(x,y,c,j),y):

eijsB2_V:=
1/2*diff(u11(x,y,c,i),x)*diff(v11(x,y,c,j),x)*epson^2+
1/2*diff(v11(x,y,c,i),x)*diff(u11(x,y,c,j),x)*epson^2+
1/2*diff(u11(x,y,c,i),y)*diff(v11(x,y,c,j),x)*epson+1/2*diff(v11(x,y,c,i),y)*diff(u11(x,y,c,j),x)*epson+
1/2*diff(u11(x,y,c,i),x)*diff(v11(x,y,c,j),y)*epson+1/2*diff(v11(x,y,c,i),x)*diff(u11(x,y,c,j),y)*epson+
1/2*diff(u11(x,y,c,i),x)*diff(v10(x,y,c,j),x)*epson+1/2*diff(v11(x,y,c,i),x)*diff(u10(x,y,c,j),x)*epson+
1/2*diff(u10(x,y,c,i),x)*diff(v11(x,y,c,j),x)*epson+1/2*diff(v10(x,y,c,i),x)*diff(u11(x,y,c,j),x)*epson+
1/2*diff(u10(x,y,c,i),x)*diff(v10(x,y,c,j),y)/epson+1/2*diff(v10(x,y,c,i),x)*diff(u10(x,y,c,j),y)/epson+
1/2*diff(u11(x,y,c,i),y)*diff(v10(x,y,c,j),y)/epson+1/2*diff(v11(x,y,c,i),y)*diff(u10(x,y,c,j),y)/epson+
1/2*diff(u10(x,y,c,i),y)*diff(v11(x,y,c,j),y)/epson+1/2*diff(v10(x,y,c,i),y)*diff(u11(x,y,c,j),y)/epson+
1/2*diff(u10(x,y,c,i),y)*diff(v10(x,y,c,j),x)/epson+1/2*diff(v10(x,y,c,i),y)*diff(u10(x,y,c,j),x)/epson+
1/2*diff(u10(x,y,c,i),y)*diff(v10(x,y,c,j),y)/epson^2+
1/2*diff(v10(x,y,c,i),y)*diff(u10(x,y,c,j),y)/epson^2+
1/2*diff(u10(x,y,c,i),y)*diff(v11(x,y,c,j),x)+1/2*diff(v10(x,y,c,i),y)*diff(u11(x,y,c,j),x)+
1/2*diff(u10(x,y,c,i),x)*diff(v11(x,y,c,j),y)+1/2*diff(v10(x,y,c,i),x)*diff(u11(x,y,c,j),y)+
1/2*diff(u11(x,y,c,i),y)*diff(v10(x,y,c,j),x)+1/2*diff(v11(x,y,c,i),y)*diff(u10(x,y,c,j),x)+
1/2*diff(u10(x,y,c,i),x)*diff(v10(x,y,c,j),x)+1/2*diff(v10(x,y,c,i),x)*diff(u10(x,y,c,j),x)+
1/2*diff(u11(x,y,c,i),x)*diff(v10(x,y,c,j),y)+1/2*diff(v11(x,y,c,i),x)*diff(u10(x,y,c,j),y)+
1/2*diff(u11(x,y,c,i),y)*diff(v11(x,y,c,j),y)+1/2*diff(v11(x,y,c,i),y)*diff(u11(x,y,c,j),y):
##------------------------------------------------------
ekmsB2:=
1/2*diff(u11(x,y,c,k),x)*diff(u11(x,y,c,m),x)*epson^2+
1/2*diff(u11(x,y,c,k),y)*diff(u11(x,y,c,m),x)*epson+
1/2*diff(u11(x,y,c,k),x)*diff(u11(x,y,c,m),y)*epson+
1/2*diff(u11(x,y,c,k),x)*diff(u10(x,y,c,m),x)*epson+
1/2*diff(u10(x,y,c,k),x)*diff(u11(x,y,c,m),x)*epson+
1/2*diff(u10(x,y,c,k),x)*diff(u10(x,y,c,m),y)/epson+
1/2*diff(u11(x,y,c,k),y)*diff(u10(x,y,c,m),y)/epson+
1/2*diff(u10(x,y,c,k),y)*diff(u11(x,y,c,m),y)/epson+
1/2*diff(u10(x,y,c,k),y)*diff(u10(x,y,c,m),x)/epson+
1/2*diff(u10(x,y,c,k),y)*diff(u10(x,y,c,m),y)/epson^2+
1/2*diff(u10(x,y,c,k),y)*diff(u11(x,y,c,m),x)+1/2*diff(u10(x,y,c,k),x)*diff(u11(x,y,c,m),y)+
```

```
1/2*diff(u11(x,y,c,k),y)*diff(u10(x,y,c,m),x)+1/2*diff(u10(x,y,c,k),x)*diff(u10(x,y,c,m),x)+
1/2*diff(u11(x,y,c,k),x)*diff(u10(x,y,c,m),y)+1/2*diff(u11(x,y,c,k),y)*diff(u11(x,y,c,m),y):
ekmsB2_V:=
1/2*diff(u11(x,y,c,k),x)*diff(v11(x,y,c,m),x)*epson^2+
1/2*diff(v11(x,y,c,k),x)*diff(u11(x,y,c,m),x)*epson^2+
1/2*diff(u11(x,y,c,k),y)*diff(v11(x,y,c,m),x)*epson+
1/2*diff(v11(x,y,c,k),y)*diff(u11(x,y,c,m),x)*epson+
1/2*diff(u11(x,y,c,k),x)*diff(v11(x,y,c,m),y)*epson+
1/2*diff(v11(x,y,c,k),x)*diff(u11(x,y,c,m),y)*epson+
1/2*diff(u11(x,y,c,k),x)*diff(v10(x,y,c,m),x)*epson+
1/2*diff(v11(x,y,c,k),x)*diff(u10(x,y,c,m),x)*epson+
1/2*diff(u10(x,y,c,k),x)*diff(v11(x,y,c,m),x)*epson+
1/2*diff(v10(x,y,c,k),x)*diff(u11(x,y,c,m),x)*epson+
1/2*diff(u10(x,y,c,k),x)*diff(v10(x,y,c,m),y)/epson+
1/2*diff(v10(x,y,c,k),x)*diff(u10(x,y,c,m),y)/epson+
1/2*diff(u11(x,y,c,k),y)*diff(v10(x,y,c,m),y)/epson+
1/2*diff(v11(x,y,c,k),y)*diff(u10(x,y,c,m),y)/epson+
1/2*diff(u10(x,y,c,k),y)*diff(v11(x,y,c,m),y)/epson+
1/2*diff(v10(x,y,c,k),y)*diff(u11(x,y,c,m),y)/epson+
1/2*diff(u10(x,y,c,k),y)*diff(v10(x,y,c,m),x)/epson+
1/2*diff(v10(x,y,c,k),y)*diff(u10(x,y,c,m),x)/epson+
1/2*diff(u10(x,y,c,k),y)*diff(v10(x,y,c,m),y)/epson^2+
1/2*diff(v10(x,y,c,k),y)*diff(u10(x,y,c,m),y)/epson^2+
1/2*diff(u10(x,y,c,k),y)*diff(v11(x,y,c,m),x)+1/2*diff(v10(x,y,c,k),y)*diff(u11(x,y,c,m),x)+
1/2*diff(u10(x,y,c,k),x)*diff(v11(x,y,c,m),y)+1/2*diff(v10(x,y,c,k),x)*diff(u11(x,y,c,m),y)+
1/2*diff(u11(x,y,c,k),y)*diff(v10(x,y,c,m),x)+1/2*diff(v11(x,y,c,k),y)*diff(u10(x,y,c,m),x)+
1/2*diff(u10(x,y,c,k),x)*diff(v10(x,y,c,m),x)+1/2*diff(v10(x,y,c,k),x)*diff(u10(x,y,c,m),x)+
1/2*diff(u11(x,y,c,k),x)*diff(v10(x,y,c,m),y)+1/2*diff(v11(x,y,c,k),x)*diff(u10(x,y,c,m),y)+
1/2*diff(u11(x,y,c,k),y)*diff(v11(x,y,c,m),y)+1/2*diff(v11(x,y,c,k),y)*diff(u11(x,y,c,m),y):
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
integrFxy:=sort( expand(Eijkm* (eijsB0_sum*ekmsB2_V+eijsB0_sumV*ekmsB2+
                        eijsB1_sumV*ekmsB1_sum+eijsB1_sum*ekmsB1_sumV ) )  ,epslon):
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
ioption:=-2;
if (ioption = -3) then
##------------------------------------------------------ (33)
    Em3:=int(coeff(integrFxy,epson,-3),x);
    Em3v11:=expand(subs(v10(x,y,c,i)=0,v10(x,y,c,j)=0,v10(x,y,c,k)=0,v10(x,y,c,m)=0,
                v10(x,y,i,j)=0,v10(x,y,k,m)=0,Em3));
    Em3v10:=expand(subs(v11(x,y,c,i)=0,v11(x,y,c,j)=0,v11(x,y,c,k)=0,v11(x,y,c,m)=0,
                v10(x,y,i,j)=0,v10(x,y,k,m)=0,Em3));
elif (ioption = -2) then
##------------------------------------------------------ (34-37)
    Em2:=int(coeff(integrFxy,epson,-2),x);
    Em2v11:=expand(subs(v10(x,y,c,i)=0,v10(x,y,c,j)=0,v10(x,y,c,k)=0,v10(x,y,c,m)=0,
                v10(x,y,i,j)=0,v10(x,y,k,m)=0,Em2));
    Em2v10:=expand(subs(v11(x,y,c,i)=0,v11(x,y,c,j)=0,v11(x,y,c,k)=0,v11(x,y,c,m)=0,
                v11(x,y,i,j)=0,v11(x,y,k,m)=0,Em2));
elif (ioption = -1) then
##------------------------------------------------------ (38-40)
    Em1:=int(coeff(integrFxy,epson,-1),x):
```

```
    Em1v11:=expand(subs(v10(x,y,c,i)=0,v10(x,y,c,j)=0,v10(x,y,c,k)=0,v10(x,y,c,m)=0,
                v10(x,y,i,j)=0,v10(x,y,k,m)=0,Em1));
    Em1v10:=expand(subs(v11(x,y,c,i)=0,v11(x,y,c,j)=0,v11(x,y,c,k)=0,v11(x,y,c,m)=0,
                v11(x,y,i,j)=0,v11(x,y,k,m)=0,Em1));
elif (ioption = 0) then
 ##------------------------------------------------------ (41-48)
   Em0:=int(coeff(integrFxy,epson, 0),x):
    Em0v11:=expand(subs(v10(x,y,c,i)=0,v10(x,y,c,j)=0,v10(x,y,c,k)=0,v10(x,y,c,m)=0,
                v10(x,y,i,j)=0,v10(x,y,k,m)=0,Em0));
    Em0v10:=expand(subs(v11(x,y,c,i)=0,v11(x,y,c,j)=0,v11(x,y,c,k)=0,v11(x,y,c,m)=0,
                v11(x,y,i,j)=0,v11(x,y,k,m)=0,Em0));
elif (ioption = +1) then
##------------------------------------------------------
    Ep1:=int(coeff(integrFxy,epson,+1),x);
    Ep1v11:=expand(subs(v10(x,y,c,i)=0,v10(x,y,c,j)=0,v10(x,y,c,k)=0,v10(x,y,c,m)=0,
                v10(x,y,i,j)=0,v10(x,y,k,m)=0,Ep1));
    Ep1v10:=expand(subs(v11(x,y,c,i)=0,v11(x,y,c,j)=0,v11(x,y,c,k)=0,v11(x,y,c,m)=0,
                v11(x,y,i,j)=0,v11(x,y,k,m)=0,Ep1));
else
print(`ioption not valid`);
fi;
```

4. CONCLUSIONS

The symbolic computation presented here use an asymptotic technique to obtain a double scale formulation for the linearized elastic stability problem of structures built of perfectly periodic microstructures. Part of these developments were presented in [10], but only now the symbolic computations are presented. The symbolic computations allows the manipulation of the double scale strain-displacement relationship and perturbation equations are generated. Even if the possible extension to non-periodic modes of instability cannot be extracted from this formulation, but in the author's opinion the Floquet-Bloch theorem gives naturally support to such extension.

The use of symbolic computation is here very important as can be appreciated by following the deduction of the equations presented. The proposed algebraic manipulation allows that the intensive symbolic computation performed be handled with a reasonable effort and without mistakes that would take long time to detect. But from the very begin, it is clear that these computations are not get automatically, instead each step requires considerable interpretation in order to select the solutions that are of interest.

REFERENCES

[1] Sanchez-Palencia, E. Non-Homogeneous Media and Vibration Theory, Lecture Notes in Physics 127, Springer, Berlin, 1980.

[2] Geymonant, G., Muller, S., Triantafyllidis, N., "Quelques remarques sur

l'homogéneisation des matériaux élastiques nonlinéares", C.R. Acad. Sci., t. 311, Série 1, p. 911-916, 1990.

[3] N Triantafyllidis, S Bardenhagen, The influence of scale size on the stability of periodic solids and the role of associated higher order gradient continuum models, Journal of the Mechanics and Physics of Solids 44 (11), 1891-1928, 1996.

[4] Bendsϕe, M.P., Triantafyllidis, N, "*Scale Effects in the Optimal Design of a Microstructured medium against Buckling*", Int. J. Solids Structures, Vol. **26,** Nº7, pp725-741, 1990.

[5] Rizzi, NL; Tatone, A. "Symbolic manipulation in buckling and postbuckling analysis", *Computers & Structures* Vol. 21 (4), pp. 691-700. 1985.

[6] Neves, M.M., Guedes, J.M., and Rodrigues, H.C., - " *Generalized Topology Design of Structures with a Buckling Load Criteria*", Structural Optimization , 10, 71-78, Spring-Verlag 1995.

[7] Novozhilov ,V.V. 1953: Foundations on the Non-linear Theory of Elasticity. Graylock Press, Rochester, New York.

[8] Guedes, J.M.; Kikuchi, N.: Preprocessing and Postprocessing for Materials Based on the Homogenization Method with Adaptive Finite Elements Methods, Computer Methods in Applied Mechanics and Engineering 83, 143-198, 1990.

[9] MAPLE, [Online] https://www.gnu.org/software/octave/ [Accessed Jan 29, 2017]

[10] MM Neves, O Sigmund, MP Bendsøe, Topology optimization of periodic microstructures with a penalization of highly localized buckling modes, International journal for numerical methods in engineering 54 (6), 809-834

SYMCOMP 2017
Guimarães, 6-7 April 2017

INFLUENCE OF SEMIQUANTIFICATION IN DATSCAN^{TM} STUDIES FOR DIAGNOSIS OF PARKINSONIAN SYNDROMES

M. Queiroga[1,2], M. Elias[1,2], D. Silva[1,3], J. Serrano[4], J.I. Rayo[4], E. Carolino[1], L. Vieira[1], E, Sousa[1]

(1) Lisbon School of Health Technology, Polytechnic Institute of Lisbon, Lisbon, Portugal

(2) High Institute of Engineering of Lisbon, Polytechnic Institute of Lisbon, Lisbon, Portugal

(3) Nuclear Medicine Department, Diana Princess of wales Hospital, Grimsby, United Kingdom

(4) Nuclear Medicine Department, Infanta Cristina Hospital, Badajoz, Spain

e-mails: {queiroga.guigui@gmail.com, mariapirao11@gmail.com, dianasilva11_vsc@hotmail.com, etcarolino@estesl.ipl.pt, lina.vieira@estesl.ipl.pt, eva.sousa@estesl.ipl.pt}

Keywords: Key Words: DaTScan^{TM}, semi-quantification, reference values, accuracy, precision.

Abstract

Healthy controls reference values (RV) of DatScan^{TM} and its semi-quantification (SQ) adapted to Infanta Cristina's Hospital in Badajoz were created. It was used a sample of 120 DaTScan^{TM} tests divided into 4 groups with n = 30. The first one of healthy controls (GI) was used for calculation of RV. It was applied a semiautomatic method for segmentation and posterior calculi of the specific uptake ratios (SUR), respectively: Left and Right Caudate Nucleus/Occipital (A) and (B); Left and Right Putamen/Occipital (C) and (D); Striatum/Occipital (E); Left and Right Striatum/Occipital (F) and (G); Putamen/Caudate Nucleus (H). The second and third groups (healthy group (GIIH) and pathological group (GIIP)) were compared to GI for validation proposes. Control chards and ROC curves (ROCC) were obtained. The fourth group of randomized patients was used to evaluate the ability of SQ classification and it was compared with the physician evaluation. Sensibility (S), Specificity (SP), Positive (PPV) and Negative (NPV) predictive values were calculated. RV ($\bar{x} \pm \delta$): A (2.60±0.40); B (2.57±0.36); C (2.29±0.36); D (2.31±0.35); E (2.44±0.35); F (2,44±0,37); G (2,44±0,34); H (0.89±0.07). The GIIH values are above the RV and the GIIP group below the RV. The AUC, S and SP values for each SUR obtained by ROCC were: A (0.805; 0.765; 0.846); B (0.787; 0.711; 0.864); C (0.907; 0.853; 0.962); D (0.933; 0.930; 0.933); E (0.884; 0.871; 0.897); F (0.860; 0.800; 0.920); G (0.868; 0.844; 0.893); H (0.866; 0.732; 1.00). Were obtained results for PPV=0.466; NPV=1; S=1 and SP=0.65. The SQ combined with the visual analysis is a good method to detect healthy patients.

1. INTRODUCTION

Parkinsonian Syndromes (PS) are a group of neurodegenerative diseases characterized by the functional loss of neurons and dopaminergic terminations in the *substantia nigra.* Parkinson Disease (PD) belongs to this group and it is the second is the second most common movement disorder (it afflicts 1-2% of the population), being thereby of great importance the capability of distinguish between PS and other pathologies with similar symptoms but with different origin and physiopathology. [1, 2]

Single Photon Emission Computed Tomography (SPECT) with ^{123}I-FP-CIT – DaTScan^{TM} – is used for assessment by imaging of the *Striatum*, allowing the evaluation of the presynaptic dopaminergic system, in which Dopamine Transporters (DaT) are responsible for the release and reabsorption of dopamine in to the synapse, for this to be stored or degraded [2-5].

Therefor DaTScan^{TM} is used in differentiation between essential tremor and PS [1-6]. *Tolosa E. et al*, reported the sensibility of 77% and specificity of 100% in this differentiation, being aware that this diminution of DaT only occurs in PS this differentiation is important for the following therapeutic management [7, 8].

In DaTScan^{TM} exams with a healthy image standard imaging it is characterized by the visualization of portions of the *Striatum Body* (SB), with the shape similar to two symmetric comas. In pathologic exams there is a diminishing of the uptake in the SB, in the early stage of these diseases this starts to be noticeable in the Putamen and in later stages this starts to occur also in the caudate. This uptake changes can be symmetric or asymmetric [1, 7].

In the clinic work is usual the evaluation of these exams for diagnostic purpose being performed only by visual assessment, although this evaluation is subjective, complex and dependent of the operator experience, it is also influenced by high levels of variability intra and operator [7-9].

In order to improve the diagnosis process, several methods of SQ of the ^{123}I-FP-CIT uptake in the interest brain structures have being developed [8, 10]. To apply these methods in DaTScan^{TM} studies, there are drawn Region of Interest (RoI) and ratios between different regions: Specific uptake ratios (SUR) are calculated between regions of specific uptake (in the Striatum) and other regions of nonspecific uptake that correspond to the background, and are regions with a low density of DaT (for example in the occipital region) and symmetry ratios (SR) between regions of specific uptake [5-8, 11]. By using the different uptake values of the segmented regions the rations are calculated, for the SUR is expressed in the equation (1).

$$SUR = \frac{\bar{x}\ \text{Striatum counts} - \bar{x}\ \text{background counts}}{\bar{x}\ \text{background counts}} \quad (1)$$

The segmentation process in this process can be by manual, automatic or semiautomatic methods [11]. The manual method is the most currently used in the clinical practice although the automatic method being the most reproducible [10, 11].

In the segmentation process by with semiautomatic RoI, these are automatically

generated, and after they are manually adjusted, diminishing the variability intra and inter operator when compared with the fully manual method [9-11].

The reproducibility of the SQ increases when it values are compared with the Reference Uptake Ratios of the same age class. The low utilization of the SQ in the clinical practice, is due to the necessity of it values to be adjusted to the population where it is being applied, and also to the exam protocol of each Nuclear Medicine Department. There are also many factors that can create bias in the process of implementation of the Reference Values (RV), and is also a process that takes much time [5, 7-12].

The purpose of the study is the creation and validation of RV obtained by recurring to SQ of DaTScan^{TM} studies, as complement to the visual assessment, in order to increase the accuracy diagnosis of the PS in DaTScan^{TM} exams.

2. METHODOLOGY

The present study was developed in the three stages following presented: creation of a Data Base with healthy controls for the determination of the RV of each Uptake Ratio (UR); the validation of the obtained RV by comparison with studies between healthy controls with ratios obtained in patients with PS; and the third stage the assessment of the influence of the SQ of the DaTScan^{TM}, through the comparison of the results of diagnostic classification by applying the RV with the previous diagnosis performed by an clinic trained physician, without recurring to SQ.

2.1. Sample Selection

A sample of 120 DaTScan^{TM} studies was collected, the subjects of the sample had ages between 60 and 75 years, and were people with indication for undergoing DaTScan^{TM} exam, and were in the database of the Xeleris^{TM} workstation in Infanta Cristina's Hospital, in Badajoz.

The first 90 studies were selected by convenience sampling, and divided into three groups with 30 elements; this separation was made by the classification of the exams clinical reports by Nuclear Medicine and Neurologist Physicians: Group I (GI) – Healthy control group, this exams had a normal uptake imaging, and this group was used for obtention of the RV; Group IIP (GIIP) – Pathologic group that presented a pathologic uptake imaging, suggestive of PS and Group IIS (GIIH) – healthy group, these exams were of subjects with normal uptake. Both GIIP and GIIH were used in the validation of the obtained RV.

In the last phase of the study was used random sampling for the selection of 30 DaTScan^{TM} studies of the same age range than the ones of the created RV, also from studies of the department workstation database. This sample was the Group III (GIII), it was guaranteed that this group didn't included repeated studies form the previously created groups.

2.2. Exams protocol

All studies were acquired by *EANM guidelines* suggested protocol, and with the same acquisition conditions. For the acquisition of the exams the patients were in supine position and had the head immobilized with a head support, to restrict movement. The exams were acquired 4 hours after the Intravenous administration of 185 MBq of ^{123}I-FP-CIT.
All the SPECT Studies performed using parallel-hole, low energy, and high resolution collimator; the rotation radius was minor than 15 cm, matrix of 128x128 pixels, and zoom of 1.25, circular acquisition orbits of 360 º in mode *step and shoot*, the Energy peak of 159 keV±10%, 64 projections with 30 seconds each [5,10].
For the image reconstruction and processing of the *DaTScan™ studies of GI, GIIP and GIIH*, each study was processed three times in the workstation *Xeleris™*, by the same operator, to eliminate any interference of interoperator variability. All data were reconstructed by Filtered Back Projection (FBP), it was used Butterworth filter (cutoff frequency of 0.4 cycles per pixel and power was 10), the ramp filter applied was quantitative, the color map applied was GE Color.[10]
It was created a summed image of three transaxial slices that comprehend the totality of the Striatum tissue. It was determined the average counts of the Left and Right Caudate (LC and RC), left and right putamen (LP and RP) and of the occipital (OC) regions respectively. This was achieved by using the semi-automated RoIs presented in the figure 1[5, 10].

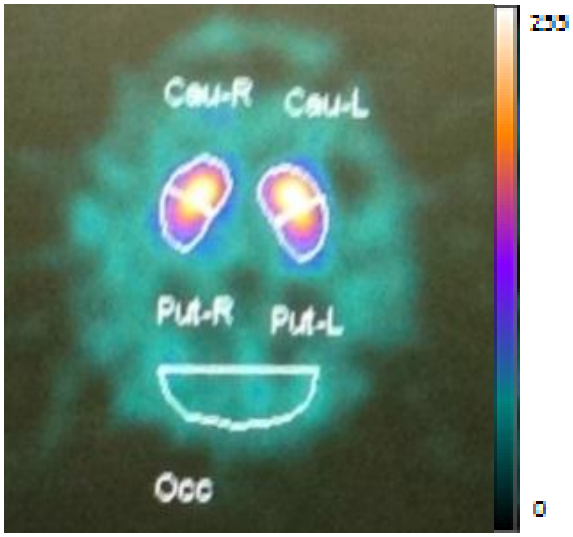


Figure 1: Summed image of three transaxial slices containing the striatum, and the semiautomatic used RoIs in LC, RC, LP, RP and OC.

2.3. Reference Values Uptake Ratios

After the assessment and exclusion of outliers trough the dispersion diagram, it was calculated the average ± standard deviation ($\bar{x} \pm \delta$) of the UR (both the SUR and symmetry ratios (SR)) to obtain the RV of the healthy controls of Infanta Cristina's Hospital.

There were calculated RV for each of the UR of each study, presented in the table 1.

Table 1: Expressions for ratios calculation for SQ of the in the *DatScan™* studies, and code of the expression.

Uptake Ratio		Short form	Code
$\frac{\textbf{Caudate}}{\textbf{Occipital}}$	$\frac{Left\ Caudate}{Occipital}$	$\frac{LC}{Oc}$	A
	$\frac{Right\ Caudate}{Occipital}$	$\frac{RC}{Oc}$	B
$\frac{\textbf{Putamen}}{\textbf{Occipital}}$	$\frac{Left\ Putamen}{Occipital}$	$\frac{LP}{Oc}$	C
	$\frac{Right\ Putamen}{Occipital}$	$\frac{RP}{Oc}$	D
$\frac{\textbf{Striatum}}{\textbf{Occipital}}$	$\frac{Right\ Putamen + Left\ Putamen + Right\ Caudate + left\ Caudate}{\frac{4}{Occipital}}$	$\frac{\frac{RP + LP + RC + LC}{4}}{Oc}$	E
$\frac{\textbf{Left Striatum}}{\textbf{Occipital}}$	$\frac{\frac{Left\ Putamen\ + Left\ Caudate}{2}}{Occipital}$	$\frac{\frac{LP + LC}{2}}{Oc}$	F
$\frac{\textbf{Right Striatum}}{\textbf{Occipital}}$	$\frac{\frac{Right\ Putamen\ + Right\ Caudate}{2}}{Occipital}$	$\frac{\frac{RP + RC}{2}}{Oc}$	G
$\frac{\textbf{Putamen}}{\textbf{Caudate}}$	$\frac{Right\ Putamen\ + Left\ Putamen}{Right\ Caudate + Left\ Caudate}$	$\frac{RP + LP}{RC + LC}$	H

2.4. Validation of the Reference values

To validation of the RV, after assessment of the normal distribution of the samples GIIH and GIIP were created control charts, with the distribution of the values of the ratios with the codes name from A to H presented in the table 1, using the RV as standard [13, 14].

The control charts were analyzed by a different, method, for more restrict classification, it was considered healthy all values which the SUR were higher than the lower value of standard deviation, and always that in the symmetry UR, this was preserved and fitted the limit ($\bar{x} \pm \delta$).

This analysis considers what from a clinic point of view makes sense in a way of increasing the rigor of the established RV. From the analysis of the control charts, to assess the sensibility and specificity of the study, were created Receiver Operating Characteristics (ROC) curves [15, 16].

2.5. Assessment of the influence of the SQ in the classification of *DaTScan*^{TM} studies

After calculation of the UR of the GIII, and by comparison of the UR from A to H, obtained with the created RV, each study was classified as pathologic or non-pathologic. This classification was after compared with the diagnosis that the Nuclear Medicine Physician (realized only with visual assessment of the *DaTScan*^{TM} studies) and with the results of the neurologic reports with the evolution of the last five years, always that this information was available. From this comparison were calculated the sensibility, specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) [14, 17, 18].

3. RESULTS

The RV created from the SQ of the Healthy control group, are presented as ($\bar{\boldsymbol{x}} \pm \boldsymbol{\delta}$), in the table 2. The UR A (2.60 ± 0.40) and B (2.57±0.36) are about the caudate dopaminergic uptake, the UR C (2.29± 0.36) and D (2.31±0.35) of the putamen right and left respectively. The UR E is the one that measures the totality of all the Striatum and it value is (2.44±0.35). The UR F (2.44±0.37), G (2.44±0.34) and H (0.89±0.07) are UR of symmetry.

Table 2: Reference Values ($\bar{\boldsymbol{x}} \pm \boldsymbol{\delta}$) obtained for the SUR A-H through the SQ of GI.

UR	Reference Value ($\bar{\mathbf{x}} \pm \boldsymbol{\delta}$)
A	2.60 ± 0.40
B	2.57±0.36
C	2.29± 0.36
D	2.31±0.35
E	2.44±0.35
F	2.44±0.37
G	2.44±0.34
H	0.89±0.07

In the control charts of all the Uptake rations calculated, obtained for validation of the RV previously calculated, only the UR C of the GIIP did not presented one normal distribution (p=0,004). The control charts allow the global visualization of the distribution of the UR of the subjects in the groups GIIH and GIIP, when compared with the established RV.

The limit ($\bar{x} - \delta$) of each RV of the UR were considered the value that allow to distinguish a pathologic *DaTScan*^{TM} from a healthy one. The UR from A to H corresponding to the GIIS places above the RV ($\bar{x} - \delta$)and in the GIIP it occurred the opposite. In the UR A and B that correspond to the caudate uptake is the UR that presents the bigger quantity of subjects of the GIIP above the average of the RV.

The UR H that corresponds to a Symmetry UR is one that also presents many subjects of the GIIP above the average, nerveless it does not present any subject of the GIIS below that limit. The UR D is the one that presents better concordance with the theoretical assumed knowledge.

Through the analysis of the control charts, was verified that 93% of the patients of the GIIH were above the limit (x - δ) of the UR and 80% of the patients that belong to the GIIP were below the same limit, for corresponding to clinical pathologic patients.

The results of the Area Under the Curve (AUC) obtained by the ROC curves of standard error, sensibility and specificity of the UR from A to H is in the figure 2. The AUC range is between 0 and 1. In this study the AUC values are comprehended between 0.787 UR B and 0.933 UR D, and the standard error range from 0.037 UR D and 0.062 UR B. The UR A is the one that presents lower value is of 0.84. The UR D of 0.930 is the one that present higher value of specificity, being the UR B 0.711, the one which presents the lower value.

Table 3: Results AUC, Standard Error, Specificity and Sensibility of the curves ROC of the UR A to H.

	A	B	C	D	E	F	G	H
AUC	0.805	0.787	0.907	0.933	0.884	0.860	0.868	0.866
Standard Error	0.059	0.062	0.042	0.037	0.048	0.051	0.051	0.046
Specificity	0.765	0.711	0.853	0.930	0.871	0.800	0.844	0.732
Sensibility	0.846	0.864	0.962	0.933	0.897	0.920	0.893	1.000

In the tables 4 and 5 are presented the concordance values between the medical report and the diagnosis performed having SQ for basis, with the sensibility values (1.000); specificity (0.652), PPV (0.467) and NPV (1.000) of the RV of the UR A to H. The sensibility is elevated, and consequently an NPV also high. On the other side the specificity is not so elevated, to which 5 corresponds a lower PPV. The elevated NPV indicates that all the subjects classified as healthy (n=15), having by support the RV of the UR from A to H,

correspond to healthy subjects also classified as healthy by the medical reports that are being used as reference [19, 20].

The lower PPV indicates that from the 15 that our classifier classified as pathologic only 7 were classified as pathologic by the physician's reports.

Table 4: Sensibility, Specificity, PPV and NPV of the RV of the UR C to H, compared to physician's diagnosis by Visual assessment.

Sensibility	1,00
Specificity	0,652
PPV	0,467
NPV	1,00

Table 5: Classification of positive or negative to the presence of PS, by comparing classification by the RV limit implementation and the visual assessment by the physician.

	Pathologic	Yes	No	Total
Test	Positive	7	8	15
	Negative	0	15	15
Total		7	23	30

4. DISCUSSION

About the RV calculated for the GI, the UR from A-B, C-D and F-G are similar, as they correspond to the homologs symmetric structure of the striatum. The UR F and G present an average value between the caudate and the left and right putamen respectively.

The UR C and D correspond to putamen uptake. This is the first structure to suffer degeneration in pathologic cases [5, 10]. The present study corroborates this.

The AUC values presented in the table 3 more elevated are relatives to the UR C (0.907) and D (0.933), corresponding also to the lower values of the obtained standard error. Through the SQ analysis of these two UR exist 90.7% and 93.3% probability, respectively; of a pathologic study present a lower uptake value than a healthy study [15, 16, 21-23]. This result needs further investigation, for understanding the reason for the asymmetry, or if it is a random probabilistic value. Although this value these two UR are considered an excellent tool for separating the pathologic form the non-pathologic subjects [15, 21-23]. These UR have also very high values of specificity and sensibility.

The UR E, F, G and H present as AUC values 0.884, 0.860, 0.868 and 0.866, respectively; being also useful for separation of the groups [16, 21-23].

In this group the UR E is the best indicator, because it accounts for total of the specific uptake, being the summed value of the UR F and G.
The UR G presents a AUC value slightly higher to the UR F, for the same reason of the difference between UR C and D.
The UR H is the one with better value of sensibility (1.000), being an excellent parameter for detecting pathologic subjects [16].
The UR A and B account the uptake of the left and right caudate, the caudate is the last structure of the *Striatum* to suffer diminishing of the uptake [5, 10]. Thereby these are the UR with lower value of AUC of 0.805 and 0.787, respectively.
In the last stage of the study only the UR C to H were taken in to consideration, for the subject's classification, for being the best discriminating ratios between pathologic and healthy. It was possible to verify the high NPV and consequently the high sensibility (1.000) of the created RF, as is presented in the tables 4 and 5. These results confirm the high capability of the SQ for detecting pathology between the subjects that are positive to the presence of the disease [16, 21]. The concordance level between the two techniques was of 100%, in this detection.
The specificity is lower, the proportion of truly healthy subjects between the ones that were considered non pathologic is lower [16], in this study it were considered 8 subjects as pathologic, that the reference clinical reports classified as healthy. These results can be explained by the SQ being able of detect pathologic changes prior to the visual assessment, but further investigations are required, with the following of the patients and extern validation by other techniques, to confirm this hypothesis. The majority of patients had no information with a 5 years follow up; this could have provided more certainty about the evolution of the patient, to confirm which classification was more accurate. The method with only one limit below the standard deviation that was used could also influenced the results and have influence in this parameters.

5. CONCLUSION

From this study we can conclude that the SQ, with the RV established for Infanta Cristina's Hospital in the *DaTScan*^{TM} studies, is precise and accurate in the classification of subjects with suspected PS, being able of use for complementing visual assessment.

In the future it would be interesting to confirm is the differences between SQ classification and the visual assessment classification could be due to an earlier detection by the SQ technique.

ACKNOWLEDGEMENTS
The authors wish to acknowledge the financial support of Project IPL, IDI &CA/SOFTIMOB.

REFERENCES

[1] A. Roussakis, P. Piccini, M. Politis, "Clinical utility of DaTSCAN (123I-Ioflupane injection) in the diagnosis of Parkinsonian Syndromes", *Degener Neurol Neuromuscul* Dis, Vol. (3) pp. 33-39, 2013.

[2] C. C. Umeh, Z. Szabo, G. M. Pontone et al., "Dopamine Transporter Imaging in Psychogenic Parkinsonism and Neurodegenerative Parkinsonism with Psychogenic Overlay : A Report of Three Cases". *Tremor Other Hyperkinet Mov*, Vol. (3) pp. 39-42, 2013.

[3] R. Marrow, "DaTscan ABSORBED DOSE PER UNIT ADMINISTERED Pancreas". pp. 8-11, 2011.

[4] N. Matsumoto, T. Hanakawa, S. Maki, et al., "Nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner". *J Neurophysiol*, pp. 978-998, 1997.

[5] J. Darcourt, J. Booij, K. Tatsch, et al., "EANM procedure guidelines for brain neurotransmission SPECT using 123I-labelled dopamine transporter ligands, version 2". *Eur J Nucl Med Mol Imaging*, pp.443-450, 2010.

[6] J. E. Ferri-de-Barros, "Doença de Parkinson". *Rev Bras Med*, Vol (69(11)) pp.113-118, 2012.

[7] L. Bolt, S. Hoffmann, P. Kemp, et al. "Quantification of [123I] FP -CIT SPECT brain Images: an accurate technique for measurement of the specific binding ratio". *Eur J Nucl Mol Im*, Vol. (33) pp.1491-1499, 2006.

[8] E. Tolosa, T. Vander Borght, E. Moreno, "Accuracy of DaTSCAN (123I-Ioflupane) SPECT in diagnosis of patients with clinically uncertain parkinsonism: 2-Year follow-up of an open-label study". *Mov Disord*, Vol. (22) pp.2346-2351, 2007.

[9] K. Badiavas, E. Molyvda, I. Iakovou, et al., "SPECT imaging evaluation in movement disorders: Far beyond visual assessment", *Eur J Nucl Med Mol Imaging*, Vol. (38) pp. 764-773, 2011.

[10] D. S. W. Djang, M. J. R. Janssen, N. Bohnen, et al., "SNM Practice Guideline for Dopamine Transporter Imaging with 123I-Ioflupane SPECT 1.0.", *J Nucl Med*. Vol.(53(1)) pp.154-163. doi:10.2967/jnumed.111.100784. 2012.

[11] D. A. B. Faria, "Segmentação, Reconstrução e Quantificação 3D de Estruturas em Imagens Médicas – Aplicação em Imagem Funcional e Metabólica", pp.76, 2010.

[12] M. Driessnack, V. D. Sousa, I. A. C. Mendes, "An Overview of Research Designs Relevant to Nursing: part 2: qualitative research designs", *Rev Lat Am Enfermagem*,Vol. (15(3)) pp.684-688, 2007.

[13] B. F. E. Lins et al., "Ferramentas de Qualidade", *Ci. Inf., Brasília*, Vol. (22(2)) pp. 153-161, 1993.

[14] J. Luis, D. Ribeiro, C. Schwengber, *Série Monográfica Qualidade Controle Estatístico Do Processo*, 2012.

[15] A.C. B. Curvas ROC, *Aspectos Funcionais E Aplicacões*. 2000.

[16] E. Z. Martinez , F. Louzada-Neto, B. D. B. Pereira, 2 A Curva ROC para Testes Diagnósticos", *Cad Saúde Coletiva*, Vol. (11) pp.7-31, 2003.

[17] S. Maria, B. Barbosa, G. C. V. Costa, C. Franco, *Probabilidade e Estatística*. Belo Horiz PUC Minas Virtual. 2003.

[18] E. Reis, I. Reis, *Avaliação de Testes Diagnósticos.* Univ Fed Minas Gerais Inst Ciências Exatas Dep Estatística. 2002.

[19] Hart M. K. H. R., *Statistical Process Control*. Statit Software, Inc. 2007.

[20] Kendall M. *Statistical Quality Control*.Vol (1947) pp. 454-454.

[21] R. Kumar, A. Indrayan , "Receiver operating characteristic (ROC) curve for medical research", *Indian Pediatr*, Vol (48) pp. 277-287, 2011;

[22] T. Fawcett, "An introduction to ROC analysis". *Pattern Recognit Lett*. Vol (27) pp. 861-874, 2006.

[23] M. Greiner, D. Pfeiffer, R. D. Smith, "Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests". *Preventive Veterinary Medicine,* Vol. (45) pp. 23-41.

SYMCOMP 2017
Guimarães, 6-7 April 2017

INFLUENCE OF COMPUTED TOMOGRAPHY ATTENUATION CORRECTION IN MYOCARDIAL PERFUSION IMAGING, IN OBESE PATIENTS: CLASSIFICATION BY SEX AND BODY MASS INDEX

O. Stakhiv[1, 2], I. Melo[1,3], M. Clarke[4], M. Aplin[4], N. Singh[4], K. Day[4], S. Dizdarevic[4], M. Jessop[4], P. Begley[4], E. Carolino[1], L. Vieira[1], E. Sousa[1]

(1) Lisbon School of Health Technology, Lisbon, Portugal

(2) United Lincolnshire Hospitals, NHS Trust, Lincoln, UK

(3) Nuclear Medicine Department, Barts Health, NHS Trust, London, UK

(4) Nuclear Medicine, Department of Imaging, Brighton and Sussex University Hospitals, NHS Trust, Brighton, UK.

e-mails: {{oleh.stakhiv@ulh.nhs.uk; Ines.Costa@bartshealth.nhs.uk; mark.aplin@bsuh.nhs.uk; nitasha.singh@bsuh.nhs.uk; katherine.day@bsuh.nhs.uk; sabina.dizdarevic@bsuh.nhs.uk; maryam.jessop@bsuh.nhs.uk; patrik.begley@bsuh.nhs.uk; etcarolino@estesl.ipl.pt; lina.vieira@estesl.ipl.pt; eva.sousa@estesl.ipl.pt}

Keywords: Myocardial perfusion imaging; Computed tomography; Attenuation correction, Body Mass Index.

Abstract

Four groups of 71 subjects 47 with body mass index (BMI) between 30 and 35 (27 Male (M1) and 20 Female (F1)) and 24 with BMI above 35 (13 Male (M2) and 11 Female (F2)) who underwent stress-rest , SPECT MPI of a two day protocol, by EANM guideline protocol, with and without the incorporation of the Attenuation correction by computed tomography (CT-AC), for stress and rest separately. For perfusion percentage quantifications, the 5 walls model of left ventricle (LV) was used: anterior (ANT), lateral (LAT), inferior (INF), septal (SEP) and apical (API), using the QGS/QPS^{TM} software. For statistical evaluation it was used the Friedman test. Statistically significant differences were found in comparison of studies with and without attenuation correction (AC) for: F1 in stress and rest studies respectively for LAT (p=0.006 and p=0.034), INF (p=0.000 and p=0.000) and in rest study for SEP (p=0.044) LV walls; F2 group of stress and rest studies respectively for INF (p=0.001 and p=0.008) walls; M1 group of stress and rest study respectively for LAT (p=0.000 and p=0.000), INF (p=0.000 and p=0.000) SEP (p=0.003 and p=0.001) walls and just in rest study for API (p=0.045) LV walls; M2 group of stress and rest studies respectively for INF (p=0.000 and p=0.000), LAT (p=0.020 and p=0.014) and in stress study for SEP (p=0.003) LV walls. The influence of CT-AC is bigger within the groups with BMI between 30 and 35.

1. INTRODUCTION

Myocardial perfusion imaging (MPI) acquired by single photon emission computed tomography (SPECT) is a well-established, non-invasive, technique that is used for the evaluation of known or suspected Coronary artery disease (CAD) [1–3]. Nevertheless, due to the attenuation artifacts (AA), the specificity of conventional SPECT MPI has remained suboptimal [1,4]. The presence of abnormal findings caused by breast, diaphragmatic and thoracic wall attenuation results in marked regional variations in myocardial activity that are not related to myocardial perfusion defects [5,4] . These variations are most often found in patients suffering from obesity which tend to have a larger volume of soft tissue and organs that attenuate gamma rays from the heart [6]. In these cases, the number of false positive results can increase, hindering a correct diagnosis.

Some of the AA are different between male and female population as they can be dependent on the quantity of attenuating material adjacent to the heart [7]. In male population it is more common to find attenuation artifacts in the inferior wall of the left ventricle, owing to sub-diaphragmatic attenuation, but in female population the artifacts appear more often in the anterior and sometimes in the lateral walls due to breast attenuation [5].

To overcome this problem, some guidelines recommend the incorporation of attenuation correction (AC) to improve diagnostic accuracy. AC methods measure the tissue interference in the photon (PT) attenuation in order to minimize its impact in the images [8]. They attempt to determinate a real tracer distribution in the tissues and the fraction of the attenuated PT. Physically the attenuation process depends on the thickness and tissue type and the intensity of the incident PT, as shown in the following equation:

$$I = I_0 e \sum_i -x_i \mu_i \quad (1)$$

where I represents the measured ray and I_0 the initial ray intensities, and the index i represents all the different tissue type regions along the trajectory, μ_i are the effective attenuation coefficients for the different tissue regions and x_i are the different thicknesses of the regions included in the study, thus the sum represents the total attenuation through all the regions.

Several AC systems were developed through the years, and the AC by computed tomography (CT-AC) is the one that was used in this study. It has been shown that this system improves the image quality and the diagnostic accuracy and it has several advantages over the others AC methods such as high PT flux, no decay of transmission source and short scans times [9,10]. CT-AC methods require the determination of an attenuation map (AM), which represents the spatial distribution of linear attenuation coefficients (ACo) for patients anatomy regions that are included in the MPI scans [8,11]. These describe the fraction of a beam of energy that is absorbed or scattered per unit of thickness of absorber, of different tissues for each individual patient, therefore the construction of AM is performed in order to correct the attenuated PT, which usually results in an increase of the density photon counts (PC) in areas affected by the attenuation effects (AE).

It is important to acquire a good quality CT AM so as to obtain a high quality myocardial

images (Figure 1), especially when scanning individuals with high BMI which tent to attenuate a significant number of gamma photons coming from the heart. In some cases CT-AC generates differences in quantification perfusion values (Figure 2) of the left ventricle (LV) when compared with non-attenuation corrected (NAC) MPI results [11].

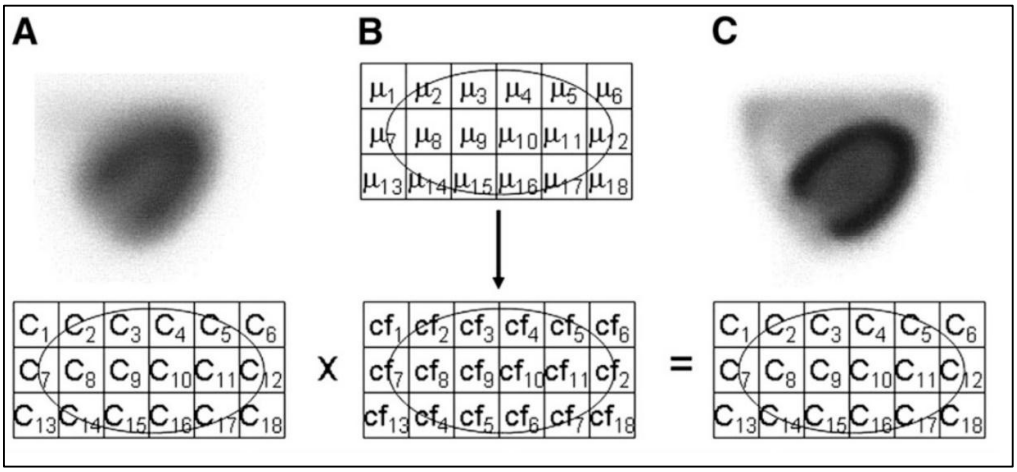


Figure 1: The AC factors (B) are obtained from ACo measurements determined by CT scan and used to correct emission data from uncorrected SPECT MPI scan (A) to generate AC myocardial images (C) with better spatial resolution which results in an improvement of the image quality. *Adapted from [8]*.

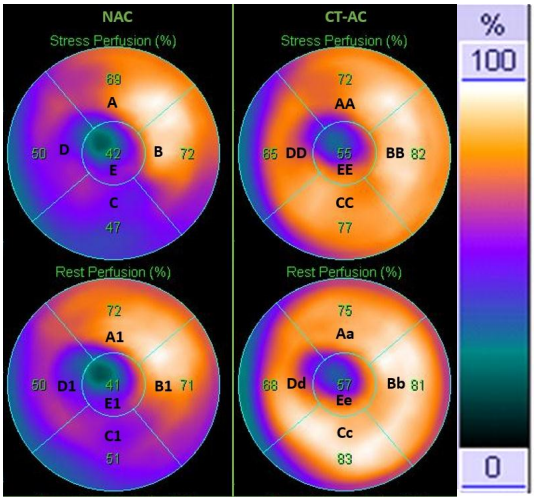


Figure 2: By analysing NAC polar maps is obtained a reduced perfusion rate in the INF (C and C1) and SEP (D and D1) LV walls, both for stress and rest respectively. CT systems correct SPECT data and as a result the perfusion rate increases in the affected areas, INF (CC and Cc) and SEP (DD and Dd) LV walls for stress and rest respectively in the CT-AC polar maps.

The assessment of MPI is performed by either qualitative or quantitative analyses and the quantification of the perfusion rate is a very important tool for the detection of CAD [12]. The most common software used for this process is Quantitative Gated SPECT/Quantitative Perfusion SPECT (QGS/QPS^{TM}) which provides automatic segmentation, quantification, analysis and display of static and gated MPI data [13,14].
Genovesi et al. study suggests that the use of AC should be limited to male patients with a BMI higher than 27, but there is no study performed for male and female individuals with the BMI superior than 30, which is considered as obese group by the Centers for Disease Control and Prevention [15].
The aim of this study is to verify the influence in MPI results between CT-AC and NAC data, in patients with BMI between 30 and 35 and higher than 35 for male and female population.

2. METHODOLOGY

2.1. Study population

Retrospective evaluation of MPI data since January 2014 until January 2015 of 71 subjects (40 male and 31 female) who underwent stress-rest, SPECT MPI of a two days protocol, with ^{99m}Tc-tetrafosmin, for suspected CAD. The selected subjects had a body mass index (BMI) above 30, and they were divided into four groups: BMI between 30 and 35 (27 male (M1) and 20 female (F1)) and BMI above 35 (13 male (M2) and 11 female (F2)).

2.2. Stress Testing

All subjects underwent pharmacological stress test with the intravenous radiotracer injection 10-20 s after regadenoson administration. During the stress induction the vital signs were monitored with an ECG signal with 12 leafs and the blood pressure was measured at the beginning and at the end of the stress testing.

2.3. SPECT data acquisition

All patients had 2 day imaging protocol, and the images were acquired 45 minutes after ^{99m}Tc-tetrafosmin intravenous injection. The injected activity varied with the patient's weight: 400 MBq were used for patients with < 90 kg, 600 MBq for patients between 90 and 110 kg and 800MBq for patients weighing more than 110 kg, as it is recommended by the Administration of Radioactive Substances Advisory Committee guidance notes.
All image acquisitions were performed using a hybrid SPECT/CT dual-head gamma camera (Infinia Hawkeye 4; GE Medical Systems). Emission data were acquired using parallel-hole, low energy, and high resolution collimator with the patients in supine position. Body contour was used for the acquisition orbits, over 180 degree arcs starting on the right anterior oblique and ending in the left posterior oblique projections, with use of 60 stops each of 3 degree. For each stop, the emission data were acquired for 25 seconds. The matrix used for the image

acquisition was 64x64 pixel. All images were acquired on the 140 keV photopeak with a 20% symmetrical window.

2.4. CT attenuation maps acquisition

The SPECT images were followed by CT examination, with acquisition parameters of 120 kV, 2.5 mA, pitch of 1.9, rotation speed of 3 s per rotation, 512x512 pixel matrix and slice thickness of 5mm.
After the CT acquisition, the obtained images were transformed into SPECT attenuation maps by the following steps: first, the CT images were transformed into a volume with the same voxel and pixel size as SPECT images, in order to perform the SPECT and CT image fusion [8]. In the second step, CT Hounsfield values were transformed into linear attenuation coefficients, with the same energy as the SPECT emission photons (140 keV) [16]. Next, the obtained attenuation maps were smoothed in all directions with a Gaussian filter to adjust the resolution to that of the SPECT images [8]. Finally, the CT-based attenuation coefficients were included to reconstruct AC SPECT data.

2.5. Acquired data processing

All studies were uniformly processed, by the same operator, with commercially available QGS/QPS™ software on a Xeleris 2.0 workstation. Both NAC and CT-AC SPECT emission images data were reconstructed by use of OSEM/MLEM™ software. Thereby, tomographic slices were generated and displayed as vertical long-axis, horizontal long-axis and short-axis slices. SPECT emission data were co-registered and fused with CT data. The co-registration of the images was semi-automatic, therefore the fusion images were visually inspected to ensure the correctness of the image alignment. The predetermined attenuation coefficients were applied to SPECT emission data to perform the CT-AC.

2.6. Quantitative Image Analysis

For further quantitative analysis it was performed the segmentation of the LV into 5 walls: anterior (ANT), lateral (LAT), inferior (INF), septal (SEP) and apical (API), as shown in the Figure 3, in order to quantify the perfusion percentage of all the walls independently. Stress and rest results were registered separately. This process was carried out for each individual subject of the study.

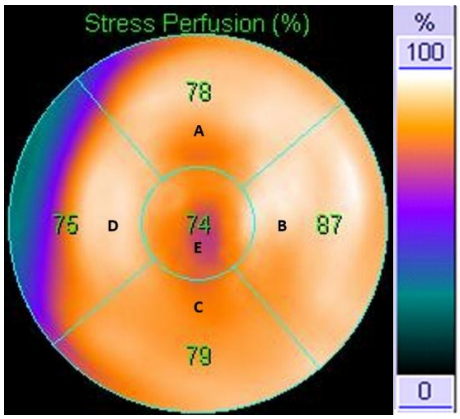


Figure 3: Segmentation of the LV into 5 walls, anterior (A), lateral (B, inferior (C), septal (D) and apical (E), with subsequent quantification of the Stress Perfusion percentage for each LV wall, performed by QGS/QPS™ software.

2.7. Statistical Analysis

All statistical analysis were executed using Statistical Package for the Social Sciences (SPSS) v.22, software. In order to execute the statistical evaluation it was used Friedman test for non-parametrical pared samples (CT-AC and NAC data), as the data did not followed a normal distribution [17]. For all the groups included in the study, a significance level of 5% was used.

3. RESULTS

Quantification of the perfusion rate of LV using CT-AC and NAC data

The findings of AC and NAC quantitative assessment, of the LV perfusion rate, for female and male groups, are shown in Table 1 and 2 respectively. As a result, CT-AC systems introduce some statistical significant differences in the perfusion quantification when applied to the NAC MPI.

For F1 group statistically significant differences between AC and NAC results were found in the LAT and INF walls respectively for the Rest (p=0.034; p=0.000) and Stress (p=0.006; p=0.000) studies and in SEP walls just for Rest (p=0.044) studies as shown in Table 1. For F2 group no significant differences were obtained except in the INF LV walls for Rest (p=0.008) and for Stress (p=0.001) as represented in Table 1.

Table 1: Results of the *p-values* obtained from the perfusion rate comparisons of the LV walls, between AC and NAC SPECT MPI data, on the female study population with BMI between 30 and 35 (F1 group) and above 35 (F2 group).

WOMEN	$30 \leq BMI \leq 35$		BMI > 35	
	AC – NAC LV Walls	p - value	AC – NAC LV Walls	p - value
STRESS	ANT	0.240	ANT	0.647
	LAT	0.006*	LAT	0.073
	INF	0.000*	INF	0.001*
	SEP	0.076	SEP	0.944
	API	0.240	API	0.504
REST	ANT	0.159	ANT	0.622
	LAT	0.034*	LAT	0.181
	INF	0.000*	INF	0.008*
	SEP	0.044*	SEP	0.833
	API	0.192	API	0.218
* statistical significant differences found between CT-AC and NAC perfusion percentage results				

For male exams, in the M1 group statistical significant changes in LV perfusion between CT-AC and NAC were observed on the LAT (p=0.000; p=0.000), INF (p=0.000; p=0.000) and SEP (p =0.001; p=0.003) walls, for Rest and Stress subsequently and on the API walls just for Rest (p=0.045) as presented in Table 2. The same occurrence can be detected in the M2

group, on the same anatomical regions, except the API walls areas. Nevertheless, significant variations among AC and NAC data on the SEP walls are only evident in the Stress study (p=0.003). The remaining differences are present on both, Rest (p=0.014, for LAT walls; p=0.000, for INF walls) and Stress (p=0.020, for LAT walls; p=0.000 for INF walls) studies as verified in Table 2.

Table 2: Results of the *p-values*, obtained from the perfusion rate comparisons of the LV walls, between AC and NAC SPECT MPI data, on the male study population with BMI between 30 and 35 (M1 group) and above 35 (M2 group)

Men	30 ≤ BMI ≤ 35		BMI > 35	
	AC – NAC LV Walls	p - value	AC – NAC LV Walls	p - value
STRESS	ANT	0.486	ANT	0.348
	LAT	0.000*	LAT	0.020*
	INF	0.000*	INF	0.000*
	SEP	0.003*	SEP	0.003*
	API	0.529	API	0.771
REST	ANT	0.500	ANT	0.627
	LAT	0.000*	LAT	0.014*
	INF	0.000*	INF	0.000*
	SEP	0.001*	SEP	0.075
	API	0.045*	API	0.698
* statistical significant differences found between CT-AC and NAC perfusion percentage results				

4. DISCUSSION

As described in the literature, the sub-diaphragmatic attenuation is the one that generates the most common AA in the INF walls of the LV, resulting in a decrease of the perfusion percentage [18]. This AA do not only affects the male study population (M1 and M2) as was previously described in literature [19], but is also present in female groups F1 and F2, due to the large abdominal area causing the diaphragmatic attenuation reducing PT count density in the INF walls [6]. This AA is corrected by CT-AC system leading to statistical significant differences in the quantification perfusion values in the affected areas.

According to the results LAT walls present statistical significant differences between CT-AC and NAC data excepting the F2. For F1 group the perfusion differences are explained by the breast attenuation since all the studies were performed in the supine position with no brassiere causing the lateral positioning of the breast tissues which attenuate gamma photons prevenient from the heart. Once again, this AA is corrected by the CT-AC system and it can explain the differences between corrected NAC quantitative results [19].

The distance of the gamma camera detectors to the patient can also be a factor that affects the quantitative results of the perfusion. As it is known if the distance between detectors and the source is increased, the scatter effect will be more intense, increasing the noise in the pictures

and degrading its contrast and spatial resolution. In consequence, the areas that are exposed to this effect will appear with perfusion variations in myocardial activity which are not real perfusion defects. AC methods correct the perfusion quantification [20], increasing perfusion percentage quantitative results in areas that are affected by improving the spatial resolution and the image contrast. For this reasons CT-AC and NAC data present differences when this problem arises. It can be observed in M1 and M2 groups where the distance problem may have caused the perfusion differences in the LAT walls. Nevertheless, some man have a large volume of breast tissue and like F1 the soft tissue attenuation is responsible for the variations of corrected and non-corrected results. Unexpectedly no statistical differences were found in the LAT walls for female population with BMI > 35. It can be explained by poor AC provided by CT systems as CT parameters were not modified even when scanning patients with very high weight in order to obtain good quality AM. Still this result need to be investigated.

Variations found in the SEP walls for M1 (Stress and Rest), M2 (Stress) and F1 (Rest) groups and API walls for M1 (Rest) group can be explained by truncation artifacts emerged from CT-AC techniques [21]. Those rise when attenuation maps do not contain the whole thoracic wall of the patient. In consequence AE in anterior direction are underestimated creating false quantitative results which differ from NAC outcome. Otherwise, the differences between CT-AC and conventional MPI results can be caused by overestimation of the AE in the SEP and API left myocardium walls [4,21].

All the groups, except the F2, present quantitative variations in in the SEP walls. M2 group just presents different results on the Stress and F1 groups on the Rest studies.

No statistical differences were found in the ANT walls as it was described by the majority of the literature, for the female population [5] probably due to the lateral positioning of the breast tissues during the scans, which do not results in the ANT LV walls attenuation but in the LAT attenuation.

CT-AC affects the MPI quantification results in several LV walls. It's very important to recognise and try to avoid AC generated artifacts by being alert for body truncation, patient motion, accurate registration of AM and emission data. Many authors suggest that it's important that both non-corrected and corrected image sets should be reviewed to understand which of the results can be accepted as real.

5. CONCLUSIONS

As demonstrated significant differences were found between CT-AC and NAC MPI data for male and female population, in various areas of the LV. Some of the differences were no expected neither described by the literature. Therefore all the members of the staff involved in MPI exams should be aware not only of the AA but also of the pitfalls of the CT-AC.

In the future this study should be continued with a larger study population and adding other parameters such as brassiere size, for females, abdominal perimeter, for males and female, and the quantification of the soft tissues thickness adjacent to the heart in order to correlate the AC and NAC MPI differences with the patients anatomy.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the financial support of Project IPL, IDI&CA/SOFTIMOB.

REFERENCES

[1] A. P. Pazhenkottil, J.-R. Ghadri, R. N. Nkoulou, M. Wolfrum, R. R. Buechel, S. M. Küest, L. Husmann, B. A. Herzog, O. Gaemperli, and P. A. Kaufmann, "Improved Outcome Prediction by SPECT Myocardial Perfusion Imaging After CT Attenuation Correction," *J. Nucl. Med.*, vol. 52, no. 2, pp. 196–200, Feb. 2011.

[2] Zaret BL, Beller GA., "Clinical nuclear cardiology: state of the art and future directions.," Philadelphia: PA: Mosby, 2005, pp. 215–355.

[3] A. Flotats, J. Knuuti, M. Gutberlet, C. Marcassa, F. M. Bengel, P. A. Kaufmann, M. R. Rees, B. Hesse, and Cardiovascular Committee of the EANM, the ESCR and the ECNC, "Hybrid cardiac imaging: SPECT/CT and PET/CT. A joint position statement by the European Association of Nuclear Medicine (EANM), the European Society of Cardiac Radiology (ESCR) and the European Council of Nuclear Cardiology (ECNC)," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 38, no. 1, pp. 201–212, Jan. 2011.

[4] E. Fricke, H. Fricke, R. Weise, A. Kammeier, R. Hagedorn, N. Lotz, O. Lindner, D. Tschoepe, and W. Burchert, "Attenuation correction of myocardial SPECT perfusion images with low-dose CT: evaluation of the method by comparison with perfusion PET," *J. Nucl. Med. Off. Publ. Soc. Nucl. Med.*, vol. 46, no. 5, pp. 736–744, May 2005.

[5] E. G. DePuey, "How to detect and avoid myocardial perfusion SPECT artifacts.," *J. Nucl. Med. Off. Publ. Soc. Nucl. Med.*, vol. 35, no. 4, pp. 699–702, 1994.

[6] R. A. Dvorak, R. K. J. Brown, and J. R. Corbett, "Interpretation of SPECT/CT Myocardial Perfusion Images: Common Artifacts and Quality Control Techniques," *RadioGraphics*, vol. 31, no. 7, pp. 2041–2057, Nov. 2011.

[7] A. Cuocolo, "Attenuation correction for myocardial perfusion SPECT imaging: still a controversial issue," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 38, no. 10, pp. 1887–1889, Aug. 2011.

[8] J. A. Patton and T. G. Turkington, "SPECT/CT Physical Principles and Attenuation Correction," *J. Nucl. Med. Technol.*, vol. 36, no. 1, pp. 1–10, Mar. 2008.

[9] S. Goetze, T. L. Brown, W. C. Lavely, Z. Zhang, and F. M. Bengel, "Attenuation correction in myocardial perfusion SPECT/CT: effects of misregistration and value of reregistration," *J. Nucl. Med. Off. Publ. Soc. Nucl. Med.*, vol. 48, no. 7, pp. 1090–1095, Jul. 2007.

[10] X. Ou, L. Jiang, R. Huang, F. Li, Z. Zhao, and L. Li, "Computed tomography attenuation correction improves the risk stratification accuracy of myocardial perfusion imaging," *Nucl. Med. Commun.*, vol. 34, no. 5, pp. 495–500, May 2013.

[11] H. Zaidi and B. Hasegawa, "Determination of the attenuation map in emission tomography," *J. Nucl. Med. Off. Publ. Soc. Nucl. Med.*, vol. 44, no. 2, pp. 291–315, Feb. 2003.

[12] J. K. Kahn, I. McGhie, M. S. Akers, M. N. Sills, T. L. Faber, P. V. Kulkarni, J. T. Willerson, and J. R. Corbett, "Quantitative rotational tomography with 201Tl and 99mTc 2-methoxy-isobutyl-isonitrile. A direct comparison in normal individuals and patients with coronary artery disease," *Circulation*, vol. 79, no. 6, pp. 1282–1293, Jun. 1989.

[13] E. P. Ficaro and J. R. Corbett, "Advances in quantitative perfusion SPECT imaging," *J. Nucl. Cardiol.*, vol. 11, no. 1, pp. 62–70, Jan. 2004.

[14] G. Germano and D. S. Berman, "Quantitative Gated SPECT," *J. Nucl. Med.*, vol. 42, no. 3, pp. 528–529, Mar. 2001.

[15] "Obesity and Overweight for Professionals: Adult: Defining - DNPAO - CDC." [Online]. Available: http://www.cdc.gov/obesity/adult/defining.html. [Accessed: 17-May-2015].

[16] S. C. Blankespoor, X. Xu, K. Kaiki, J. K. Brown, H. R. Tang, C. E. Cann, and B. H. Hasegawa, "Attenuation correction of SPECT using X-ray CT on an emission-transmission CT system: myocardial perfusion assessment," *IEEE Trans. Nucl. Sci.*, vol. 43, no. 4, pp. 2263–2274, Aug. 1996.

[17] J. N. G. Maria Helena Pestana, "ANÁLISE DE DADOS PARA CIÊNCIAS SOCIAIS A Complementaridade do SPSS 6 ª EDIÇÃO Revista, Atualizada e Aumentada MARIA HELENA PESTANA JOÃO NUNES GAGEIRO," 2014.

[18] J. Wheat and G. Currie, "Recognising And Dealing With Artifact In Myocardial Perfusion SPECT," 2007.

[19] S. Burrell and A. MacDonald, "Artifacts and pitfalls in myocardial perfusion imaging," *J. Nucl. Med. Technol.*, vol. 34, no. 4, pp. 193–211; quiz 212–214, Dec. 2006.

[20] M. Garcheva-Tsacheva, "SPECT-CT in myocardial perfusion scintigraphy," *Arch. Oncol.*, vol. 20, no. 3–4, pp. 132–135, 2012.

[21] S. J. C. Timothy M Bateman, "Attenuation correction single-photon emission computed tomography myocardial perfusion imaging.," *Semin. Nucl. Med.*, vol. 35, no. 1, pp. 37–51, 2005.

[22] R. M. T. Giubbini, S. Gabanelli, S. Lucchini, G. Merli, E. Puta, C. Rodella, F. Motta, B. Paghera, P. Rossini, A. Terzi, and F. Bertagna, "The value of attenuation correction by hybrid SPECT/CT imaging on infarct size quantification in male patients with previous inferior myocardial infarct," *Nucl. Med. Commun.*, vol. 32, no. 11, pp. 1026–1032, Nov. 2011.

SYMCOMP 2017
Guimarães, 6-7 April 2017

DATA SELECTION TO IMPROVE SAMPLES QUALITY AND TO OVERCOME THE CURRENT PREDICTIONS

Antonio J. Tallón-Ballesteros[1,2]* and António Ruano[2,3]

1: Department of Languages and Computer Systems
Higher Technical School of Computer Science Engineering
University of Seville (Spain)
Reina Mercedes Av. 41012-Seville (Spain)
e-mail: atallon@us.es, web: http://www.lsi.us.es

2: Department of Electronics and Computer Science Engineering
Faculty of Science and Technology
University of Algarve (Portugal)
Campus de Gambelas, 8005-139, Faro (Portugal)
e-mail: {ajballesteros, aruano}@ualg.pt, web: http://deei.fct.ualg.pt/

3: IDMEC
Instituto Superior Técnico
Universidade de Lisboa, 1049-001 Lisboa, (Portugal)
e-mail: aruano@idmec.ist.utl.pt, web: http://www.idmec.ist.utl.pt/

Keywords: Data reduction, Data pre-processing, Classification, Convex hull, Data quality

Abstract *Classification task performance is affected by the data quality. Representative samples are the key ingredient of any good training procedure that eventually could conduct to an accurate assessment. This paper makes emphasis on convex hull algorithm to pick up the most relevant instances. The empirical study covers some machine learning algorithms from different paradigms such as, mainly, decision trees, lazy and rule-based classifiers. Generally speaking, some enhancements take place. Nevertheless the behavior in some cases is very close to the starting point. The advantage is clear because we do not need to work with all the samples and it would reduce the computational cost in terms of time. Moreover, for the future, we may opt to store new samples only if the historical data quality could be improved through the aforementioned samples. Experimentation is supported by some problems with up to almost four thousands of instances and no more than one hundred of features. The results are very promising in term of the performance increase. We plan to extend to other real-world problems and also to try to take the advantage of the current proposal with the successive application of other data preparation procedures.*

REFERENCES

[1] Duda, R.O., Hart, P.E., Stork, D. *Pattern Classification*, second ed., Wiley, 2001

[2] Khosravani, H.R., Ruano, A.E., Ferreira, P.M. "A convex hull-based data selection method for data driven models". *Applied Soft Computing* 47, pp. 515-533, 2016.

[3] Tallón-Ballesteros, A.J., Riquelme, J.C. "Deleting or keeping outliers for classifier training?" *Nature and Biologically Inspired Computing (NaBIC), 2014 Sixth World Congress on*. IEEE, pp. 281-286, 2014.

[4] Michie, D., Spiegelhalter, D.J., Taylor, C.C. Machine Learning, Neural and Statistical *Classification*, Ellis Horwood, 1994.

[5] Bache, K., Lichman, M. "UCI machine learning repository", *URL: http://archive. ics. uci. edu/ml*, 2013.

[6] Quinlan, J. *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, CA, 1993.

[7] Cover, T., Hart, P. "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory* Vol. 13(1), pp. 21-27, 1967.

[8] Aha, D., Kibler, D., Albert, M.K. "Instance-based learning algorithms", *Machine Learning* Vol. 6, pp. 37-66, 1991.

[9] Vapnik, V. *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[10] Frank, E., Witten, I.H. "Generating accurate rule sets without global optimization". *ICML 1998, Proceedings of the fifteenth international conference on machine learning*. Madison, Wisconsin, USA: Morgan Kaufmann, pp. 144-151, 1998.

SYMCOMP 2017
Guimarães, 6-7 April 2017

USING WOLFRAM MATHEMATICA IN SPECTRAL THEORY

Ana C. Conceição* and José C. Pereira

Center for Functional Analysis, Linear Structures and Applications (CEAFEL)[1]
Faculdade de Ciências e Tecnologia
Universidade do Algarve
8005-139 Faro
e-mail: aicdoisg@gmail.com

Keywords: Spectral theory, Wolfram Mathematica, symbolic computation, numeric computation

Abstract. *The development of the spectral theory is motivated by the need to solve problems emerging from several fields in Mathematics and Physics. Some progress has been achieved for some classes of singular integral operators (SIOs) whose properties allow the use of particular strategies in the study of the spectral problem but there is still no general and explicit method for obtaining the spectrum of any given SIO. Furthermore, the existing algorithms allow, in general, to study the spectrum of some kind of SIOs but they are not designed to be implemented on a computer. The main goal of this work is to present our spectral algorithms [ASpec-Hankel], [ASpecPaired-Scalar], and [ASpecPaired-Matrix] which use the symbolic and numeric computation capabilities of the computer algebra system Mathematica to explore the spectra of some classes of SIOs, defined on the unit circle. These analytical algorithms allow us to check, for each considered SIO, if a complex number (chosen arbitrarily) belongs to its spectrum.*

[1]This research is supported by Fundação para a Ciência e Tecnologia (Portugal) through the Center for Functional Analysis, Linear Structures and Applications.

1 INTRODUCTION

In recent years, several software applications were made available to the general public with extensive capabilities of symbolic computation. These applications, known as computer algebra systems (CAS), allow to delegate to a computer all, or a significant part, of the symbolic calculations present in many mathematical algorithms. In our work we use the CAS *Mathematica*[2] to implement on a computer analytical algorithms developed by us and other authors within operator theory. In the last years we designed and/or implemented analytical algorithms for solving integral equations [2, 5], analytical algorithms to factorize scalar and matrix functions [4, 5], calculation techniques to compute singular integrals [3], and more recently analytical algorithms to study the spectrum [7] and the kernel [6] of several classes of singular integral operators. It is our belief that the construction and implementation on a computer of these kind of analytical algorithms is a very interesting line of research. In fact, the development of the spectral theory is motivated by the need to solve problems emerging from several fields in Mathematics and Physics. Some progress has been achieved for some classes of singular integral operators (SIOs) whose properties allow the use of particular strategies in the study of the spectral problem but there is still no general and explicit method for obtaining the spectrum of any given SIO. Furthermore, the existing algorithms allow, in general, to study the spectrum of some kind of SIOs but they are not designed to be implemented on a computer.

The main goal of this work is to present our spectral algorithms [ASpec-Hankel], [ASpecPaired-Scalar], and [ASpecPaired-Matrix] which use the symbolic and numeric computation capabilities of *Mathematica* to explore the spectra of some classes of SIOs, defined on the unit circle. These analytical algorithms allow us to check, for each considered SIO, if a complex number (chosen arbitrarily) belongs to its spectrum. It is considered the one-dimensional and the matrix cases.

2 MAIN DEFINITIONS

Let $\mathbb{T}$ denote the unit circle in the complex plane. Let $\mathbb{T}_+$ and $\mathbb{T}_-$ denote the open unit disk and the exterior region of the unit circle (∞ included), respectively. Let $L_\infty(\mathbb{T})$ be the space of all essentially bounded functions defined on the unit circle. Let $H_\infty(\mathbb{T})$ be the class of all bounded analytic functions in the interior of the unit circle.

2.1 Singular Integral Operators

Let us consider the singular integral, defined almost everywhere on $\mathbb{T}$, associated with the singular integral operator $S_\mathbb{T}$,

$$S_\mathbb{T}\varphi(t) = \frac{1}{\pi i} \int_\mathbb{T} \frac{\varphi(\tau)}{\tau - t}\, d\tau,\ t \in \mathbb{T}, \tag{1}$$

[2]Wolfram *Mathematica* is a symbolic mathematical computation program used in many scientific, engineering, and computing fields. It was conceived by Stephen Wolfram and is developed by Wolfram Research.

with Cauchy kernel, defined on the Lebesgue space $L_2(\mathbb{T})$. It is known that (1) is a selfadjoint, unitary, and bounded linear operator in $L_2(\mathbb{T})$ (see, for instance, [9]). Thus, we can associate with $S_{\mathbb{T}}$ two complementary projection operators

$$P_{\pm} = (I \pm S_{\mathbb{T}})/2, \tag{2}$$

where I represents the identity operator.

Let $\varphi, \psi \in [L_\infty(\mathbb{T})]_{n,n}$. Operators of the form $T = \varphi I + \psi S_{\mathbb{T}}$ and $\widetilde{T} = \varphi I + S_{\mathbb{T}}\psi I$ are linear and bounded SIOs (see, for instance, [9]). In the following, these operators will be written in a more convenient form as

$$T_{\{a,b\}} = aP_+ + bP_- \tag{3}$$

and

$$\widetilde{T}_{\{a,b\}} = P_+aI + P_-bI, \tag{4}$$

where $a = \varphi+\psi$ and $b = \varphi-\psi$. We will call these operators, paired SIOs, with coefficients a and b.

Obviously, since $S_{\mathbb{T}} = P_+ - P_-$, $S_{\mathbb{T}}$ is a paired SIO that belongs to classes (3) and (4).

Let us also consider the special class of self-adjoint SIOs

$$T_+ = P_+\varphi P_-\overline{\varphi}P_+, \tag{5}$$

where $\varphi \in L_\infty(\mathbb{T})$.

2.2 Generalized Factorization Concept

Let us now introduce the generalized [3] factorization concept in $L_2(\mathbb{T})$.

A matrix function $A \in [L_\infty(\mathbb{T})]_{n,n}$, that is, a matrix function whose entries are essentially bounded functions on the curve $\mathbb{T}$, admits a left (right) generalized factorization in $L_2(\mathbb{T})$ if it can be represented as

$$A = A_+\Lambda A_- \qquad (A = A_-\Lambda A_+), \tag{6}$$

where

$$A_+^{\pm 1} \in [\mathrm{im}P_+]_{n,n}\,, \quad A_-^{\pm 1} \in [\mathrm{im}P_- \oplus \mathbb{C}]_{n,n}\,, \quad \Lambda(t) = \mathrm{diag}\{t^{\varkappa_i}\}_{j=1}^n, \tag{7}$$

$\varkappa_j \in \mathbb{Z}$, $j = \overline{1,n}$, with $\varkappa_1 \geq \varkappa_2 \geq ... \geq \varkappa_n$, and $A_+P_+A_-I$ $(A_-P_+A_+I)$ represents a bounded linear operator in $[L_2(\mathbb{T})]_n$.

The number $\varkappa = \sum_{j=1}^{n} \varkappa_j$ is called the left (right) factorization index of the matrix function A. The integers $\varkappa_j$ are called the left (right) partial indices of A. If $\varkappa_j = 0$, $j = \overline{1,n}$, then A is said to admit a left (right) canonical generalized factorization.

[3] Although many of the results presented in this paper can be generalized (see, for instance, [1], [8]) to the space $L_p(\Gamma)$, where Γ is a closed Carleson curve, we decided to state them only for $L_2(\mathbb{T})$ due to the application of symbolic computation for the construction of nontrivial examples.

Let $\mathcal{R}(\mathbb{T})$ be the algebra of rational functions without poles on $\mathbb{T}$ and let $\mathcal{R}_{\pm}(\mathbb{T})$ denote the subsets of $\mathcal{R}(\mathbb{T})$ whose elements are without poles in $\mathbb{T}_{\pm}$.

Any non-singular rational matrix function $A \in [\mathcal{R}(\mathbb{T})]_{n,n}$ admits a left (right) generalized factorization of the form (6), where

$$A_{+}^{\pm 1} \in [\mathcal{R}_{+}(\mathbb{T})]_{n,n}\,, \quad A_{-}^{\pm 1} \in [\mathcal{R}_{-}(\mathbb{T})]_{n,n}\,.$$

The [ARFact-Matrix] algorithm [4] can be used to compute explicit left and right factorizations for non-singular rational matrix function defined on the unit circle.

For the particular rational scalar case

$$\varkappa = z_{+} - p_{+}, \tag{8}$$

where z_{+} is the number of zeros of A in $\mathbb{T}_{+}$ (with regard to their multiplicities) and p_{+} is the number of poles of A in $\mathbb{T}_{+}$ (with regard to their multiplicities) (see, for instance, [4]).

The [ARFact-Scalar] algorithm [4] can be used to compute explicit factorizations for any factorable rational function defined on the unit circle. In particular, the index (8) is always obtained explicitly.

3 SPECTRAL ALGORITHMS

In this section we present our spectral algorithms [ASpec-Hankel], [ASpecPaired-Scalar], and [ASpecPaired-Matrix] which use the symbolic and numeric computation capabilities of the computer algebra system *Mathematica* to explore the spectra of some classes of SIOs, defined on the unit circle.

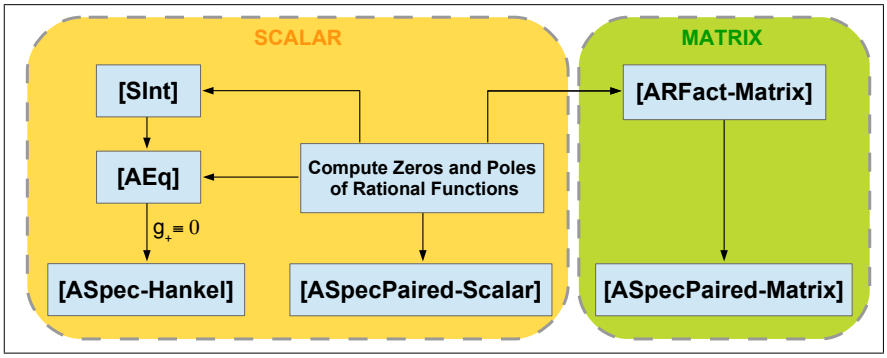


Figure 1: General flowchart.

One of the most important steps in our algorithms is the calculation of zeros and poles of rational functions. The computer algebra system *Mathematica* uses *Root* objects to represent solutions of algebraic equations in one variable, when it is impossible to find explicit formulas for these solutions. The *Root* object is not a mere denoting symbol but

rather an expression that can be symbolically manipulated and numerically evaluated with any desired precision.

Figure 1 represents a General Flowchart that also includes the auxiliary operator theory algorithms for the creation and implementation of our spectral algorithm: [SInt], [AEq], and [ARFact-Matrix].

3.1 Auxiliary operator theory algorithms

In this subsection it is given a brief description of the algorithms [SInt], [AEq], and [ARFact-Matrix].

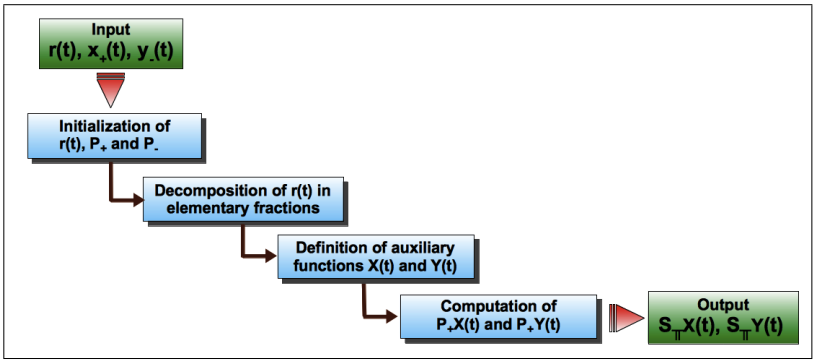


Figure 2: [SInt] flowchart.

The [SInt] algorithm described in [3] computes the Cauchy type singular integrals $S_{\mathbb{T}}\varphi(t)$, $P_+\varphi(t)$, and $P_-\varphi(t)$, when the essentially bounded function φ can be represented as $\varphi(t) = r(t)[x_+(t) + y_-(t)]$, where $x_+, \overline{y_-} \in H_\infty(\mathbb{T})$ and $r \in \mathcal{R}(\mathbb{T})$ (see Figure 2).

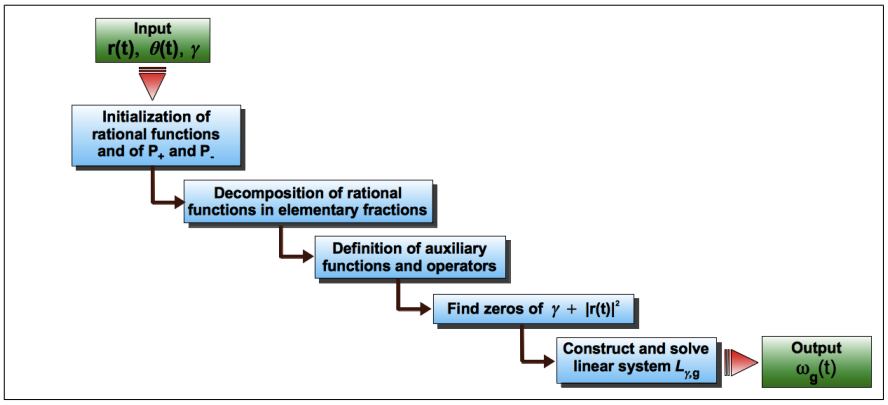


Figure 3: [AEq] flowchart.

The [AEq] algorithm presented in [5] computes explicit solutions of non-homogeneous

singular integral equations related with the SIO T_+ defined in (5), with $\varphi \in L_\infty(\mathbb{T})$ (see Figure 3).

[ARFact-Matrix]

Input the entries of $A(t)$, where $A(t)$ is a $n \times n$ rational matrix, invertible in the unit circle.
if $A(t) \notin [\mathcal{R}_+(\mathbb{T})]_{n,n}$
  then $r(t) \leftarrow$ monic polynomial with zeros equal to the poles of $A(t)$ which lie in $\mathbb{T}_+$, and with corresponding maximum multiplicity.
    $A(t) \leftarrow r(t)A(t)$
$\{(z_s, m_s)\}_{s\leftarrow 1}^{q-1} \cup (0, m_q) \leftarrow$ list of zeros of the function $\det(\mathrm{A(t)})$ and their corresponding multiplicities.
if 0 is not a zero of $\det(\mathrm{A(t)})$
  then $m_q \leftarrow 0$
$\tilde{A}(t) \leftarrow A(t)$
for $s \leftarrow 1$ to q
  do $f_j(t) \leftarrow$ j-th row of $\tilde{A}(t)$
    $(p_{js})_{j\leftarrow 1}^{n} \leftarrow$ multiplicities of z_s as a zero of vector function $f_j(t)$.
    Permute the rows of $\tilde{A}(t)$ such that $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_n$.
    while $\sum_{j=1}^{n} p_j < m_s$
      do Find constants $c_1, c_2, \ldots, c_l$ where $l \leqslant n$ and $c_l = 1$, such that the function $f(t) = \sum_{j=1}^{l} \frac{c_j f_j(t)}{(t-z_s)^{p_j}}$ vanishes at $t = z_s$.
        Replace row l of $\tilde{A}(t)$ with the vector function $(t - z_s)^{p_l} f(t)$.
        Recompute p_l and permute the rows of $\tilde{A}(t)$ such that $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_n$.
    if $s < q$
      then $D_s(t) \leftarrow \left(\left(\frac{t}{t-z_s}\right)^{p_j} \delta_{jk}\right)_{j,k\leftarrow 1}^{n}$ ▷ δ_{jk} is the Kronecker delta symbol.
        $\tilde{A}(t) \leftarrow D_s(t)\tilde{A}(t)$
        $m_q \leftarrow m_s + m_q$
      else $D_q(t) \leftarrow (t^{p_j}\delta_{jk})_{j,k\leftarrow 1}^{n}$
        $\tilde{A}_+(t) \leftarrow D_q^{-1}(t)\tilde{A}(t)$
        $\tilde{A}_-(t) \leftarrow A(t)\tilde{A}^{-1}(t)$
if $A(t) \notin [\mathcal{R}_+(\mathbb{T})]_{n,n}$
  then $\hat{D}(t) \leftarrow r^{-1}(t)D_q(t)$ ▷ $\hat{D}(t)$ is a diagonal matrix with nonzero entries $d_j(t) = t^{p_j} r^{-1}(t) \in \mathcal{R}(\mathbb{T})$
    for $j \leftarrow 1$ to n
      do $zeros \leftarrow$ list of zeros of $d_j(t)$
        $poles \leftarrow$ list of poles of $d_j(t)$
        $d_j(t) \leftarrow$ ARFact-Scalar($zeros$,$poles$,1)
    $\hat{D}_-(t) \leftarrow (d_{j-}(t)\delta_{jk})_{j,k\leftarrow 1}^{n}$ ▷ The ARFact-Scalar procedure computes the scalar factorization $d_j(t) = d_{j-}(t)\dot{t^{p_j}} d_{j+}(t)$
    $\hat{D}_+(t) \leftarrow (d_{j+}(t)\delta_{jk})_{j,k\leftarrow 1}^{n}$
    $A_-(t) \leftarrow \tilde{A}_-(t)\hat{D}_-(t)$
    $\overset{*}{D}(t) \leftarrow \left(\dot{t^{p_j}}\delta_{jk}\right)_{j,k\leftarrow 1}^{n}$ ▷ The final factorization is
    $A_+(t) \leftarrow \hat{D}_+(t)\tilde{A}_+(t)$ $A(t) = A_-(t)\overset{*}{D}(t)A_+(t)$
  else $A_-(t) \leftarrow \tilde{A}_-(t)$
    $\overset{*}{D}(t) \leftarrow D_q(t)$ ▷ The final factorization is
    $A_+(t) \leftarrow \tilde{A}_+(t)$ $A(t) = A_-(t)\overset{*}{D}(t)A_+(t)$

Figure 4: [ARFact-Matrix] pseudocode.

The factorization algorithm [ARFact-Matrix] described in [4] computes explicit factorizations for given factorable rational matrix functions (see Figure 4).

3.2 Spectral algorithms

In this section we present our spectral algorithms [ASpec-Hankel], [ASpecPaired-Scalar], and [ASpecPaired-Matrix].

3.2.1 [ASpec-Hankel] algorithm

The [ASpec-Hankel] algorithm checks if a complex number λ (chosen arbitrarily) belongs to the spectrum of the SIO T_+ defined by (5), with essentially bounded coefficients defined on the unit circle. In the design of this spectral algorithm we used the [AEq] algorithm to compute explicit solutions of non-homogeneous singular integral equations.

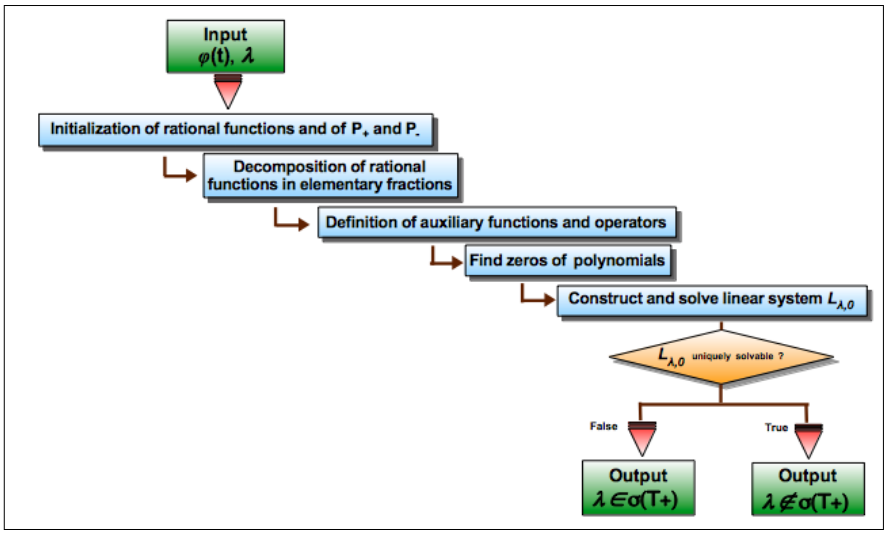


Figure 5: [ASpec-Hankel] flowchart.

For each inputed complex value λ and function $\varphi \in L_\infty(\mathbb{T})$ the algorithm gives one of the *outputs* (see Figure 5):

$$[Output\ 1] \quad \lambda \in \sigma(T_+)$$
$$[Output\ 2] \quad \lambda \notin \sigma(T_+)$$

3.2.2 [ASpecPaired-Scalar] algorithm

The [ASpecPaired-Scalar] algorithm checks if a complex number λ (chosen arbitrarily) belongs to the spectra of the one-dimensional paired SIOs $T_{\{a,b\}}$ and $\widetilde{T}_{\{a,b\}}$, with rational coefficients defined on the unit circle. The symbolic computation capabilities of *Mathematica*, and the pretty-print functionality, allow the [ASpecPaired- Scalar] code to be very simple (see Figure 6).

For each pair of inputed functions $a, b \in \mathcal{R}(\mathbb{T})$ the algorithm gives one of the *outputs*:

$$[\textit{Output 1}] \quad \lambda \in \sigma(T_{\{a,b\}}) \text{ and } \lambda \in \sigma(\widetilde{T}_{\{a,b\}})$$
$$[\textit{Output 2}] \quad \lambda \notin \sigma(T_{\{a,b\}}) \text{ and } \lambda \notin \sigma(\widetilde{T}_{\{a,b\}})$$

[ASpecPaired-Scalar]

```
1  Input rational functions a(t) and b(t), and complex value λ.
2  if a(t) ≡ λ or b(t) ≡ λ
3     then λ ∈ σ(T_{a,b})                                   ▷ λ ∈ σ(T̃_{a,b})
4     else  r(t) ← (a(t)−λ)/(b(t)−λ)
5           {z_i}_{i←1}^m ← list of zeros of function r(t)
6           {p_j}_{j←1}^n ← list of poles of function r(t)
7           if at least one zero or pole of r(t) lies in 𝕋
8              then λ ∈ σ(T_{a,b})                          ▷ λ ∈ σ(T̃_{a,b})
9              else  z_+ ← number of zeros of r that lie in 𝕋_+
                           with regard to their multiplicities
10                   p_+ ← number of poles of r that lie in 𝕋_+
                           with regard to their multiplicities
11                   ϰ ← z_+ − p_+
12                   if ϰ = 0
13                      then λ ∉ σ(T_{a,b})                ▷ λ ∉ σ(T̃_{a,b})
14                      else  λ ∈ σ(T_{a,b})               ▷ λ ∈ σ(T̃_{a,b})
```

Figure 6: [ASpecPaired-Scalar] pseudocode.

3.2.3 [ASpecPaired-Matrix] algorithm

The [ASpecPaired-Matrix] algorithm checks if a complex number λ (chosen arbitrarily) belongs to the spectra of the paired SIOs $T_{\{a,b\}}$ and $\widetilde{T}_{\{a,b\}}$, with rational matrix coefficients defined on the unit circle. As in the scalar case, the implementation of this spectral algorithm with *Mathematica* makes the results of lengthy and complex calculations available in a simple way. In the design of this spectral algorithm we used the factorization algorithm [ARFact-Matrix] to compute explicit factorizations for auxiliar rational matrix functions, defined on the unit circle.

For each pair of inputed matrix functions $a, b \in [\mathcal{R}(\mathbb{T})]_{n,n}$ the algorithm gives one of the *outputs* (see Figure 7):

$$[\textit{Output 1}] \quad \lambda \notin \sigma(T_{\{a,b\}}) \text{ and } \lambda \notin \sigma(\widetilde{T}_{\{a,b\}})$$
$$[\textit{Output 2}] \quad \lambda \notin \sigma(T_{\{a,b\}}) \text{ and } \lambda \in \sigma(\widetilde{T}_{\{a,b\}})$$
$$[\textit{Output 3}] \quad \lambda \in \sigma(T_{\{a,b\}}) \text{ and } \lambda \in \sigma(\widetilde{T}_{\{a,b\}})$$
$$[\textit{Output 4}] \quad \lambda \in \sigma(T_{\{a,b\}}) \text{ and } \lambda \notin \sigma(\widetilde{T}_{\{a,b\}})$$

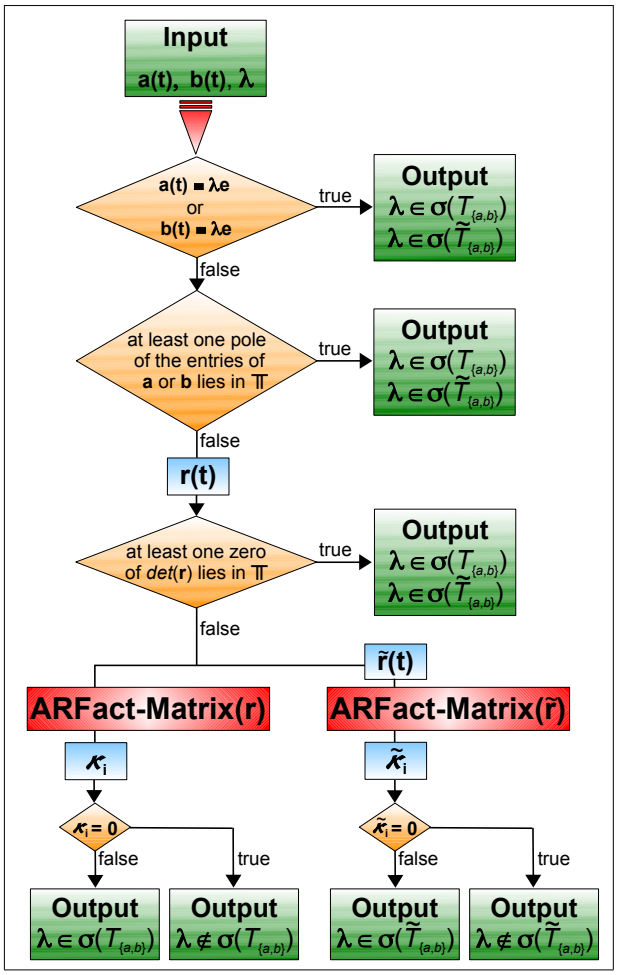


Figure 7: [ASpecPaired-Matrix] flowchart.

4 CONCLUSIONS

The design of our analytical algorithms is focused on the possibility of implementing on a computer all, or a significant part, of the extensive symbolic and numeric calculations present in the algorithms. The methods developed rely on innovative techniques of operator theory and have a great potential of extension to ever more complex and general problems. Also, by implementing these methods on a computer, new and powerful tools are created making the results of lengthy and complex calculations available in a simple way to researchers of different areas.

REFERENCES

[1] Calderón, A.P., "Cauchy integrals on Lipschitz curves and related operators", *Proceedings of the National Academy of Sciences of the United States of America* Vol. **74(4)**, pp.1324-1327, 1977.

[2] Conceição, A.C., Kravchenko, V.G., Pereira, J.C., "About explicit factorization of some classes of non-rational matrix functions", *Mathematische Nachrichten*, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim Vol. **280(9-10)**, pp.1022-1034, 2007.

[3] Conceição, A.C., Kravchenko, V.G., Pereira, J.C., "Computing some classes of Cauchy type singular integrals with *Mathematica* software", *Advances in Computational Mathematics*, Springer US Vol. **39(2)**, pp. 273-288, 2013.

[4] Conceição, A.C., Kravchenko, V.G., Pereira, J.C., "Rational Functions Factorization Algorithm: a symbolic computation for the scalar and matrix cases", *Proceedings of the 1st National Conference on Symbolic Computation in Education and Research*, Lisboa - Portugal, 2012.

[5] Conceição, A.C., Kravchenko, V.G., Pereira, J.C., "Factorization algorithm for Some special non-rational matrix functions", *Topics in Operator Theory, Operator Theory: Advances and Applications*, Birkhäuser, Basel, Ed. Ball, J., Bolotnikov, V., Rodman, L., Helton, J., Spitkovsky, I. Vol. **202**, pp. 87-109, 2010.

[6] Conceição, A.C., Marreiros, R.C., Pereira, J.C., "Symbolic Computation Applied to the Study of the Kernel of a Singular Integral Operator with Non-Carleman Shift and Conjugation", *Mathematics in Computer Science*, Springer International Publishing Vol. **10(3)** , pp. 365-386, 2016.

[7] Conceição, A.C., Pereira, J.C., "Exploring the Spectra of Some Classes of Singular Integral Operators with Symbolic Computation", *Mathematics in Computer Science*, Springer International Publishing Vol. **10(2)** , pp. 291-309, 2016.

[8] Davis, G., "Opérateurs intégraux singuliers sur certaines courbes du plan complexe", *Annales scientifiques de l'École normale supérieure* Vol. **17(1)**, pp.157-189, 1984.

[9] Gohberg, I., Krupnik, N. *One-Dimensional Linear Singular Integral Equations*, Operator Theory: Advances and Applications, Birkhäuser Basel Vol. **53**, 1992.

SYMCOMP 2017
Guimarães, 6-7 April 2017

VERY HIGH ORDER FINITE VOLUME APPROXIMATION FOR THE 1D STEADY-STATE EULER SYSTEM

Gaspar J. Machado[1]*, Stéphane Clain[1] and Raphael Loubère[2]

1: Centro de Matemática, Universidade do Minho, Portugal
2: CNRS and Institut de Mathématiques de Bordeaux (IMB), Université de Bordeaux, France

e-mails: gjm@math.uminho.pt ; clain@math.uminho.pt ; raphael.loubere@u-bordeaux.fr

Keywords: Finite volume, MOOD, very high order, Euler system

Abstract: *We propose a finite volume numerical scheme devoted to solve the one-dimensional steady-state Euler system.*

High-accuracy (up to the sixth-order presently) is achieved thanks to polynomial reconstructions while stability is provided with an \textit{a posteriori} MOOD method which control the cell polynomial degree for eliminating non-physical oscillations in the vicinity of discontinuities.

Such a procedure demands the determination of a chain detector to discriminate between troubled and valid cells, a cascade of polynomial degrees to be successively tested when oscillations are detected, and a parachute scheme corresponding to the last, viscous, and robust scheme of the cascade.

The obtained results demonstrate that the scheme manages to retrieve smooth solutions with optimal order of accuracy but also irregular solutions without spurious oscillations.

SYMCOMP 2017
Guimarães, 6-7 April 2017

VERY HIGH ORDER FINITE VOLUME APPROXIMATION FOR THE 1D BIHARMONIC OPERATOR

Hélder C. Barbosa[1]*, Ricardo Costa[2] and Gaspar Machado[1]

1: Centro de Matemática, Universidade do Minho, Portugal

2: Institute for Polymers and Composites/I3N, Universidade do Minho, Portugal

e-mails: a77835@alunos.uminho.pt ; pg24046@alunos.uminho.pt ; gjm@math.uminho.pt

Keywords: Finite volumen, polynomyal reconstruction, very high order, biharmonic equation

Abstract *We propose a very high order finite volume numerical scheme devoted to solve the one-dimensional biharmonic operator with different types of boundary conditions.*

The scheme is based on polynomial reconstructions of the solutions of degree at least 3, which take into account the prescribed boundary conditions.

Different strategies to include the boundary conditions in the reconstructions are tested.

We present a large set of numerical tests to assess the accuracy and stability of the scheme

SYMCOMP 2017
Guimarães, 6-7 April 2017

EXPLORING ACYCLIC NEURAL NETWORKS IN CLASSIFICATION PROBLEMS. WHICH ACTIVATION FUNCTION COULD WE CHOOSE?

Antonio J. Tallón-Ballesteros[1]* and María Rodríguez-Romero[2]

1: Department of Languages and Computer Systems
Higher Technical School of Computer Science Engineering
University of Seville (Spain)
Reina Mercedes Av. 41012-Seville (Spain)
e-mail: atallon@us.es, web: http://www.lsi.us.es

2: Higher Technical School of Computer Science Engineering
University of Seville (Spain)
Reina Mercedes Av. 41012-Seville (Spain)
e-mail: marrodrom8@alum.us.es, web: http://www.eii.us.es

Keywords: Supervised Machine Learning, Data Mining, Hyperbolic tangent units, Sigmoid units

Abstract *Multilayer perceptron is one the most extended type of neural networks. In most of the cases, it is composed of hidden nodes whose activation function is a sigmoid one. The typical back propagation (BP) algorithm considers the number of epochs, the learning rate and the momentum as the main parameters to train the neural network model. This paper studies other alternatives such as hybrid neural networks where two kinds of activation functions are included at different levels of the neural network or other variations of the classical BP procedure. The empirical analysis is conducted on medium-size supervised machine learning problems. The research shed light on that the hyperbolic tangent units although not having being used extensively in the previous research are also a very powerful way to deal with pattern recognition problems.*

REFERENCES

[1] Bishop, C.M. *Neural networks for pattern recognition*, Oxford University Press, New York, 1995.

[2] Duda, R.O., Hart, P.E., Stork, D. *Pattern Classification*, second ed., Wiley, 2001

[3] Lippmann, R.P. Pattern classification using neural networks, *IEEE Communications Magazine* Vol. 27, pp. 47-64, 1989.

[4] Tallón-Ballesteros, A.J., Riquelme, J.C. "Data Mining Methods Applied to a Digital Forensics Task for Supervised Machine Learning". *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications*. Studies in Computational Intelligence (SCI) 555. Springer International Publishing, 413-428, 2014.

SYMCOMP 2017
Guimarães, 6-7 April 2017

NUMERICAL STUDY OF THE BLOOD FLOW IN THE RENAL ARTERIES

Daniela A. L. Martins[1]*, Pedro A. M. Lobarinhas[2] and Senhorinha F. C. F. Teixeira[3]

1: Department of Mechanical Engineering
School of Engineering
University of Minho
Portugal
e-mail: a75949@alunos.uminho.pt

2: Department of Mechanical Engineering
School of Engineering
University of Minho
Portugal
e-mail: pl@dem.uminho.pt

3: Department of Production and Systems Engineering
School of Engineering
University of Minho
Portugal
e-mail: st@dps.uminho.pt

Keywords: Cardiovascular, Computational Fluids Dynamics, Wall Shear Stress, Velocity

Abstract *The characterization of the blood flow in the arteries is an important strategy to establish a relationship between hemodynamic behavior and the development of diseases in the arterial tree. In the present study, it was analyzed blood flow in the abdominal aorta artery, at the level of renal bifurcation. The blood flow was simulated using a transient velocity profile. For that purpose, a comparison was made between laminar and three different turbulence models, that are important in this type of studies, due to the complexity of the arterial region and the use of transient velocity. The comparison between the different models of turbulence, allowed to study the differences in the results, and which model is more suitable to be applied to the blood flow in the study geometry. By analyzing, the obtained results for the different models under study, it is possible to state that the difference on the velocity fields was relevant to the problem under study. As for the walls shear stresses, as expected, they were larger in the recirculation zones.*

1. INTRODUCTION

Cardiovascular diseases are the leading causes of death, particularly in the developed countries. The circulatory system is very complex, therefore the study of blood flow in the arteries is an important step in understanding the relationship between hemodynamics and the occurrence of the cardiovascular diseases [1].

The composition and structure of the blood, have a prominent role in blood rheology. The plasma, a Newtonian fluid, and a suspension of various cells, such as erythrocytes, leucocytes, and platelets, are the main components of blood. The blood cells, make up about 45% by blood volume, forming a non-Newtonian fluid. The behavior of the cells change shear-thinning viscoelastic behavior, such as aggregation, deformation, and alignment of the erythrocytes.

However, in most CFD (Computational Fluids Dynamics) simulations, blood is modelled as a Newtonian fluid, particularly in large arteries, because the influence of shear-thinning properties is not significant. The viscoelastic properties of blood are also often ignored [2] [3].

The blood flow in the arteries is characterized by variable and unstable nature and these characteristics are often the consequence of vascular system geometry. The fact of the vascular geometry presents complex shapes, as in case of the bifurcations, ramifications, and curvatures, it becomes more favorable for the development of serious diseases, as the case of atherosclerosis, which is one of the more common pathologies [4].

Atherosclerosis is characterized by deposition of plaques in the inner wall of medium and large blood vessels. These deposits within blood vessels, make the inner surface irregular and the lumen narrow, hindering the normal blood flow [5]. This obstruction, in the vessel, is designed as stenosis. In the renal arteries, the existence of stenosis can preclude the blood flow to the kidney and thus active the renin-angiotensin system, which can lead to severe hypertension [6].

Although advances in the diagnosis and treatment of these pathologies are significant, there is still a clear need for more effective and safe therapies. In this way, mathematical models, and computer-assisted simulations open new ways of understanding the complexity of cardiovascular pathologies and their treatment, as well as better surgical planning. Thus, the application of CFD methods is an important strategy to investigate the blood flow in the arteries and consequently the development of the cardiovascular diseases. CFD methods have the capacity to simulate velocity fields, pressures and walls shear stress in the virtual models of the cardiovascular system [7] [8].

The development of a model for CFD studies involves different steps. Firstly, is essential to define the region of interest (in this case, anatomical structures) and to create the computational model of the region. The next step is the generation of a grid adequate at the geometric domain. The following part consist in the of a computational solution and extraction of the relevance hemodynamics information for the problem in study [7].

The purpose of the present study is used FLUENT to simulate the blood flow in the aortic artery, at the level of renal bifurcation, using a transient velocity profile, and making a comparison between the laminar model and different turbulence models. The importance of the use of turbulence models, is due to the complexity of the arterial zone and to the use of transient velocity. In addition, the comparison between the laminar model and the turbulence models will allow to evidence if the use of the turbulence models is relevant in this problem.

2. METHODS

2.1. Governing equations and numerical solutions

The conservation equations, which describe the laws of fluid mechanics, were solved numerically. A three-dimensional mathematical model was used, for an incompressible and Newtonian fluid and the equations for the mass conservation (1) and *momentum* (2) are:

$$\nabla \vec{v} = 0 \tag{1}$$

$$\frac{\partial}{\partial t}(\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \vec{v}) = -\nabla p + \nabla \cdot \bar{\bar{\tau}} \tag{2}$$

Being $\vec{v}$ the vector of the velocity of the fluid, ρ the density, p the static pressure, and $\bar{\bar{\boldsymbol{\tau}}}$ the tensor of the tensions [4], [9].

FLUENT software was used to solve the continuity equations, *momentum* and turbulence. In the finite volume method, these partial differential equations were approximated by a set of algebraic equations on the computational domain that were subsequently solved [4]. The solutions are obtained iteratively using the solver with the SIMPLE algorithm, and the converge was accepted when the residuals reach values in the order of $1e^{-04}$.

2.2 Geometry model and grid generation

The schematic representation of the three-dimensional geometry, considered in this investigation, and the relevant details are shown in figure 1. In this case, an idealized geometry representative of the real geometry, was constructed through literature data.

The bifurcation of the abdominal aorta in the renal arteries was modeled like a rigid wall with a circular cross-section. The renal bifurcation is asymmetric, since that the right and left renal arteries, are not located in the same axial plane and have different bifurcation angles (60º right renal and 65º left renal) [10].

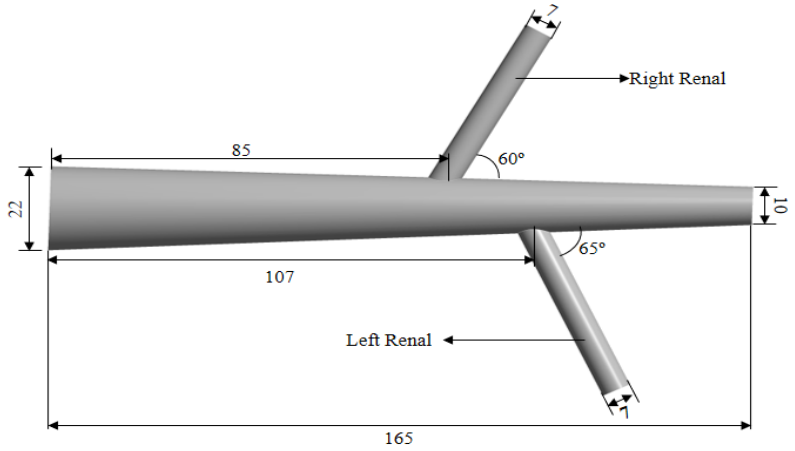


Figure 1. Geometry of the abdominal aorta with arteries renal.

The geometry was created using the SOLIDWORKS software. For the creation of the mesh, the ANSYS 17.1 was used, considering the requirements for this type of the problem. The accuracy of a CFD solution is related to the choice of the type of elements used and cells number used in the mesh.
In this case, the mesh was more refining next to walls due to sliding of the fluid, being also more refined in the renal arteries, since there was a significant decrease of the diameter, which was necessary, so the results obtained does not depend on the mesh type [11]. For that reasons, some tests were carried out with different types of meshes, to verify when the refinement of the mesh, did not influence the obtained results. The grid was created with 248 436 tetrahedral elements (Figure 2).

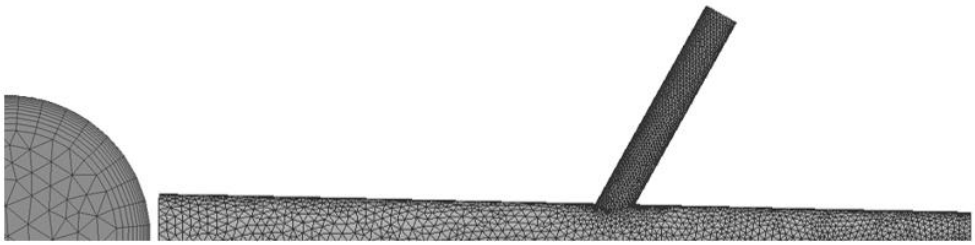


Figure 2. Computational mesh the model, in the inlet and abdominal aorta with renal artery.

2.3. Boundary conditions

The fluid was assumed as a Newtonian and incompressible fluid, with a viscosity of 0.00345 kg/m and a density of 1056 kg/m^3 [10]. The velocity-inlet was adjusted based in the cardiac cycle at the beginning of the aorta artery [12], represented in Figure 3.
Concerning the outlets, the condition of pressure-outlet was defined, where the constant pressure of 11 500 Pa was assumed for the main outlet. For the right and left renal artery, a constant pressure of 11 000 Pa was admitted [1].
The inlet flow is laminar or turbulent. When turbulence was defined, two models were used because it is possible to do a comparison within two turbulence models and the laminar model. The turbulence models used in this study, SST k-ω and Reynolds Stress models. The SST k-ω is a model that has two equations and SST model incorporates a damped cross-diffusion derivative term, which makes this model more accurate and reliable for a wider class of flows. The Reynolds Stress model is a turbulence model more complex with seven equations, generally used in cases where the fluid is more complex.

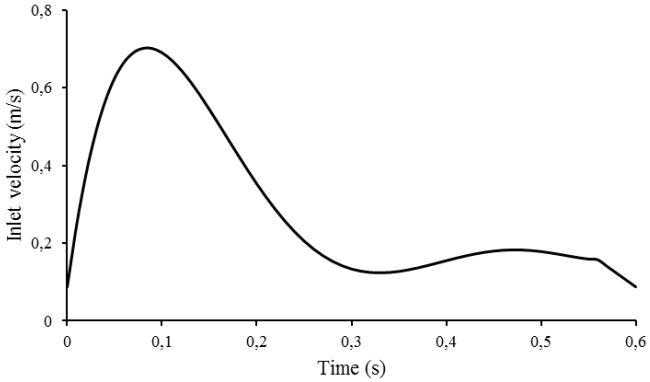


Figure 3. Time variation of inlet velocity.

3. RESULTS

In this section, the results for the velocity distributions and the wall shear stress (WSS), are described at the maximum point of the cardiac cycle. As stated before, the laminar model and different turbulence models are studied.

3.1. Velocity

The velocity distribution, for laminar and turbulence models, are shown in Figure 4. The velocity distribution was analysed over a plane that divides the developed artery geometry into two equal parts (z=0). It was considered a maximum point of the cardiac cycle which corresponds to a systolic peak.

By the analysis and study of the velocity distribution diagrams, on Figure 4, it was verified that doesn't exist considerable differences in the distributions of the velocity on the applied plane. The maximum velocity was reached in all cases, nearby the renal arteries ramifications, that phenomenon may be the result of the reduction in the cross-sectional area. Regarding the maximum velocity, the values have changed from model to model, but the recorded values didn't present significant variations. Notwithstanding, the Reynolds Stress was the model that reached higher maximum velocity, being this 2,55 m/s. In other hand, the model that presented lower maximum velocity was the laminar model, 2,46 m/s.

As expected, it could be noted that the regions of recirculation were on renal branches, being the right renal branch, where the recirculation zone is more prominent.

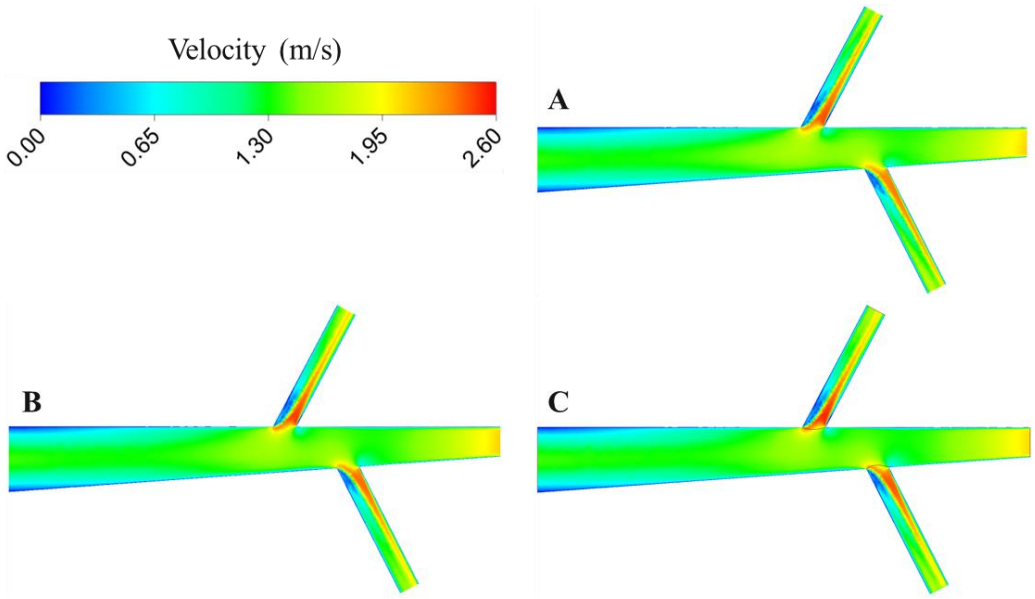


Figure 4. The velocity in plane z=0 in different models in the peak maximum of the cycle cardiac: A- Laminar model; B- SST k-ω model; C- Reynolds Stress model.

3.2. Wall Shear Stress (WSS)

WSS perform a critical role in the vessel wall behaviour, and their variations are associated with atherosclerosis appearance. For the study of the WSS in the renal arteries it was necessary to define two paths on the wall of the right renal artery, those planes were created from the medial plane. The paths and corresponding directions are shown in Figure 5.

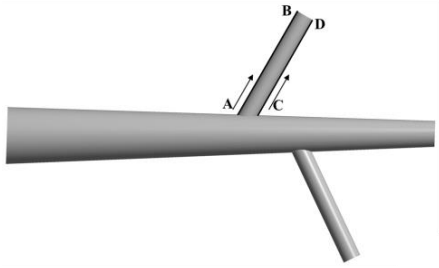


Figure 5. Paths and corresponding directions on the right renal artery wall.

The WSS represented on figure 6, 7 and 8, displays the variations along the defined paths in the different models, on the point that corresponds to the maximum inlet velocity in the cardiac cycle.

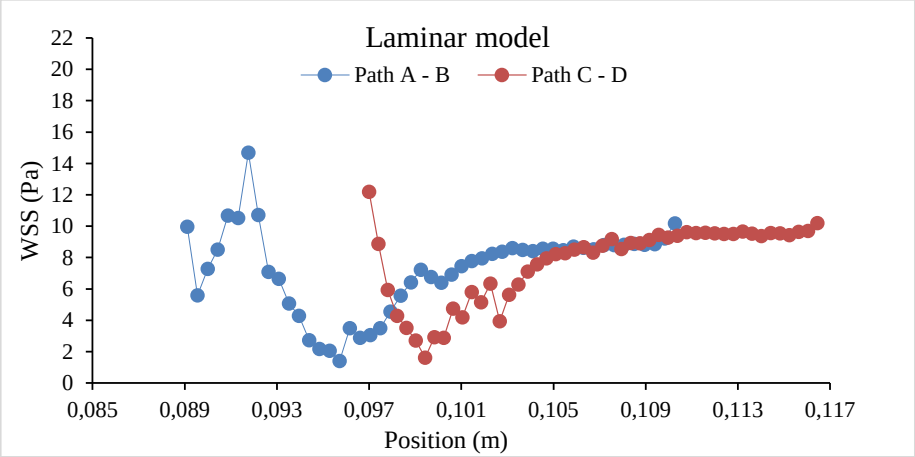


Figure 6. WSS distributions for laminar model.

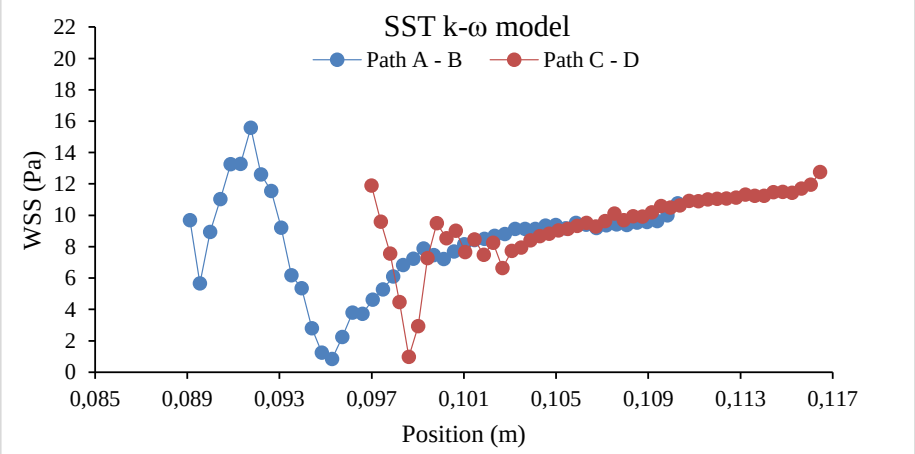


Figure 7. WSS distributions for k-omega model.

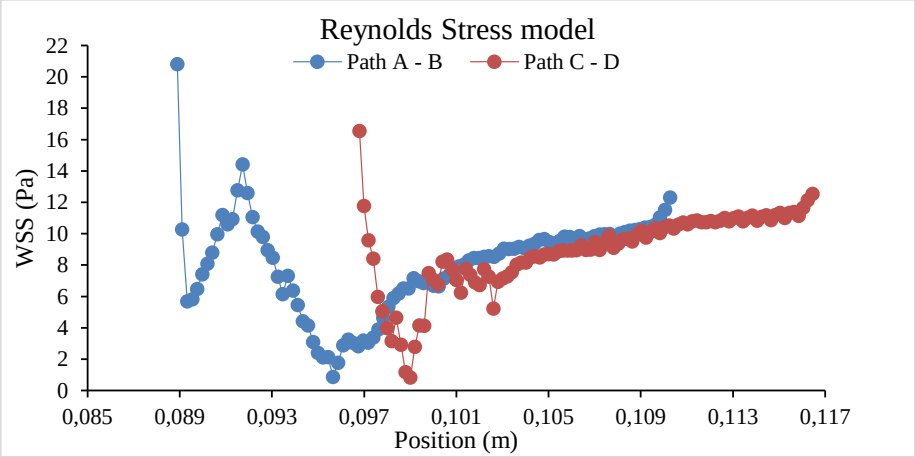


Figure 9. WSS distributions for Reynolds Stress model.

Trough the study of the distribution plots of the WSS on the different models, was verified that on the laminar and k-omega models the variations on the distribution of WSS, also the reached maximums were identical. However, the SST k-ω model still presented a more stable distribution along the paths then the laminar model, which meets the expected as the k-omega model is advisable for studies that consider the geometry walls.
The Reynolds Stress model presents an evolution tendency similar to the aforementioned models, although possesses a higher maximum shear stress on the beginning of the paths.
Besides these models was also used the k-ε model, yet the results for this model are not presented because the behaviour is different of the other models. This behaviour could be related to the inefficiency of the model on the current study, once the equations which characterize the model where developed to evaluate the evolution of the flux on the centre of the geometry.

4. CONCLUSIONS

The obtain results show that although of the differences are not more significant, a comparison between the laminar and turbulence models reveals, that the use of the turbulence models is important, which analyse the velocities and WSS. The STT k-ω model is the more indicate for using in this study.
It was also possible to conclude that the branches angles considered on the geometry have higher impact on the obtained results, namely on the accumulation of WSS forming a recirculation zone. Furthermore, one of the limitations of this study is the use of blood as a Newtonian fluid.

REFERENCES

[1] Z. Mortazavinia, A. Zare, and A. Mehdizadeh, "Effects of renal artery stenosis on realistic model of abdominal aorta and renal arteries incorporating fluid-structure interaction and pulsatile non-Newtonian blood flow," *Appl. Math. Mech. (English Ed.*, vol. 33, no. 2, pp. 165–176, 2012.

[2] J. Y. Moon, D. C. Suh, Y. S. Lee, Y. W. Kim, and J. S. Lee, "Considerations of blood properties, outlet boundary conditions and energy loss approaches in computational fluid dynamics modeling.," *Neurointervention*, vol. 9, no. 1, pp. 1–8, 2014.

[3] F. Y. and M. Y. Gundogdu, "A critical review on blood flow in large arteries ; relevance to blood rheology , viscosity models , and physiologic conditions rheology , viscosity models , and physiologic conditions," *korea-Australia Rheol. J.*, vol. 20, no. 4, pp. 197–211, 2008.

[4] A. Javadzadegan, A. Simmons, and T. Barber, "Spiral blood flow in aorta – renal bifurcation models," *Comput. Methods Biomech. Biomed. Engin.*, vol. 5842, no. November, pp. 1–13, 2015.

[5] C. VanPutte, J. Regan, A. Russo, B. Manager, and A. Reed, *Seeley's Anatomy & Physiology 9th Lab Edition*. 2016.

[6] M. A. R. Mortazavinia Z., Arabi S., "Numerical Investigation of Angulation Effects in Stenosed Renal Arteries," *J. Biomed. Phys. Eng.*, pp. 115–122, 2014.

[7] P. D. Morris *et al.*, "Computational fluid dynamics modelling in cardiovascular medicine.," *Heart*, p. heartjnl-2015-308044-, 2015.

[8] S. N. Doost, D. Ghista, B. Su, L. Zhong, and Y. S. Morsi, "Heart blood flow simulation : a perspective review," *Biomed. Eng. (NY).*, pp. 1–29, 2016.

[9] D. L. Martins, J. C. Pires, A. A. Soares, and L. Morgado, "Hemodinâmica em modelos simplificados da bifurcação da artéria carótida com estenose," pp. 11–12, 2014.

[10] F. Carneiro, A. E. Silva, S. F. C. F. Teixeira, J. C. F. Teixeira, and A. M. Pedro, "The influence of renal branches on the iliac arteries blood flow," pp. 1–7, 2008.

[11] S. F. C. F. Silva, Ana E. S., Lobarinhas, Pedro A. M., Teixeira, "The influence of different grid approaches on a cardiovascular computational model," in *Proceedings of the 17th IASTED International Conference APPLIED SIMULATION AND MODELLING (ASM 2008)*, 2008, pp. 225–228.

[12] E. Soudah, E. Y. K. Ng, T. H. Loong, M. Bordone, U. Pua, and S. Narayanan, "CFD Modelling of Abdominal Aortic Aneurysm on Hemodynamic Loads Using a Realistic Geometry with CT," *Hindawi Comput. Math. Methods Med.*, vol. 2013, 2013.

SYMCOMP 2017
Guimarães, 6-7 April 2017

MODELS ON VARIATIONAL METHODS FOR IMAGE PROCESSING

J. A. Rodrigues

Área Departamental de Matemática
ISEL Instituto Superior de Engenharia de Lisboa,
Instituto Politécnico de Lisboa
Rua Conselheiro Emídio Navarro,
1959-007 Lisboa, Portugal

Keywords: Variational method, Finite element method, Image processing

Abstract. *Variational image-processing models offer high-quality processing capabilities for imaging. They have been widely developed and used in the last two decades, enriching the fields of mathematics as well as information science. Mathematically, several tools are needed: energy optimization, regularization, partial differential equations, level set functions, and numerical algorithms. With this work we analyze some variational image-processing models.*

1 INTRODUCTION

Image processing and image analysis refers to some aspects of the process of computing with images. This process has been made possible by the advent of computers powerful enough to cope large image data.

Image restoration is an important challenging inverse problem in image analysis. The problem consists in reconstructing an image u from a data f. Energy minimization method has demonstrate to be a powerful approach to handle this kind of problems image restoration problems are usually severely ill posed and a Tykhonov-like regularization process is needed. The foundations of these approximation methods were laid by Tikhonov in 1943, when he generalized the classical definition of well-posedness (this generalization is now commonly referred to as conditional well-posedness). The aim of this technique is to specify a set of correctness on which it is known a priori that the considered problem has a unique solution. In 1963, Tikhonov [10] and [11] suggested what is nowadays commonly referred to as Tikhonov regularization. The abstract setting of regularization methods presented there already contains all of the variational methods that are popular in imaging.

Here we presents the a first order variational model, the so-called Rudin- Osher-Fatemi model.

2 VARIATIONAL MODELS PRINCIPLE

The general form of variational models consists in the minimization of an energy functional :

$$\mathcal{F}(u) = \|u - f\|_X + \mathcal{R}(u), \quad u \in X, \tag{1}$$

where X is a (real) Banach space, $\mathcal{R}$ is a regularization operator f is the observed image and u is the image to recover. The first term is the fitting data term and the second one permits to get a problem which is no longer ill posed via a regularization process.

We can illustrate the regularization operator effect with the following case:

Let Ω be an open bounded subset of $\mathbb{R}^2$ and $X = H^2(\Omega)$ and $L^2(\Omega)$ the usual Sobolev spaces (see [2]) endowed with the usual norms:

$$\|v\|_2 = \|v\|_{L^2(\Omega)} \text{ and } \|v\|_X^2 = \|v\|_2^2 + \|\nabla v\|_2^2, \ v \in X. \tag{2}$$

We consider the following fitting data problem :

$$(\mathcal{P}) \begin{cases} \text{Find } u \in X \text{ such that} \\ \\ \|u - f\|_2^2 = \min\limits_{v \in X} \|v - f\|_2^2 \end{cases} \tag{3}$$

where $f \in L^2(\Omega)$. It is easy to see that the functional

$$v \mapsto \|v - f\|_2^2 \tag{4}$$

is not coercive on X.

For example:

if $\Omega =]0,1[$, $u_n(t) = t^n$ and $f = 0$. Then $\|u_n\|_2 = \dfrac{1}{\sqrt{2n}}$ and $\|u_n^{'}\|_2 = \dfrac{n}{\sqrt{2n-1}}$.

So

$$\lim n\|u_n\|_X = +\infty \text{ and } \lim n\|u_n\|_2 = 0 \tag{5}$$

Therefore, we do not even know if $(\mathcal{P})$ has (at least) a solution. this fact take us to define the regularized problem as

$$(\mathcal{P}_\alpha) \begin{cases} \text{Find } u \in X \text{ such that} \\ \\ \|u - f\|_2^2 + \|\nabla u\|_2^2 = \min\limits_{v \in X} \|v - f\|_2^2 + \|\nabla v\|_2^2 \end{cases} \tag{6}$$

with $\alpha > 0$. This new minimization problem, allow us to obtain u which fit the data f, but ask for the gradient to be small.

We can prove the following result (see [7]):

Theorem 1 *For every $\alpha > 0$, problem $(\mathcal{P}_\alpha)$ has a unique solution u_α. Moreover, assuming that $(\mathcal{P})$ has at least a solution, then one can extract a subsequence of the family u_α that weakly converges in X to a solution u of $(\mathcal{P})$ as $\alpha \to 0$.*

We remark that problem $(\mathcal{P}_\alpha)$ has a unique solution u_α because the functional

$$v \mapsto \mathcal{J}_\alpha(v) = \|v - f\|_2^2 + \|\nabla v\|_2^2 \tag{7}$$

is coercive, continuous and strictly convex.

We want use Finite element Method to compute u_α numerically. As $\mathcal{J}_\alpha$ is strictly convex, u_α satisfies the necessary and sufficient optimal condition

$$\mathcal{J}_\alpha^{'}(u_\alpha) = 0. \tag{8}$$

A classical calculus, using Green's formula and a null Dirichlet condition to u_α, gives

$$\begin{aligned} \frac{1}{2}\mathcal{J}_\alpha^{'}(u_\alpha) \cdot v &= \int_\Omega (u_\alpha(x) - f(x))\, v(x)dx + \int_\Omega \nabla u_\alpha(x) \nabla v(x) dx \\ &= \int_\Omega (u_\alpha(x) - f(x) - \Delta u_\alpha(x))\, v(x)dx \end{aligned} \tag{9}$$

for all $v \in X$

Thus, the solution u_α satisfies the Euler-Lagrange equation:

$$u - f - \Delta u = 0,\ u \in H_0^1(\Omega) \tag{10}$$

This is the most simple regularization term (Tykhonov regularization), however is not well adapted to image restoration: the reconstructed image is too smoothed because the Laplacian is an isotropic diffusion operator. In particular, the edges are degraded which is not acceptable to perform a good segmentation.

3 THE RUDIN-OSHER-FATEMI MODEL

A better approach is the use of a regularization term that preserves contours. This implies to deal with functions that can be discontinuous (the jump-set describes the contours). Such functions cannot belong to $H^1(\Omega)$ any longer since their distributional derivative may be Dirac measures.

The most famous model is the Rudin-Osher-Fatemi denoising model (see [1] and [8]). This model involves a regularization term that preserves the solution discontinuities, what a classical H^1-Tykhonov regularization method does not. This model relies the assumption that the space of bounded variation functions ($BV(\Omega)$) is a good space to study images The observed image to recover is splitted in two parts $f = f_s + f_n$ where f_n represents the oscillating component (noise or texture) and f_s is the smooth part. So we look for the solution as $u = u_s + u_n$ with $u_s \in X = BV(\Omega)$ and $u_n \in L^2(\Omega)$. The regularization term involves only the smooth component u_s, while the remainder term $u_n = fu_s$ represents the noise to be minimized. We get

$$(\mathcal{P}_1)\begin{cases} \text{Find } u_s \in X \text{ such that} \\ \\ \mathcal{F}_1(u_s) = \min\limits_{v \in X} \mathcal{F}_1(v) \end{cases} \tag{11}$$

where

$$\mathcal{F}_1(v) = \frac{1}{2}\|v - f\|_2^2 + \lambda \Phi(v) \tag{12}$$

$\Phi(v)$ is the total variation of v and $\lambda > 0$.

Theorem 2 *Theorem 2.3. Problem $\mathcal{P}_1$ has a unique solution in X.*

4 THE MEYER MODEL

In [6], Y. Meyer shows some limitations of the above model. In particular, if f is the characteristic function of a bounded domain, with a regular boundary the, then f is not preserved by the Rudin-Osher-Fatemi model. So Meyer propose a new model

$$(\mathcal{P}_2)\begin{cases} \text{Find } u_s \in X \text{ and } u_n \in Y \text{ such that} \\ \\ \mathcal{F}_2(u_s, u_n) = \min\limits_{v \in X} \mathcal{F}_2(u, v) \\ \\ \text{subject to } f = u_s + u_n \end{cases} \tag{13}$$

Here the solution is still splitted as $u = u_s + u_n$.
Where

$$\mathcal{F}_2(u, v) = \|v\|_y + \lambda \|\nabla u\|_2. \tag{14}$$

$X = BV(\Omega)$, $Y = G(\mathbb{R}^2)$, the Banach space of the signals with large oscillations and thus in particular textures and noises. and $\lambda > 0$.

5 ACKNOWLEDGEMENTS

This research was supported by project V2MIP, IDI& CA-IPL 2017.

REFERENCES

[1] Acar, R., Vogel, C. "Analysis of bounded variation penalty methods for ill-posed problems". Inverse Problems 10 6 pp 1217-1229, 1994.

[2] Adams, R. A. Sobolev spaces. Academic Press, Springer Verlag, 1978.

[3] Brezis, H. Analyse Fonctionnelle. Masson, Paris, 1987.

[4] Esedoglu, S., Shen, J.., "Digital image inpainting by Mumford-Shah-Euler model", *European Journal of Applied Mathematics*, Vol. **13**, pp. 353-370, 2002.

[5] Hecht, F., "New development in freefem++," *Journal of Numerical Mathematics*, Vol. **20**, pp. 251-265, 2002.

[6] Meyer, Y., "Oscillating patterns in image processing and in some nonlinear evolution equations". The Fifteenth Dean Jacquelines B. Lewis Memorial Lectures, March ,2001.

[7] Osher, S., Sole, A., L., V. "Image decomposition and restoration using total variation minimization and the H^1 norm". SIAM Journal on Multiscale Modeling and Simulation 1-3, pp 349370, 2003.

[8] Osher, S., Fatemi, E., Rudin, L., "Nonlinear total variation based noise removal algorithms". Physica D 60, pp 259268, 1992.

[9] Rudin, L., Osher, S., Fatemi, E., "Nonlinear total variation based noise removal algorithms", *Physica*, Vol. **60**, pp. 259-268, 1992.

[10] Tikhonov, A. N.,"Regularization of incorrectly posed problems", *Soviet Math. Dokl.*, **4**, pp. 1624-1627, 1963.

[11] Tikhonov, A. N.,"Solution of incorrectly formulated problems and the regularization methods", *Soviet Math. Dokl.*, **4**, pp. 1035-1038, 1963.

SYMCOMP 2017
Guimarães, 6-7 April 2017

NUMERICAL APPROACH OF SOME DELAYED-AVANCED DIFFERENTIAL EQUATIONS

M. Filomena Teodoro[1,2*]

1: CEMAT, Center for Computational and Stochastic Mathematics
Instituto Superior Técnico
Lisbon University
Avenida Rovisco Pais, 1, 1048-001 Lisboa, Portugal

2: CINAV, Naval Research Center
Portuguese Naval Academy
Portuguese Navy
Base Naval de Lisboa
Alfeite, 1910-001 Almada, Portugal
e-mail: maria.alves.teodoro@marinha.pt

Keywords: Mixed-type functional differential equations, non-linear equations, numerical approximation, method of steps

Abstract. *The mixed type functional differential equations (MTFDEs), equations with both delayed and advanced arguments, appear in many mathematical models in several contexts of applied sciences such as biology, physics, economy, control, acoustics A brief overview about the solution approximation of some advanced-retarded differential equations will be given. In some cases, the analysis of delay differential equations can be extended to MTFDEs. Algorithms relatively to some non-linear advanced-retarded equations have been carried out and will be presented.*

1 Introduction

In applied sciences, many mathematical models show up functional differential equations with delayed and advanced arguments, the mixed type functional differential equations.

Functional differential equations with delay-advanced argument appear in a wide array of different areas of knowledge such as optimal control [15, 16], economic dynamics [17], nerve conduction [2, 3, 4, 11] and traveling waves in a spatial lattice [1, 14].

We are particularly interested in the numerical approximation of the a multi-delay-advance differential equation form (1)

$$x'(t) = F(t, x(t), x(t-\tau_1), \ldots, x(t-\tau_n)), \tag{1}$$

where the shifts τ_i may take negative or positive values.

Some recent numerical methods to approximate the solution of a linear MTFDE with form (2) were introduced in [6, 7] and improved in [9, 18, 19]. More recently, these methods were adapted and used to solve numerically a nonlinear MTFDE [11, 20], the FitzHugh-Nagumo equation.

$$x'(t) = F(t, x(t), x(t-\tau), x(t+\tau)), \quad \tau > 0. \tag{2}$$

Actually, we pretend to calculate numerical solutions of a particular case of a nonlinear MTFDE which models the vibration of some elastics tissues, by the interaction of a flowing fluid (air, blood,...) with an elastic structure tissue, denominated aeroelastic oscillatory phenomena (AOP). The AOP occurs frequently in physiology. Particularly, the considered model characterizes the oscillation superficial wave propagating through the tissues in the direction of the flow. This model was initially proposed by the author of [23, 24]. He proposed a mucosal wave model where a surface wave represents the motion of vocal tissues. Later, some variants of the this model where introduced in several studies in phonation dynamics. In [21], a preliminary approach was introduced where a numerical scheme was adapted from algorithms resulting from earlier work in [11, 20]. In this article, we extend the previous results using COL, FEM, method of steps (MS) and Newton method (NM) to obtain the approximate solution of the mucosal wave model.

2 The problem

We intend to compute the numerical solution of an equation from acoustics which is associated to mucosal wave model of the vocal oscillation during phonation.

2.1 Basic principles

Before proceed, it becomes necessary to introduce some basic principles of vocal fold vibration.

As described in [23], the first principle is that vocal fold oscillation is flow induced. The glottal airstream and the yielding duct wall, the vocal folds, consist in a mechanical system

which can show some instability under some flow conditions. In this situation, a transfer of energy from the glottal airstream to the tissue will overcome frictional energy losses. The range of oscillation is determined by combination of inertial and elastic properties (mass and stiffness) and geometry of vocal folds. When the net aerodynamic driving force as a component in phase with tissue velocity, it is realized a positive flow of energy from the airstream. If we take a system which consists in a mass-spring oscillator

$$M\xi'' + B\xi' + K\xi = f(\xi', \xi, t) \tag{3}$$

where t is the time, M, B and K are mass, damping and stiffness, respectively, ξ, ξ' and ξ'' are displacement, velocity and acceleration, f the diving force. The situation of interest is when we get an autonomous differential equation. It occurs when f do not depend on time. It means that the system oscillates by itself. Another important issue for oscillation is the way how f is related with ξ'. If f and ξ' have the save direction energy is transmitted to the mass, in opposite situation, energy is taken out the mass.

In [8], Libermann describes some of the ways so the glottal airstream can provide a driving force which depends on velocity. In some way, the system needs to change the effective driving force on alternate cycles. The same force sucks the vocal folds together prior to closure works and invert direction so it can cancel partially the impulse resulting from prior to closure.

This process of reverting the driving force direction is done using different mechanisms which can be simultaneous, such as deforming the glottal geometry so can exist different intraglottal pressure distributions or making use of the oppositely phased the supraglottal and subglottal pressures.

2.2 The model

Returning to our objective, we intend to compute the numerical solution of the equation which is associated to mucosal wave model of the vocal oscillation during phonation.

This model was proposed by Titze in [24].

It is assumed left-right symmetry and the motion of tissues is done in the horizontal direction. It can be shown that the wave propagates through the superficial tissues, in the upward direction of the airflow. In the simplest case, these waves can be represented using an one-dimensional wave equation with wave velocity c:

$$\frac{\partial^2 \xi}{\partial t^2} = c^2 \frac{\partial^2 \xi}{\partial z^2}. \tag{4}$$

The solution of (4) is given by the general d'Alembert solution. The expression of tissue displacement is given by (5)

$$\xi(z, t) = x(t - \frac{z}{c}), \tag{5}$$

where $x(t) = \xi(0, t)$ is the tissue displacement at midpoint of glottis. We can notice from (5) that the propagation of mucosal wave causes a time delay from bottom to top of vocal

fold. Titze verifies in [24] that this delay helps to get some necessary instabilities for the oscillation of vocal fold.

If the prephonatory glottis has a linear (trapezoidal) dependence, we get

$$\xi_0(z) = \frac{(\xi_{01} + \xi_{02})}{2} - (\xi_{01} - \xi_{02})\frac{z}{T}, \tag{6}$$

where ξ_{01} and ξ_{02} are the inferior and superior glottal half widths; T is the vocal fold thickness.

Geometrically, it is assumed a very simple case, where the vocal fold width is constant along glottis when in rest position.

The mathematical model which describes the displacement of tissue is obtained imposing the following assumptions:

(i) The pressure at exit of glottis (P_g) equals the atmospheric pressure;

(ii) The sub-glottal pressure equals the lung pressure (P_l);

(iii) The air flow is incompressible, frictionless and stationary;

(iv) The glottis is open.

$x(t)$ is the displacement of tissues at the midpoint of the glottis, so we can get the equation of motion, a nonlinear MTFDE with deviating arguments, with the form (7)

$$Mx''(t) + Bx'(t) + Kx(t) = P_g \tag{7}$$

or in form (8)

$$Mx''(t) + Bx'(t) + Kx(t) = \frac{P_L}{k_t}\frac{x(t-\tau) - x(t+\tau)}{x_0 + x(t+\tau)}, \tag{8}$$

where $x_0 + x(t+\tau) > 0$.

The parameters M, B, K, are, respectively, the effective mass, damping and stiffness per area unit of vocal fold medial surface. The model (8) is also applied in other physiological systems such as avian syrinx, snore, or a flow passing a constricted channel (artery, lips, soft palate, nostrils).

Equation (8) can be transformed in a non-dimensional model after an adequate change of variable introduced in [12], $u = x/x_0$,

$$u''(t) + \alpha u'(t) + \omega^2 u(t) = p\frac{u(t-\tau) - u(t+\tau)}{1 + u(t+\tau)}, \tag{9}$$

where $1 + u(t+\tau) > 0$, $p = \frac{P_L}{k_t x_0 M}$, $\alpha = B/M$ and $\omega = \sqrt{K/M}$.

The model (9) can also be represented as a bidimensional form

$$\begin{cases} u'(t) = v(t), \\ v'(t) = -\alpha v(t) - \omega^2 u(t) + p\frac{u(t-\tau)-u(t+\tau)}{1+u(t+\tau)}, \end{cases} \tag{10}$$

where $1 + u(t + \tau) > 0$.

As a brief note, an advantage of the rectangular glottis configuration is that it can clarify the importance of prephonatory glottal width x_0 for oscillation threshold through relation (11),

$$P_L = \frac{k_t}{T} B_c x_0. \tag{11}$$

The closer the vocal folds are brought together, the easier is to begin of small amplitude oscillation.

Another detail is that Titze [24] assumes small values of τ.

When we consider that mucosal wave has a small time delay [24], equation (9) becomes an autonomous ordinary differential equation, analytically solvable for some values of parameters. This model is similar to the one introduced in [5]. The author of [12, 13] considered a more realistic issue: an arbitrary time delay for mucosal wave.

3 Numerical methods

Some preliminary work introduced different computational methods to numerical solution of autonomous and non-autonomous linear MTFDEs (1) with symmetric delay and advance, using collocation (COLL), least squares and finite element method (FEM), presented in [9, 10, 18]. In [11, 20, 22], it was took into account the numerical solution of a nonlinear MTFDE with deviating arguments arising from nerve conduction theory, taking the form (2).

In particular, to solve numerically (9), it is developed a numerical scheme based on work presented in [9, 10, 22]. Before proceed, we notice that (9) is a second order equation, so we can use the bi-dimensional formulation (10), and adapt the method of steps presented in [22] to system (10).

The main goal of this section is to extend to the bi-dimensional case the formula (7) presented in Section 2 of [9]. It is based on Bellman's method of steps for differential equations. In the linear case [9], one solves the equation over successive intervals of unitary length. In the case of [22], the equation is solved for successive intervals of length τ. Doing some algebraic manipulation and simplification, we get the formula (12)

$$u(t + \tau) = -p_n(t)(u''(t) + \alpha u') + u(t - \tau) + g(u(t)), \quad t \in \mathbb{R} \tag{12}$$

where $g(u(t)) = -p_n(t)\omega^2 u(t)$ and $p_n(t)$ is a polynomial function with order $n \in \mathbb{N}$.

Supposing that all the derivatives of u exist in $(a - 2\tau, a]$, in order to simplify the calculations, we can use the simpler formula (12) to extend the solution for equation (9)

on an interval $[a, a+k\tau]$ (where k is an integer and a some adequate value), starting from its initial values on $[a-2\tau, a]$; these starting values are calculated using the solution of equation (9) taking into account the small amplitude approximation, which can be found in formula (25) of [24]. After some computation, we may obtain explicitly the expressions for the solution successively in intervals $(a, a+\tau]$, $(a+\tau, a+2\tau)$, ... starting with (13)

$$u(t+\tau) \quad = -p_n(t)(u''(t)+\alpha u'(t)) + u(t-\tau) + g(u(t)), \quad t \in (a, a+\tau]. \tag{13}$$

Using this process, we can extend the solution to any interval, provided that the initial functions in the first two intervals with length τ are smooth enough functions and satisfy some simple relationships.

The problem is reduced to a BVP on a limited interval, using the solution of equation (9) under the approximation proposed by Titze (small amplitude delay) in formula (25) of [24] as a boundary function. A numerical solution of the problem (9) subject to some natural constrains is computed.

The nonlinear problem can be reduced to a sequence of linear problems by means of the NM. Can be found in [22] a detailed description about the NM iterative process. In order to enable the convergence of the Newton iteration process, we consider different values of a set of parameters. We also impose regularity conditions and boundary conditions. The system and all parameters can be updated for each iterate of the NM. The values of u in this system are computed, using the MS, and assuming that u satisfies the boundary conditions. Then we can define u on a specific limited interval and extend it to the closest intervals using a recurrence formula (13).

The numerical schemes described here are generalizations from the algorithm presented in [11, 19, 22], using a uniform mesh. Once the boundary functions can be defined using the Titze approach, we are able to apply the more recent approaches and techniques using the adapted method of steps (MS) for the nonlinear case (9). Using MS, we can extend the solution to any interval, provided that the initial functions in the first two intervals with a specific length (τ) are smooth enough functions and satisfy some simple relationships.

4 Numerical results

The parameters of model were chosen accordingly with [24], pages 1548. In table 4 are presented the absolute error ϵ_N (2-norm) and the estimated order of convergence $p = log_2\epsilon_{2N}/log_2\epsilon_N$ of approximate solution of (9) by COLL, when it is considered an uniform mesh, $x_0 = 0.04\ cm$ and $x_0 = 0.16\ cm$.

In table 4 are also computed the absolute error and the estimated order of convergence of approximate solution of (9), but using FEM, with $x_0 = 0.04\ cm$ and $x_0 = 0.16\ cm$. At each table, the results are accurate.

When it is applied the COL method, the absolute error is of order 2×10^{-4} for both set of parameters, with a partition of 128 sub-intervals. The estimated order of convergence p is compatible with the expected one, $p \approx 2$.

N	$\epsilon^{(1)}$	$p^{(1)}$	$\epsilon^{(2)}$	$p^{(2)}$
16	$1.26e-2$		$1.00e-2$	
32	$3.15e-3$	1.98	$2.50e-3$	1.99
64	$7.86e-4$	1.99	$6.06e-4$	1.99
128	$1.96e-4$	2.00	$1.50e-4$	2.01

Table 1: Absolute error ϵ and estimated order of convergence p for estimate solution of (9) by COL, using Titze approximation. Parameters defined in [24], page 1548; $^{(1)}$ $x_0 = 0.04$ cm (on left) and$^{(2)}$ $x_0 = 0.16$ cm (on right). Partition size: N subintervals.

n	$\epsilon^{(1)}$	$p^{(1)}$	$\epsilon^{(2)}$	$p^{(2)}$
16	$2.9e-4$	2.00	$8.55e-3$	1.89
32	$7.26e-5$	2.02	$2.18e-4$	1.97
64	$1.80e-5$	2.02	$5.48e-5$	1.99
128	$4.46e-6$	2.09	$1.37e-5$	2.00

Table 2: Absolute error ϵ and estimated order of convergence p for estimate solution of (9) by FEM, using Titze approximation. Parameters defined in [24], page 1548; $^{(1)}$ $x_0 = 0.04$ cm (on left) and$^{(2)}$ $x_0 = 0.16$ cm (on right). Partition size: N subintervals.

By other hand, when we apply the FEM, the results are more accurate. The absolute error is of order 4×10^{-6} for first set of parameters and 1×10^{-5} for the second set of parameter, when we take a partition with 128 sub-intervals. The estimated order of convergence p is lower than the expected one, $p \approx 2$ for the two set of parameters.

5 Final remarks

Our initial proposal was to apply the more recent approaches and techniques using an adapted method of steps (MS) for the nonlinear case (9).

The method introduced previously in [20] and extended from [19, 11], using a numerical scheme based on an adapted method of steps, was rebuilt and re-adapted, using a uniform mesh. Using MS, we could extend the solution to any interval, and provided that the initial functions in the first two intervals were smooth enough.

To solve numerically the equation on study, we consider an issue about a symmetrical system. Two different sets of parameters, the same that Titze used in [24], were tested and, in general, the results obtained by COLL and by FEM were accurate. The COLL method gave results consistent with the expected order when the convergence is guaranteed. The FEM conduced to order of convergence estimates lower than expected. It is still necessary to test another set of parameter values. As it happens when we consider the linear case, in larger intervals, the numerical solution is less accurate. The computation of numerical solution with a nonuniform mesh by COLL, FEM and finite differences is already completed.

Acknowledgements: This work was supported by Portuguese funds through the Center

of Naval Research (CINAV), Naval Academy, Portuguese Navy, Portugal and the Center for Computational and Stochastic Mathematics (CEMAT), The Portuguese Foundation for Science and Technology (FCT), University of Lisbon, Portugal, project UID/Multi/046-21/2013.

REFERENCES

[1] K. A. Abell, C. E. Elmer, . A. R. Humphries, E. S. Van Vleck, Computation of mixed type functional differential boundary value problems, SIADS - Siam Journal on Applied Dynamical Systems 4 3, 755 (2005)

[2] J. Bell, Behaviour of some models of myelinated axons, IMA journal of mathematics applied in medicine and biology 1, 149 (1984)

[3] J. Bell, C. Cosner, Threshold conditions for a diffusive model of a myelinated axon, Journal of Mathematical Biology 18, 39 (1983)

[4] H. Chi, J. Bell, B. Hassard, Numerical solution of a nonlinear advance-delay-differential equation from nerve conduction, Journal of Mathematical Biology 24, 583 (1986)

[5] K. Ishizaka, M. Matsudaira, Fluid Mechanical Considerations of Vocal Cord Vibration. Speach Commun Res Lab. CA. Monog 8 (1972)

[6] V. Iakovleva and C. Vanegas, On the Solution of differential equations withe delayed and advanced arguments, Electronic Journal of Differential Equation, Conference 13, 57-63, (2005)

[7] N. J. Ford and P. M. Lumb, Mixed-type functional differential equations: a numerical approach, Journal of Computational and Applied Mathematics, 229(2), 471-479 (2009) doi: 10.1016/j.cam.2008.04.016

[8] P. Liebermann, Speech Physiology and Acoustic Phonetics (MacMillan, New York, 1977)

[9] P. M. Lima, M. F. Teodoro, N. J. Ford, P. M. Lumb , Analytical and Numerical Investigation of Mixed Type Functional Differential Equations, Journal of Computational and Applied Mathematics, 234 9, 2732 (2010)

[10] P. M. Lima, M. F. Teodoro, N. J. Ford, P. M. Lumb , Finite Element Solution of a Linear Mixed-Type Functional Differential Equation, Numerical Algorithms 55, 301 (2010)

[11] P. M. Lima, M. F. Teodoro, N. J. Ford, P. M. Lumb , in: S. Pinelas et al. (Ed.) Analysis and Computational Approximation of a Forward-Backward Equation Arising

in Nerve Conduction, International Conference on Differential and Difference Equations with Applications, Oct. 4-8 2011, Ponta Delgada-Azores, Portugal (Springer Proceedings in Mathematics & Statistics, New York 2013), 47, 475

[12] J. C. Lucero, Advanced-Delay Equations for Aerolastics Oscillations in Physiology, Biophysical Reviews and Letters 3 1, 125 (2008)

[13] J. C. Lucero et al., A lumped mucosal wave model of vocal folds revisited: Recent extensions and oscillation hysteresis, The Journal of the Acoustical Society of America 129 3, 1568 (2011).

[14] J. Mallet-Paret, The Global Structure of Traveling Waves in Spattially Discrete Dynamical Systems, Journal of Dynamics and Differential Equations 11 1, 49 (1999)

[15] L. S. Pontryagin, R. V. Gamkreledze, E. F. Mischenko, The mathematical Theory of Optimal Process, (Interscience, New York, 1962)

[16] A. Rustichini, Functional differential equations of mixed type: The linear autonomous case, Journal of Dynamics and Differential Equations 1 2, 121 (1989)

[17] A. Rustichini, Hopf bifurcation for functional differential equations of mixed type, Journal of Dynamics and Differential Equations 1 2, 145 (1989)

[18] M. F. Teodoro, N. J. Ford, P. M. Lima, P. M. Lumb, New approach to the numerical solution of forward-backward equations, Frontiers of Mathematics on China 4 1, 155 (2009)

[19] M. F. Teodoro Computational Methods for Functional Differential Equations with Deviating Arguments, Ph. D. thesis, Instituto Superior Técnico, (Universidade Técnica de Lisboa, Portugal, 2002)

[20] M. F. Teodoro, Numerical approximation of a nonlinear delay-advance functional differential equations by a finite element method, In: T. E. Simos (Ed.), International Conference on Numerical Methods and Applied Mathematics, Sept. 19-25, 2012, Kos, Greece (AIP Proceedings, Melville-NY, 2012) 1479, 406

[21] M. F. Teodoro, Numerical Approximation of a Delay-Advanced Equation from Acoustics, In J. Vigo-Aguiar (Ed.), International Conference on Mathematical Methods on Sciences and Engineering, July 6-10, Rota-Cadiz, Spain, (Proceedings of CMMSE 2015, 2015)

[22] M. F. Teodoro, Numerical Solution of a Forward-Backward Equation from Physiology, submitted.

[23] I. R. Titze, Principles of Voice Production, (Prentice-Hall, Englewood Cliffs, 1994)

[24] I. R. Titze, The Physics of Small Amplitude Oscillation of the Vocal Folds, The Journal of the Acoustical Society of America 83, 1536 (1988)

SYMCOMP 2017
Guimarães, 6-7 April 2017

A FINITE VOLUME METHOD IN THE FRAMEWORK OF PROPER GENERALIZED DECOMPOSITION FOR THE CONVECTION-DIFFUSION-REACTION EQUATION

Ricardo Costa[1]*, João Nóbrega[1]

1: Universidade do Minho, Portugal

e-mail: pg24046@alunos.uminho.pt ; mnobrega@dep.uminho.pt

Keywords: Convection-diffusion-reaction equation, Proper generalized decomposition method, Finite volume method, Manufactured solution method

Abstract *Proper generalized decomposition (PGD) is a model order reduction technique which consists in representing a multi-dimensional function by a finite sums decompositions of functions defined in different dimensions.*

Such approach is an appealing strategy for reducing the computer resources and the calculation costs by reducing drastically the number of degrees of freedom that the functional approximation involve.

In this work we present a PGD formulation for the bi-dimensional steady-state convection-diffusion-reaction equation, and the resulting model is discretized with a finite volume scheme.

An extensive benchmark of numerical tests with manufactured solutions is presented in order to assess the efficiency of the method compared to the classical ones.

SYMCOMP 2017
Guimarães, 6-7 April 2017

MATHEMATICAL MODELS OF CONTROL SYSTEMS

João M. Lima[1]

1: Department of Electronic Engineering and Informatics
Faculty of Sciences and Technologies
University of Algarve
Campus de Gambelas
8005-139 FARO
e-mail: jlima@ualg.pt, web: http://w3.ualg.pt/~jlima/

Keywords: Control Systems, Mathematical Models, differential equations, MATLAB/SIMULINK simulation

Abstract *Automatic Control Systems are each time more important in all industrialised and advanced societies. From small plants to complex space systems, the design of devices that lead the overall system achieve some specified performance measures is a compulsory task. Normally, during the design of control systems the optimization and modelling issues should be solved. This presentation show how useful mathematical models are for analysis and design of control systems.*

No matter the system is linear or nonlinear, early studies of automatic control were based upon solution of differential equations. In spite of numerous techniques have been developed in the past decades, mastering the former approach is compulsory in both plan, theoretical and practical. A set of nth-order differential equations is normally established for modelling a continuous physical system. If the system is defined in discrete time a discrete counterpart is used: set of nth-order difference equations.

The establishment of a model is a normal task no matter our goal is only analyse a plant or tuning its controller. When a mathematical model is stablished the subsequent tasks will be accomplished independent of the nature of the started physical system. This issue is important for whom aims at developing unified methodologies.

Modelling task are definitely needed as the physical system could be inaccessible, for example, in a lab environment.

In this talk some examples of mathematical models are presented. This models are sets of differential equations that describe different physical systems, electrical and mechanical. This unified approach is suitable for building a Matlab/Simulink system and show all the numerical and graphical potentialities of this tool. This useful tool open a wide range of potentialities in terms of analyses and design in time or frequency domains.

1. INTRODUCTION

Control systems are far and wide used in all modern and industrialised societies.
Devices designed to control automatized tasks are present into small plants and large industrial buildings as well. For this reason, in many universities, the study of system theory and control systems is compulsory in many branches of sciences and technologies, such that (but not limited to):

- Electronic
- Chemistry
- Aeronautics
- Mechanical
- Economy
- Politics
- Ecology

We realise that, in spite of the former approach aims at using the control system theory to solve problems in the electrical and mechanical fields, nowadays these techniques are also extended to human sciences fields as economy, politics and ecology. So, we can define system theory as a set of formal techniques that provides a unified representation, analysis and design of control systems independently of its physical or social nature. System theory is then a transversal area of knowledge.
The first approach of the design of a controller takes into account the model of the system to be controlled rather than the physical system itself, this issue has a consequence that an accurate model is needed. In order to analyse a dynamic system, an accurate mathematical model that describes the system must be determined, so, modelling is one of the most important step in control system design [1].
In general, a control problem can be divided into the following steps:

- Specifications: Establishment a set of performance measures to be accomplished by the system.
- Modelling[i]: Formulation of a set of differential equations (or difference equations for discrete systems) that describe the physical system to be controlled.
- Analysis: exploiting the system behaviour under certain conditions.
- Verification: having accomplished the previous step, we should be able to verify if the 1st step (establishment a set of performance measures) is or isn't accomplished.
- Design: if the previous step fails, a controller should be designed.

Now, it is easy to understand that hardware and software tools are compulsory in a normal day of a control engineer.
Modelling, in the context explained in the steps above, is the goal of this paper.
From personal computers to specific applications processors, there are several devices that assist the control design. On the other hand, in a point of view of the software, it is unthinkable

[i] In spite of a set of differential equation could be considered as a model, modelling by itself could be much more than differential or difference equations. Modelling studies, by itself, are subjects that can be teaching in more than a semester.

to build a controller without any mean of simulation. In fact, since the establishment and verification of differential or difference equations to accomplishment the analysis step; or from the test of a controller initial solution to improvement of this solution, an environment of simulation is compulsory. *Matlab* software with *Simulink* graphical facilities [2] is a powerful tool for a control engineering.

Focus on the Modelling step inside the control system design, this paper has the following outline. Section 2 a brief description of systems representation techniques is presented. In section 3 some examples of systems are showed and the corresponding *Matlab* implementation is performed. In section 4 some results produced by the *Matlab*/*Simulink* models are presented. The paper ends with section 5 where some conclusions are pointed out.

2. CONTROL SYSTEM REPRESENTATION

Normal physical systems are usually nonlinear, the linearity property is a simplification of the reality. This simplification could leads to more or less accurate models, it depends on the system and the operating point. However, studies of linear systems have great advantages in the point of view of mathematical tools. So, frequently, linear approximation is accurate and the analysis and design of control systems become much more feasible in terms of mathematical skills.

The difficulty of *Matlab*/*Simulink* implementation doesn't depend on the fact that system be linear or nonlinear.

In this section it is considered continuous Linear Time Invariant Systems (LTI Systems), for simplicity reasons.

Let's think about an arbitrary system that can be described by the following differential equation:

$$\begin{aligned} &a_0 y(t)^{(n)} + a_1 y(t)^{(n-1)} + \ldots + a_{n-2}\ddot{y}(t) + a_{n-1}\dot{y}(t) + a_n y(t) = \ldots \\ &b_0\, \mathrm{u}(t)^{(m)} + b_1\, \mathrm{u}(t)^{(m-1)} + \ldots + b_{m-2}\ddot{u}(t) + b_{m-1}\dot{u}(t) + b_m\, \mathrm{u}(t), \quad n > m \end{aligned} \qquad \textbf{(1)}$$

This model describe a n[th] order continuous-time LTI system, no matter what is its physical nature; the input is $u(t)$ and the output is $y(t)$.

During a normal control design it is useful simulate this system; we can do this using the transfer function or a state space model, as I will show in the following.

Taking the Laplace transform of **(1**) we obtain the transfer function **(2**).

$$\frac{Y(s)}{\mathrm{U}(s)} = \frac{b_0 s^m + b_1 s^{m-1} + \ldots + b_{m-2} s^2 + b_{m-1} s + b_m}{a_0 s^n + a_1 s^{n-1} + \ldots + a_{n-2} s^2 + a_{n-1} s + a_n} \qquad \textbf{(2)}$$

This rational function is an external representation because it describe the system in a point of view of the input $u(t)$ and output $y(t)$.

Depending on which control techniques are more suitable for a given problem, it could be preferable deal with an internal representation, named, state space model representation instead of using the transfer function representation. This representation show the input and output as the previous approach **(2**) together with internal variables named state variables. These

variables (functions of the time) can have physical meaning or not, they can be only abstract mathematical variables. According this generic example an internal representation can be the one presented by the simulation diagram, Figure 1.

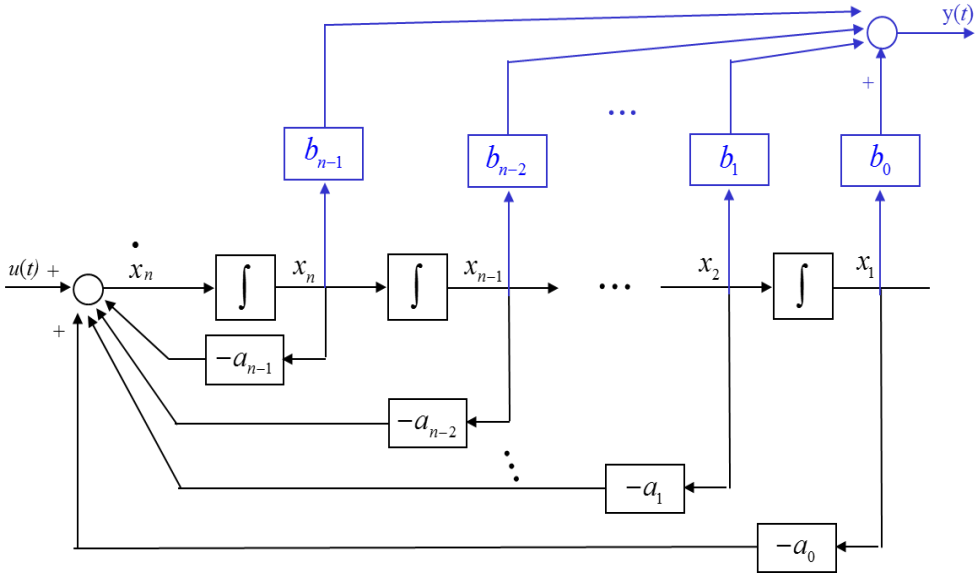


Figure 1. Simulation diagram.

The input $u(t)$ and the output is $y(t)$ are displayed in this internal model together with internal variables. For this n[th] order system these internal variables define the n[th] order state vector $X(t)=\left[x_1(t) \quad x_2(t) \quad \ldots \quad x_n(t)\right]$.

A direct reading of the simulation diagram (Figure 1) allow us to stablish the following dynamic equations **(3)** and **(4)**.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix} . \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u(t) \qquad \textbf{(3)}$$

$$y(t) = \begin{bmatrix} b_0 & b_1 & \cdots & b_{m-1} & b_m & 0 & \cdots & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \\ x_{m+1} \\ x_{m+2} \\ \vdots \\ x_n \end{bmatrix} \quad \textbf{(4)}$$

The par of equations **(3**) and **(4**) can be written in a compact form **(5**).

$$\begin{cases} \dot{X}(t) = AX(t) + BU(t) \\ Y(t) = CX(t) + DU(t) \end{cases} \quad \textbf{(5)}$$

Where *A* is the dynamic matrix $[n \times n]$, *B* is the input matrix $[n \times q]$, *q* is the number of inputs, *C* is the output matrix $[p \times n]$, *p* is the number of outputs, and *D* is a $[p \times q]$ matrix ($D = 0$, for this case).
Matlab/Simulink software is prepared to implement formal descriptions like **(2**) or **(5**).

3. *MATLAB* IMPLEMENTATION

In previous section a general system representation approach was developed. It was shown for continuous-time LTI systems that, transfer function and state space model are alternatives in a point of view of external and internal representations, respectively.
This section aims at show how can we develop a *Matlab* system including a *Simulink* model useful for system simulation.
The translation between the formalism developed into section 2 and *Matlab* software is accomplished using a nonlinear mechanical system: a crane [3]. A simplified model of a crane is depicted in a *x-y* plane, Figure 2.
This simplified model is composed by a mass m_2 moving along the x-axis on a frictionless surface which is connected to the reference by a spring with constant *k*. Another mass m_1 is suspended by means of a light rod of length *l*. The pendulum m_1 is constrained to pivot in the vertical x-y plane. The mass m_2 and the pendulum m_1 displacements are represented by length *x* and angle θ respectively; thus, we can conclude, this simplified model has 2 degrees of freedom.
The current approach demonstrate the powerfulness of the *Matlab* software consists on implementation of 2 versions of this system:

- Nonlinear model
- Linear approximation

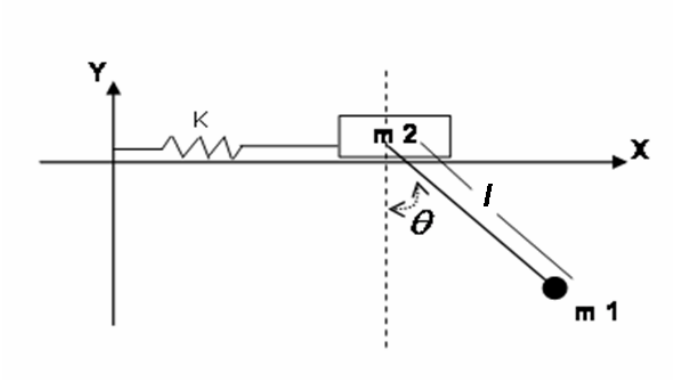


Figure 2. Simplified model of a crane.

As we can see after, both version are equally easy to implement in *Matlab* and the corresponding models are powerfully tools for testing the validity of linear approximation.

3.1. Nonlinear model of the crane

We aim at establish a set of differential equations that describe the crane dynamic, that is, we have to define a Model of the system. According to the Lagrangian formulation, after the evaluation of the kinetic and potential energies for every component of the system [4], and we can obtain the dynamic equations of the motion **(6)**.

$$\begin{cases} (m_1+m_2)\ddot{x}+m_1 l\ddot{\theta}\cos\theta-m_1 l\dot{\theta}^2\sin\theta+kx=0 \\ \ddot{x}\cos\theta+l\ddot{\theta}+g\sin\theta=0 \end{cases} \qquad \textbf{(6)}$$

The nonlinear behaviour of this mechanical system is shown by equations **(6)**; however, building the corresponding Simulink model is an easy task. To do this systematically we first define an internal representation similar of the one presented in **(3)**. However, it should be noted that **(3)** was written for a nonhomogeneous linear system and the crane is nonlinear without any disturbance as input. So, instead of having a linear state equation $\dot{X}(t)=AX(t)+BU(t)$ (see equation **(5)**), we define a non-linear vector field f.

Let's define the state variables set: $\left\{z_1=x,\quad z_2=\dot{x},\quad z_3=\theta,\quad z_4=\dot{\theta}\right\}$, so, the state vector is defined by Z and its first derivative is $\dot{Z}\equiv f(Z)$. Taking into account **(6)** we obtain the state vector first derivative given by **(7)**.

The corresponding *Simulink* model is the direct implementation of the equation **(7)**. This 4th order system is then implemented using 4 integrators whose outputs provide the time evolution of the 4 state variables: $\left\{z_1=x,\quad z_2=\dot{x},\quad z_3=\theta,\quad z_4=\dot{\theta}\right\}$, see Figure 3.

$$\dot{Z} = \begin{bmatrix} z_2 \\ \dfrac{m_1 \sin z_3 \left(g \cos z_3 + l z_4^2 \right) - k z_1}{m_2 + m_1 \sin^2 z_3} \\ z_4 \\ \dfrac{k z_1 \cos z_3 - \dfrac{1}{2} m_1 l z_4^2 \sin\left(2 z_3\right) - g \sin z_3 \left(m_1 + m_2 \right)}{m_2 l + m_1 l \sin^2 z_3} \end{bmatrix} \quad \textbf{(7)}$$

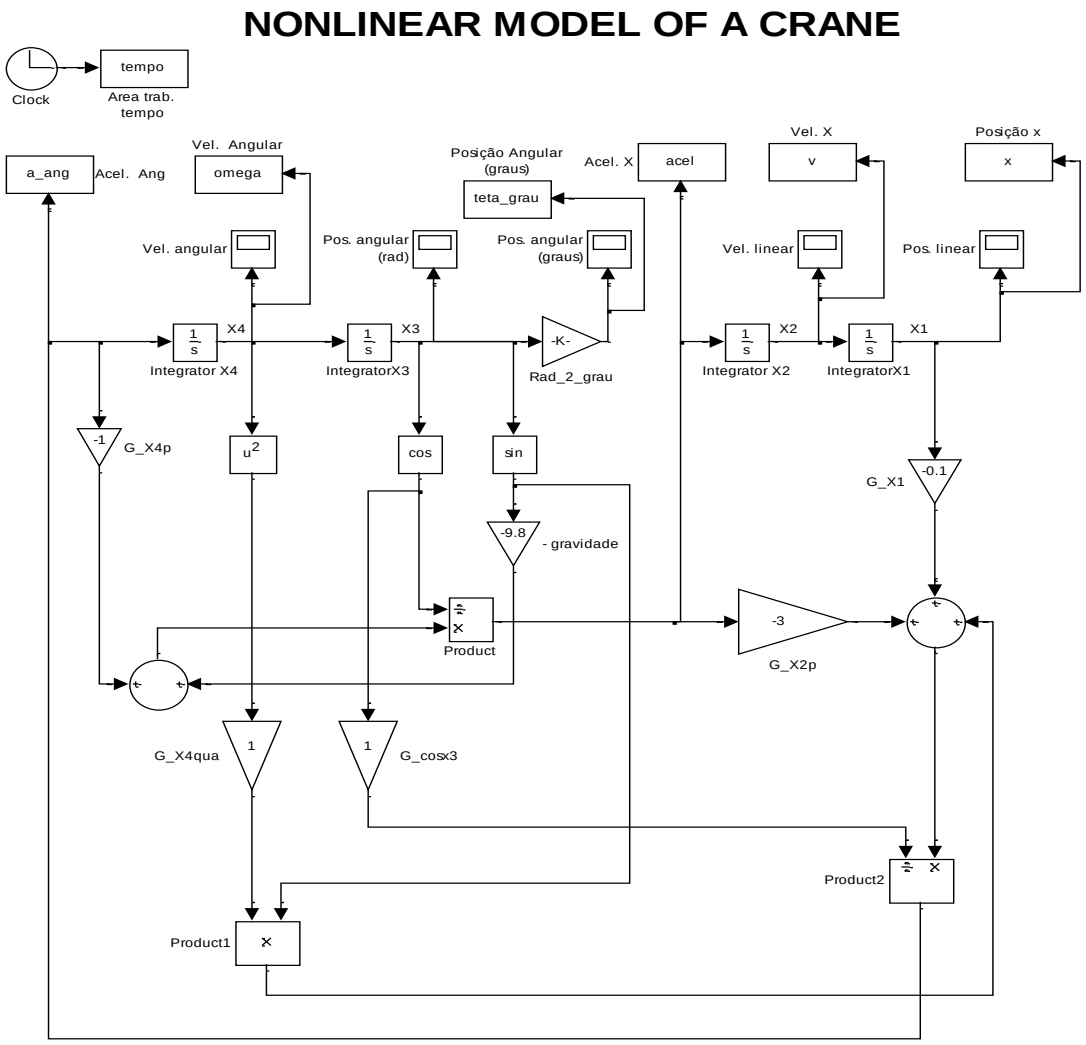


Figure 3. Nonlinear Simulink model of a crane.

Looking the 4 integrators from right to left, we have the corresponding outputs:

1. linear position of the mass m_2 ,
2. speed of the mass m_2 ,
3. angular position of the pendulum m_1
4. angular speed of the pendulum m_1.

This model runs without any external disturbance like applied force, so, the motion can begin by setting to nonzero the initial conditions of the state variables.

3.2. Linear model for the crane

Usually we can assume linearity of many physical systems over a reasonably range of value the variables [5].This example is not an exception; we will look for which the largest values of initial displacements can be used for x and θ that linear model can be considered accurate.

We aim at linearize the *f* functions as presented in **(7)**: $\dot{Z} \equiv f\left(Z\right)$. To accomplish this we will use the linear term of the vector field *f* expanded in Taylor series around the operating point z_0. This approach leads to a linearized model, $\dot{Z} = AZ$, where *A* matrix is a partial derivative matrix defined by **(8**) and $n = 4$ because we deal with a 4th order system.

$$A \equiv F_Z = \left[\begin{array}{ccc} \frac{\partial f_1}{\partial z_1} & \cdots & \frac{\partial f_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial z_1} & \cdots & \frac{\partial f_n}{\partial z_n} \end{array}\right]_{z=z_0} \qquad \textbf{(8)}$$

The operating point z_0 is defined as an equilibrium point: the state vector remains constant, $\left(\dot{Z} = 0\right)$.

Taking into account **(7**) we choose $Z_0 : z_1 = z_2 = z_3 = z_4 = 0$, that is, both m_1 and m_2 are stopped, spring is unstressed and the pendulum is vertical. So, using **(7**) and **(8**) we can evaluate the dynamic matrix of linearized system *A* given by **(9**).

$$A = \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ \frac{-k}{m_2} & 0 & \frac{gm_1}{m_2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{k}{m_2 l} & 0 & \frac{-g\left(m_1 + m_2\right)}{m_2 l} & 0 \end{array}\right] \qquad \textbf{(9)}$$

In an analogous manner to what happened with the nonlinear version (equation **(7**)), the linear approach now obtained **(9**) is equally easy to implement into the *Simulink* environment as presented in Figure 4.

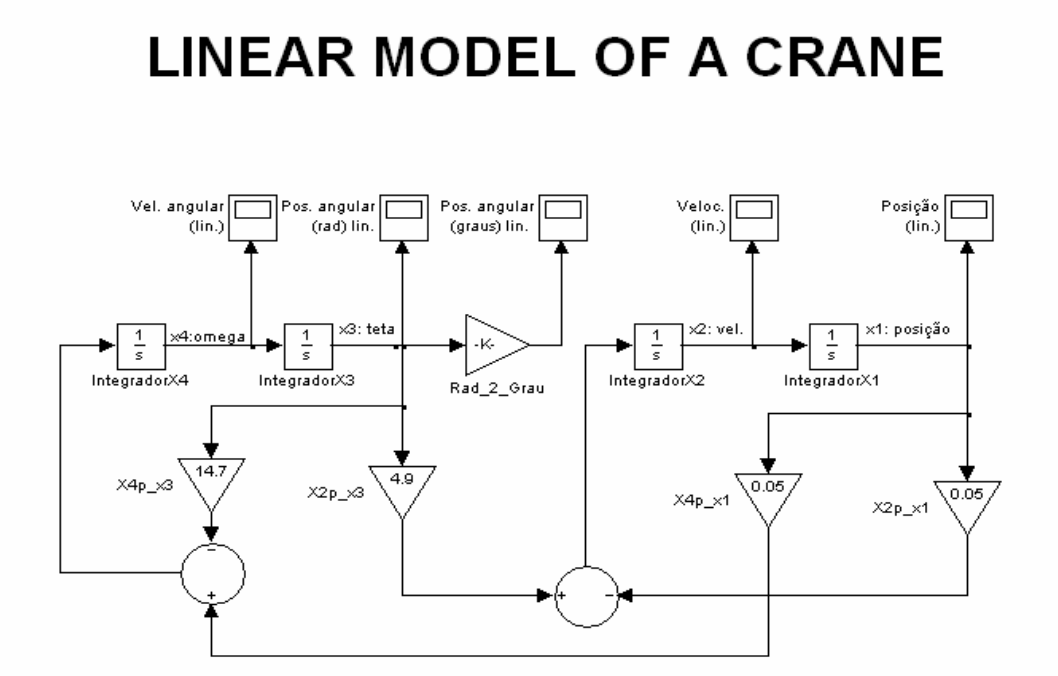


Figure 4. Linear Simulink model of a crane.

As was it happens with the nonlinear version, the linear model is also built with 4 integrators because we deal with a 4th order system. The linear model has less components then the nonlinear one, as expected.

This linear Simulink model together with the nonlinear model presented before is accommodated into the main software developed for comparison purpose. The main program is prepared for running only one model (linear or nonlinear) or both. Several sets of system parameters were used for testing the linear approximation. For each set, different initial conditions were experimented. Some examples of time evolution of the physical variables will be presented into the next session.

4. *MATLAB* AND *SIMULINK* RESULTS

Matlab system was tested using different sets of the crane parameters: $\{m_1, \quad m_2, \quad k, \quad l\}$.

For each set, different initial conditions were used to produce motion. The time evolution of the state variables was also visualized and analysed. *Matlab* code was developed to accommodate these test facilities.

In which concern nonlinear model, Figure 5 shows a normal simulation over 40*s*.

The set of parameters is $m_1 = 1Kg$, $m_2 = 2Kg$, $l = 1m$ and $k = 0.1Nm^{-1}$; the motion is produced by the initial angle of the pendulum be $\theta_0 = 45^o$ (the spring is initially unstressed, $x_0 = 0m$).

We can see the time evolution of the 4 state variables: position x of the mass m_2, angular position θ of the pendulum and them first derivatives.

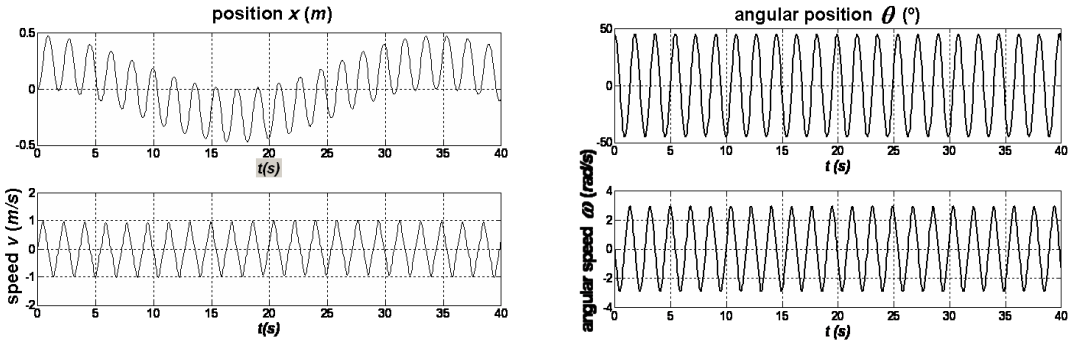


Figure 5. Time evolution of state variables using the nonlinear model.

In which concern the linear model, using the same parameters and initial conditions as before: $m_1 = 1Kg$, $m_2 = 2Kg$, $l = 1m$ and $k = 0.1Nm^{-1}$; initial angle of the pendulum is $\theta_0 = 45^o$ (the spring is initially unstressed, $x_0 = 0m$), the simulation of the linear *Simulink* model is displayed in Figure 6.

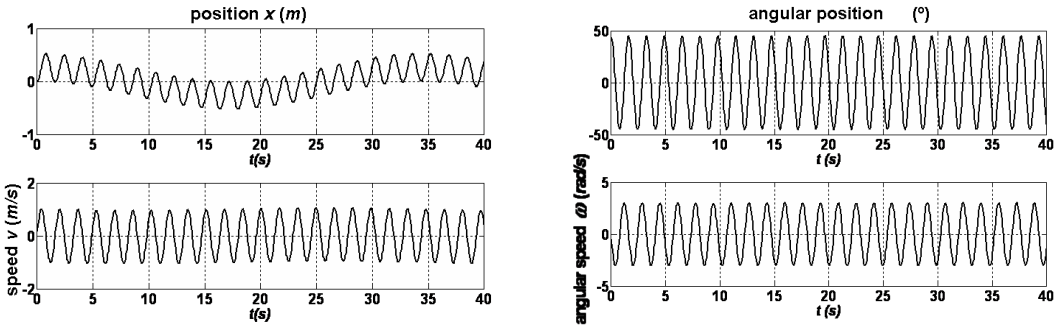


Figure 6. Time evolution of state variables using the linear model.

Figure 5 and Figure 6 are presented as examples of an arbitrary simulations of both models: nonlinear and linear.

The system developed in *Matlab/Simulink* is now prepared for a large set of experiments oriented to time domain or frequency domain.

One of the experiment that can be done consist on evaluation of the maximum radius of the equilibrium point neighbourhood allowed for the linear approximation can be considered acceptable. So, remaining constant the set of parameters, both systems was testes with $\theta_0 = 10^o$ and $\theta_0 = 45^o$ (unstressed spring in the beginning for both cases). The time evolutions of the state variables are superimposed for the nonlinear model and linear approximation.

In the Figure 7 we can see what's happened with the position and speed of mass m_2 when the initial angle of the pendulum θ_0 is 10º (left side) and 45º (right side). In the first case ($\theta_0 = 10^o$) the time evolution produced by both models (linear and nonlinear) are practically superimposed (left side of Figure 7), so, the linear approximation can be consider accurate.

On the other hand, when we use $\theta_0 = 45^o$ (right side of Figure 7) the curves are clearly not superimposed denoting that a considerable error is produced when we use de linear approximation.

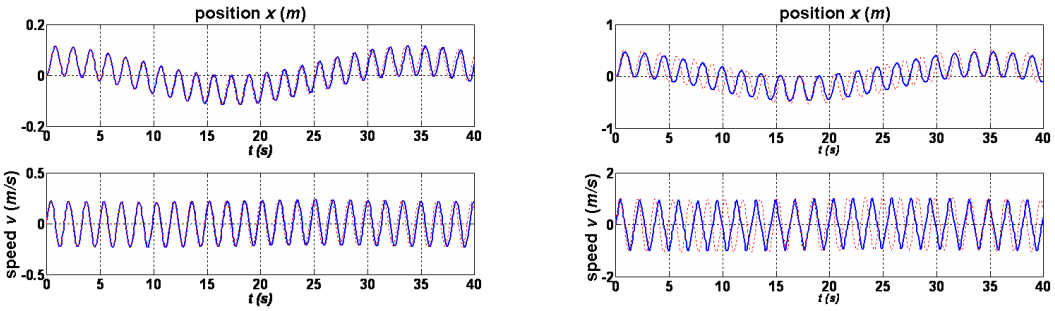


Figure 7. Position and speed: comparison between nonlinear system and linear approximation. Left $\theta_0 = 10^o$, right $\theta_0 = 45^o$

Same considerations can be pointed out for the angular position of the pendulum and its first derivative, as illustrated in Figure 8.

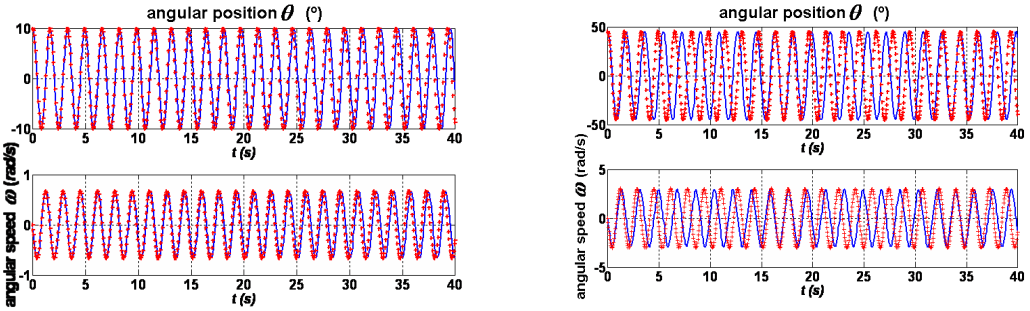


Figure 8. Angle and angular speed: comparison between nonlinear system and linear approximation. Left $\theta_0 = 10^o$, right $\theta_0 = 45^o$.

These preliminary experiments are examples of how useful is the *Matlab* environment to analyse systems. For the actual example, we can conclude that radius of neighbourhood of the equilibrium point should be between 10º and 45º for producing linear accurate models.

5. CONCLUSIONS

This paper presents one of among large number of examples that are currently studied in engineering curricula.

System theory and control systems are transversal areas of knowledge. Nowadays, with the improvement of processors capabilities it is unthinkable making any analyses or synthesis of a system without any aid of a computer or dedicated processor. Many software tools have been

developed in the last decades, among them Matlab/Simulink reveals as a powerful tool for a control engineer.
The simple example chosen for this paper demonstrate how useful is the *Matlab* software to solve the problem of linearization. In fact, the linear approximation is only useful if it is enough accurate; with the presented approach we can numerically evaluate what we can define as neighbourhood of an operating point. This concept has a great importance in many branches of sciences and technologies, particularly, in Electronic Engineering.

REFERENCES

[1] D'Azzo, J. J., Houpis, C. H., *Linear Control System Analysis and Design Conventional and Modern*, Mc GRAW-HILL INTERNATIONAL EDITIONS Electrical and Electronic Engineering Series, 1988.
[2] Math Works Inc., Matlab User's Guide.The Math Works Inc., South Natick, MA01760, USA, 2004.
[3] Dorf, R. C., Bishop, R. H. *Modern Control Systems*, Addison-Wesley Publishing Company, USA, 1995.
[4] Marion, J. B., Thornton, S. T. *Classical Dynamics of Particles and Systems*, Brooks/Cole Pub Co., 2003.
[5] Kadiyala, R., "A Toolbox for Approximate Linearization of Nonlinear Systems", IEEE Control Systems, pp. 47-56, 1993.

SYMCOMP 2017
Guimarães, 6-7 April 2017

AN APPROACH FOR DESIGNING VERTICAL AXIS WIND TURBINES USING NUMERICAL METHODS AND GENATIC ALGORITHIMS

Karim Hamouda[1], Amir Abdelmawla[1] and Tarek M. Hatem[1,2*]

1: Centre for Simulation Innovation and Advanced Manufacturing (SIAM)
The British University in Egypt
El Sherouk Campus, 11837 Cairo, Egypt
E-mail: karim.hamouda92@gmail.com,
amir.abdelmawla@bue.edu.eg
tarek.hatem@bue.edu.eg

2: Microstructure Physics and Alloy Design
Max-Planck-Institut fur Eisenforschung Düsseldorf, 40237, Germany
E-mail: t.hatem@mpie.de

Keywords: Vertical Axis Wind Turbine, Blade Element Momentum, Genetic Algorithm, Wind Energy, Computational Engineering

Abstract *In the last decade, the rise of environmental concerns inspired an increasing demand for renewable energy resources. This has increased the interest for the use of Vertical Axis Wind turbines, particularly due to their potential of power generation in remote areas and low capacity applications. However, the design of wind turbines is complicated due to a vast amount of design parameter combinations between blade length, airfoil profile and swept area that could be employed to satisfy an application. Moreover, the structural integrity of the turbine influences these parameters selection. Therefore, in this work, a new screening method was developed that could be able of initially testing a large number of VAWT parameters while making sure the structural integrity is satisfied in order to design a 5KW Darrieus type Vertical Axis Wind Turbine. The method in question combines the Blade Element Momentum model and a simplified Beam analysis model with a Genetic Algorithm Optimization scheme in order to minimize the number of possible turbine parameters with the least computational requirement. The parameters selected this stage of the design can be further studied with CFD and FEM models in order to select the final design parameters for prototyping.*

1. INTRODUCTION

Clean, renewable and sustainable energy is of paramount importance nowadays with the ever growing demand for energy, not only from growing cities, but also for remote areas

where installing large power plants is not feasible. Wind power is one of the most prominent renewable energy sources available. Generally, there are two main ways for wind energy generation that are commonly used nowadays; Horizontal Axis Wind Turbines (HAWTs) and Vertical Axis Wind Turbines (VAWTs) [1][2]. Although HAWTS are more expensive to build, they are more commonly used due to their ability to operate at higher rotational speeds and they produce more power than VAWTS.

However, VAWTs are cheaper, insensitive to wind, therefore, require no yawing mechanisms, and can be used in urban areas because they are do not produce noise. Moreover, the drive train for VAWTs can be placed at the bottom of the tower allowing easier installation and maintenance compared to HAWTs. Their power can be used through a direct drive via a vertical shaft, making it also possible for them to deliver mechanical power for pumps or other applications directly. Their blades require no tapering or twisting; thus it is easier to manufacture them through extrusion.

There are two main types of VAWTs, Darrieus and Savonius. Savonius wind turbines achieve rotation through drag, which makes them high torque machines with high reliability yet low efficiency. On the other hand, Darrieus turbines achieve rotation through lift; capable of achieving higher speeds and higher efficiencies. However, they have a low starting torque and therefore usually require some sort of starting mechanism. For the purposes of this study, only Darrieus turbines will be considered.

VAWTs have shown the potential capability of decreasing the cost of energy while providing high reliability. The modern advancements in technology and materials science are growing the possibility of introducing VAWTs as a major source for wind energy production. However, one of the main hurdles in VAWT design remains to be its highly demanding computational power and complicated aerodynamic analysis. Simplified models such as Blade Element Momentum (BEM) have been simplified and improved over the years to produce highly reliable results, however, their applications are limited to low solidity rotors.

On the other hand, VAWTs can have their structure analysis done through finite element analysis. FEA is capable of producing highly reliable results mainly due to the fact that centrifugal forces are mostly contributing to the rotor load system. In that sense, a coupling between BEM and structural analysis model can be used in order to assess whether a particular VAWT configuration can be further studied or not. By coupling the previous model with a Genetic Algorithm (GA) tool, the solution space can be scanned for particular configurations that satisfy a particular fitness function.

Genetic Algorithms have been successfully used to design HAWTS. In the work done by M.S.Selig [3] , the author notes that previously turbine designs relied heavily on direct approaches where it has been required to perform a large number of trade studies to then be able to optimize the parameters of the final blade design. Then the process required complicated analysis of the trade-offs that would be required, ie. the blade chord vs lift distribution, rotor radius and rpm and others. GA has been able to successfully iterate over

these tradeoffs much quicker and randomly than gradient-based algorithm which was heavily reliant on the user's input and experience. In his work, the author attempted to use GA to study the effects of changing the blade pitch, chord and twist on the annual energy production. The design of five turbines was done by tailoring the blade for particular wind speeds at the installation site.

There are many techniques that have been developed in order to seek out the required output, but Genetic algorithms such as the one presented by Castelli et al. shows great promise. Castelli et al. [5] used order to improve the performance of a three bladed wind turbine using a NACA 0021 airfoil. Casteli et al. used evolutionary GAs to find the optimum TSR in order to maximize the turbines power coefficient and torque.

Similarly, Donha et. al [9] used GAs for tuning the PID controller of a 300 kw HAWT when facing changing winds. The results produced by this method have shown that controllers tuned via GAs have similar and sometimes better responses than controllers tuned using more traditional tuning schemes such as the Ziegler-Nichols method.

The lay out of this report is as follows. The second chapter will discuss the BEM model implemented. The third chapter will discuss

2. AERODYNAMIC MODEL OF VAWT

In order to model the performance of a vertical-axis wind turbine the Blade Element Momentum (BEM) is the most common.

The model chosen for the aerodynamic analysis is the double-multiple stream tube with variable interference factor, often abbreviated DMST (see Figure 1), which is enclosed in the momentum models category and is based in the conservation of momentum principle based on the actuator disc theory, which can be derived from the Newton's second law of motion. It has been used successfully to predict overall torque and thrust loads on Darrius rotors [10] [11].

The main advantage of momentum models is that their computer time needed is said to be much less than for any other approach, the main dis advantages of BEM it doesn't take a dynamic stall effects into considerations. In addition, it doesn't give us pressure distribution to be able to do a blade stress analysis [14] [16]

2.1. Double Multiple Stream Tube model (DMST) for VAWT

The double multiple stream tube model is an extension of the multiple stream tube model. The model presents the same momentum balances as the single and multiple stream tube models.

The double multiple stream tube model divides the turbine into upstream and downstream [10]. The incoming wind in the downstream is assumed to be the upstream wake velocity. This provides a better account for the wake effects (see Figure 1).

The swept volume of the rotor is divided into adjacent, aerodynamically independent stream tubes, each one is identified by his middle $\Delta\varphi$ angle as shown (see Figure 2).

The analysis of the flow conditions is made on each stream tube using a combination of the momentum and blade element theories, the former uses the conservation of the angular and linear momentum principle [16].

It is assumed that the wind velocity is decelerated near the rotor by induction factors a and a' for both up and down stream regions respectively corresponding to equations (1) and (3), for each stream tube has its own induction factors if we represent the front and rear part of the turbine by two disks in series, the velocity will be decelerated two times, one for the upstream and the other for the downstream.

$$V_u = (1-a)V_\infty \tag{1}$$

$$V_e = (1-2a)V_\infty \tag{2}$$

$$V_d = (1-2a)(1-a')V_\infty \tag{3}$$

where V_∞ is the free stream velocity, V_u is the velocity in upwind region, V_e is the velocity in equilibrium region, and V_d is the velocity in downwind region while a and a' are the interference factors for up and down stream zones respectively as presented (see Figure 1).

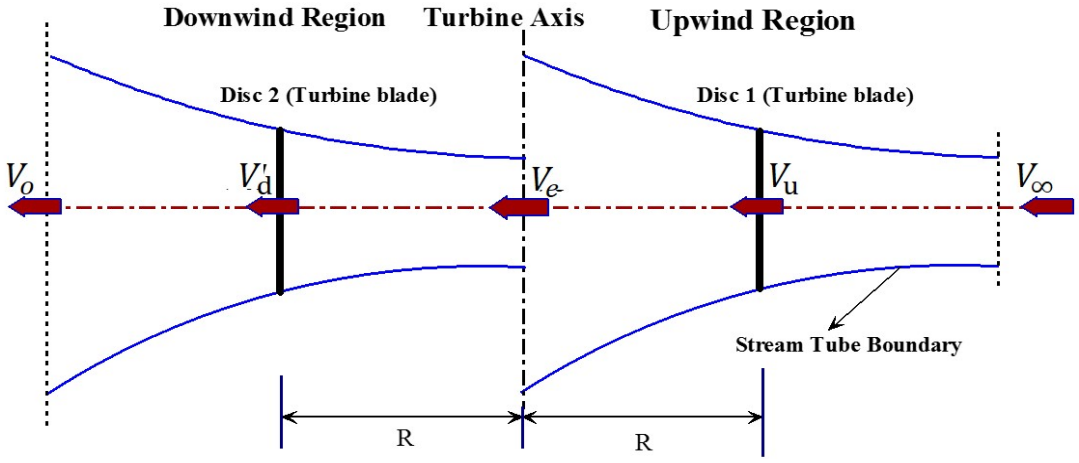


Figure 1. Double actuator disc model

Due to the wind velocity deceleration two times during the passage through the VAWT rotor, the max power can be taken by the rotor from wind does not exceed 59.26 %, this ratio for single actuator disc or Horizontal Axis Wind Turbine (HAWT) and This is what is known Betz limit. It is a theoretical limit in the efficiency of a wind turbine determined by the deceleration the wind suffers when going across the turbine. For double actuator disc model or VAWT the maximum power coefficient for an ideal rotor is 0.64 [14] [18].

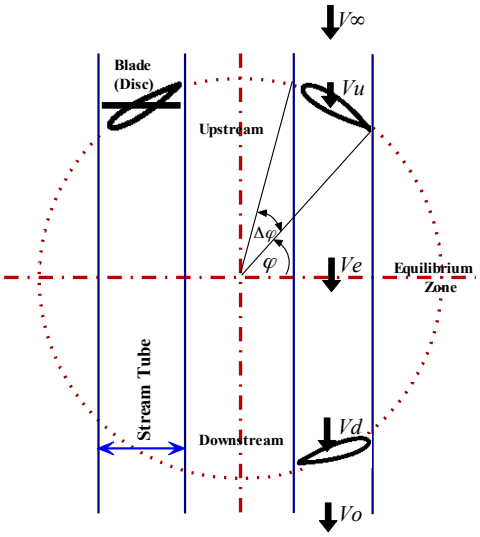


Figure 2. Double multiple stream tubes model for VAWT

2.2. Aerodynamic Model Methodology

First of all, the model calculates induced wind velocities (absolute and relative), the angle of attack (AOA) can be obtained corresponding the blade velocity triangle according to (see Figure 3).

According to the turbine blade cross section that it was selected for the computation.

Lift and drag coefficients are obtained from wind experimental data based on AOA and Reynolds number.

The aerodynamic coefficients are used to calculate the axial thrust force from equation [10].

By using an iterative process, the model simultaneously solves two equations for the stream-wise force at the actuator disk, one obtained by conservation of momentum and other based on the aerodynamic coefficients.

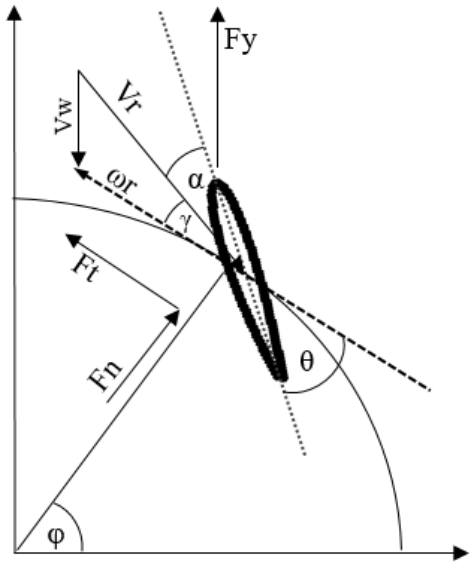


Figure 3. VAWT blade diagram

$$\alpha = \theta - \gamma \tag{4}$$

$$C_T = -C_L \sin\gamma - C_D \cos\gamma \tag{5}$$

$$C_N = C_L \cos\gamma - C_D \sin\gamma \tag{6}$$

$$C_Y = C_T \cos\varphi + C_N \sin\varphi \tag{7}$$

$$C_{Ym} = 4a\left(1-a\right) \tag{8}$$

where: C_L, C_D are the lift and drag forces coefficients respectively and C_T, C_N are the tangential and normal forces coefficients while C_Y is the thrust force coefficient.

These equations are solved twice, for the upwind with an interference factor a and for the downwind with an interference factor a' [10]. Once the accurate aerodynamic forces are obtained, turbine torque and the associated power are computed form equations (5), (9), and (10).

$$T = \frac{1}{2}\rho V_r^2 c\, r\, h\, C_T \tag{9}$$

$$P = T\,\omega \tag{10}$$

where: c is the chord length, r is the rotor radius and h is blade length while V_r is the relative wind speed and ρ is the air density.

The model was built on MATLAB based on the above mentioned aerodynamic theory. The model follows the flow chart shown (see Figure 4).

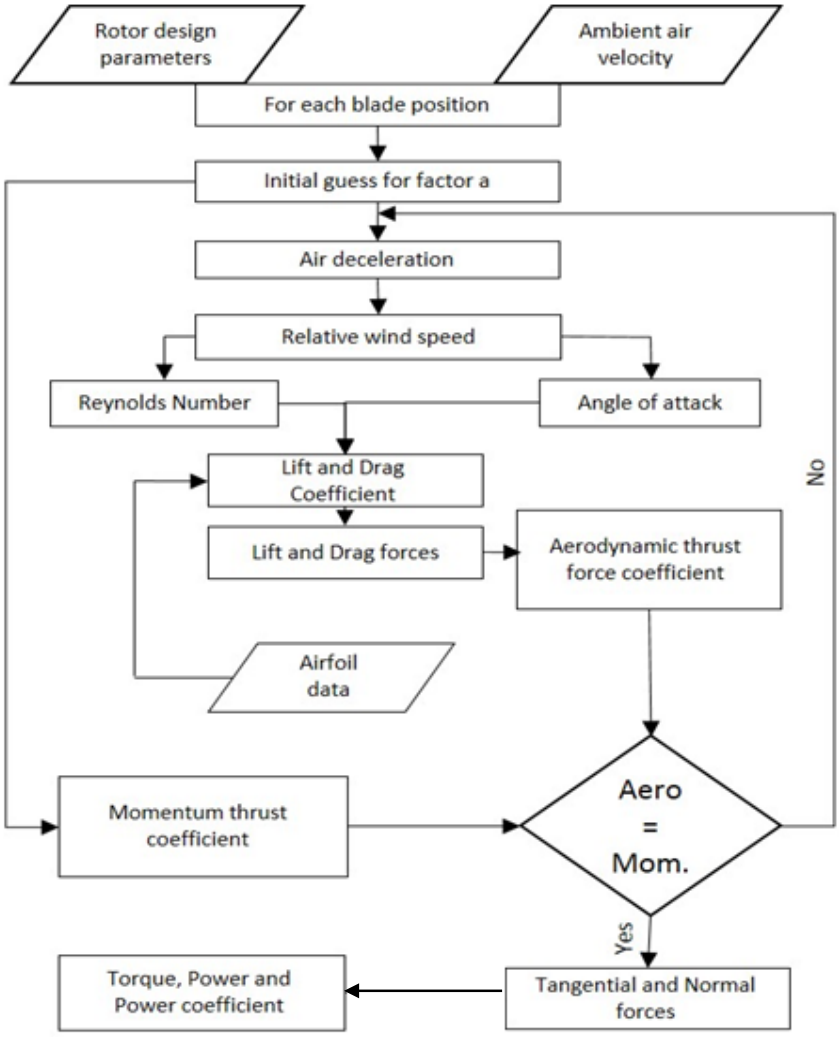


Figure 4. Aerodynamic model computation flow chart

Figure 4 illustrates a computation flowchart describing the calculation of aerodynamic loads, torque and power coefficient. For a given rotor geometry, tip speed ratio and free stream velocity, a first set of calculations from is made using initial guess $a = 0$, then with the new value of a the procedure is repeated until the initial and final a value are similar. This is repeated for each stream tube position and once the upstream induced velocities have been calculated, the downstream half is calculated using the same set of formulas, interchanging the upstream induced velocity V_u by the downstream induced velocity V_d (see equations (1), (3)).

3. OPTIMIZATION OF VAWT PARAMETERS USING GENETIC ALGORITHIM

3.1. Introduction for Genetic Algorithm

Genetic Algorithms (GA) are used in order to search for parameters using the Darwinian concept of natural selection. It does so by operating on a set of solutions instead of just one solution as is used by other calculus based optimization techniques. A solution in the set chosen by GA is referred to as an individual, and the set is referred to as the generation or population. In that essence, GA requires an objective function; also known as a fitness function, that can be used to assess the fitness of each individual in the generation. The fittest individuals are then used to help define the second generation based on their characteristics. Therefore, it uses the concept of "survival of the fittest" to be able to iterate until it reaches the desired solution.

Of course, GA is not a perfect optimization technique, but its weaknesses usually depend on how you define and constraint the problem. If the problem isn't defined correctly GA will just choose the borderline parameters with disregard of what is needed. However, GA does not really have any other main weaknesses. MATLAB provides an excellent easy to use GUI that can be utilized to deal with our problem [19]. This tool box allows us to include M script which includes our fitness function, as well as, constraint our inputs to suit our needs (see Figure 8).

3.2. VAWT Blade Airfoil Selection

Airfoil selection remains to be one of the most complicated steps required of the optimization process. Generally, referring to previous work and currently designed VAWTS, it has been preferred to use symmetrical airfoils most commonly NACA 0018 [5] [6] [20]. As mentioned there has been several reasons for that:

1. Symmetrical airfoils have a constant performance throughout the azimuth rotation.
2. NACA 0018 provides the optimum ratio between structural failure concern and increasing drag with increasing thickness.
3. Symmetrical Airfoils don't produce negative torque in the downstream.

The above reasons don't necessarily mean that symmetrical airfoils are the best. Cambered airfoils produce higher lift in the upstream region of rotation, however, they may produce negative torque in the downstream region. Also, cambered airfoils have zero pitch lift which is advantageous. However, since variable pitch will be included in this work the study of cambered airfoils have been dismissed.

There are several other commonly used airfoils which are highlighted in a lot of literature which seem appropriate for use in VAWTs [5] [6] [20]

- Symmetrical, NACA 4 digits, 0012, 0015, 0018 and 0021
- Cambered, NACA 4 digits, 2418, 4418 and 6418
- DU-06-W-200
- NACA 6 digits, 633018, 633418, 633618 and 633618

Figure 5 shows the profiles of the selected airfoils plotted over one another for comparison.
The variety of airfoils available all seem promising and all have their advantages and disadvantages. Therefore, a criteria must be created to choose between the airfoil profiles had to be used to choose between them.

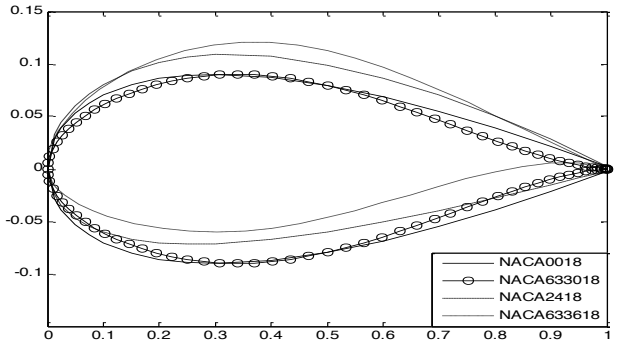


Figure 5. Airfoil profiles plotted over one another

Using XFOIL, coefficient of lift and coefficient of drag were produced for each airfoil [21]. Then the first possible criteria to minimize the airfoil choices was to see the change in the aerodynamic ratio C_L/C_D curve across 180 degrees angle of attack [22]. Lift based straight blade vertical axis wind turbines usually demand maximum lift for minimum drag at each angle of attack to produce more power.
The data of the aerodynamic ratio was then calculated from that and they were plotted for all airfoils as shown (see Figure 6). The graph showed that the airfoil profiles with higher Cl/Cd were cambered airfoils with thicker sections and the one with the max airfoil was the 6 digit 633618.

Airfoils selection using genetic algorithms remains computational demanding.
The process of coupling both the turbine design parameters and the airfoil selection to the genetic algorithm will be extremely computational and may lead to misleading results.
The NACA 4 digit family has several prominent candidates that can be used in this work. Due to the reason mentioned above NACA 0012, 0015, 0018, and 6418 will be studied for how eligible they are to be used and the most optimum will be selected.

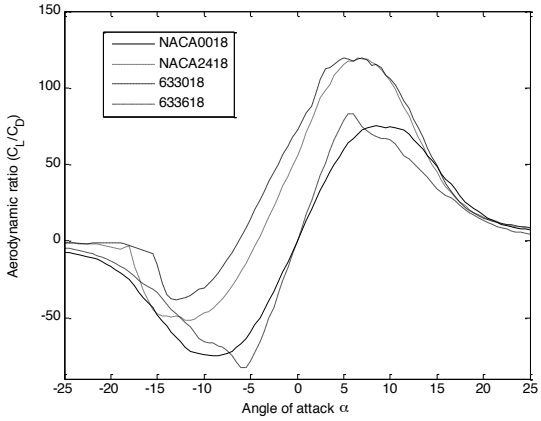


Figure 6. The aerodynamic ratio curves for the selected airfoils plotted over each other

The lift and drag coefficient curves were produced using XFOIL Due to computation concerns, a turbine will be designed for each airfoil and then the best airfoil will be selected by comparison between based on the cheapest design. The design process is discussed below.

3.3. Selection of Turbine Parameters Based on Genetic Algorithm

The turbine parameters to be selected here are the tip speed ratio, the solidity, the blade length and the diameter.
For a 5 kW VAWT the aerodynamic model predicts an approximate area ranging from 15-25 m^2, where the swept area = blade length × turbine diameter. The genetic algorithm basically works by optimization through minimizing a particular fitness/cost function.
For VAWTs, in general, the cost function should be the price required to produce the power. In general, that doesn't mean that the initial cost should be cheap, however, it means that the wind turbine should survive for as long as possible with minimum maintenance even if the initial cost is high. The optimization process will therefore be performed based on several fitness functions.

3.3.1. Fitness Functions

As mentioned above the main purpose is to design a 5kW VAWT and the purpose of the optimization is to determine the optimum rotor parameters that can achieve that task. The main rotor parameters as mentioned above are the Tip Speed Ratio (TSR, λ), solidity, turbine diameter, blade length and the rated wind velocity.
Since the TSR and the solidity of the VAWT are inversely proportional based on the graph shown (see Figure 7) the determination of one parameter is enough for determining the other from the graph. Based on that 3 fitness functions are going to be tested. The first is to minimize the rated wind speed and area which are required to produce 5 kW.

<table>
<tr><th>Optimization</th><th>Fitness functions</th><th colspan="2">Inputs</th></tr>
<tr><td rowspan="3">First Optimization</td><td>Minimum Moment of Inertia</td><td>Tip speed ratio (λ)</td><td>Turbine Diameter</td></tr>
<tr><td>Minimum Mass</td><td rowspan="5">2 – 4</td><td rowspan="5">3 – 18</td></tr>
<tr><td>Minimum Blade Length</td></tr>
<tr><td rowspan="4">Second Optimization</td><td>Moment of Inertia</td></tr>
<tr><td>Maximum Average Torque</td></tr>
<tr><td>Minimum Area; A_{FB}</td></tr>
<tr><td>Minimum Blade Length</td></tr>
<tr><td rowspan="4">Third Optimization</td><td>Minimum Torque STD</td><td>Turbine Diameter</td><td>Number of Blades</td></tr>
<tr><td>Maximum Average Torque</td><td>3 – 7</td><td>3 – 5</td></tr>
<tr><td>Minimum Blade Length</td><td rowspan="3"></td><td rowspan="3"></td></tr>
<tr><td>Minimum Area</td></tr>
<tr><td rowspan="3">Fourth Optimization</td><td>Minimum Torque STD</td></tr>
<tr><td>Maximum Average Torque</td></tr>
<tr><td>Minimum Blade Length</td><td colspan="2">Chord Length</td></tr>
<tr><td></td><td>Minimum Area</td><td colspan="2">0.4 – 0.6</td></tr>
</table>

Table 1. Genetic Algorithm input parameters

The second will be to minimize the area at the rated 8 m/s wind speeds. Finally the third uses a simplified stress model to measure the dimensions of the wind turbine and make them as cheap as possible.

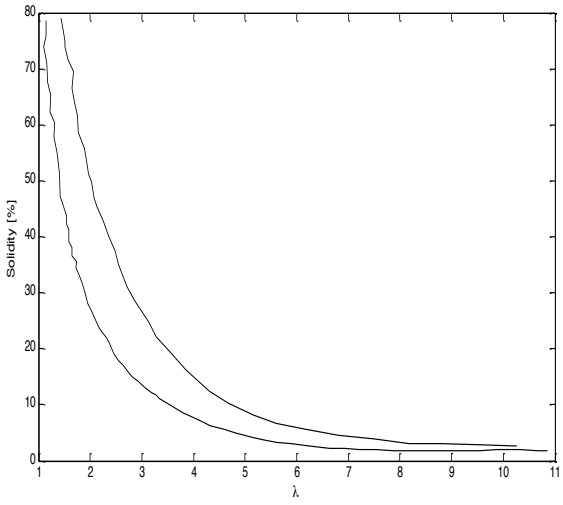


Figure 7. Graph representing the optimum solidity for each TSR

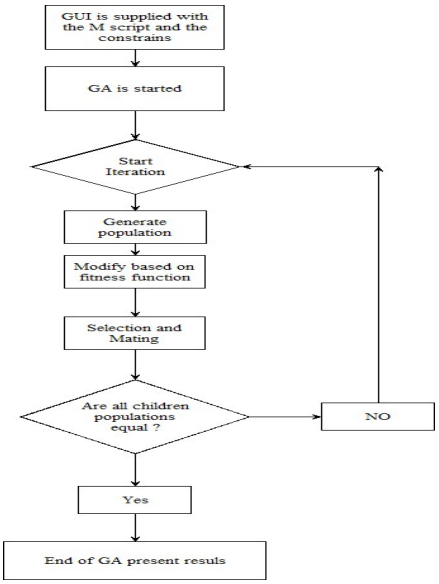


Figure 8. Genetic Algorithm flowchart

3.3.2. Stress Model

Figure 9 shows the main structural components of the VAWT. The Stress model will consist of studying the three main components (the blade, the supporting strut and the central column) separately.

In order to determine the VAWT parameters a simple stress model will be generated.

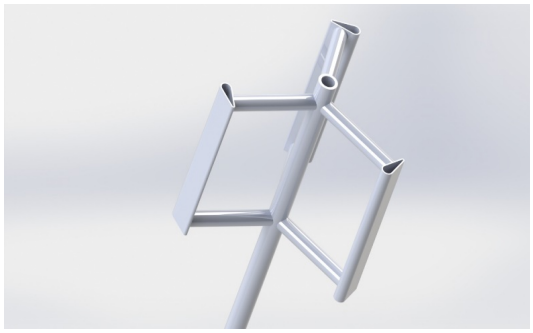


Figure 9. Basic VAWT Structure

An exact model cannot be used since the BEM model does not produce all of the actual stresses on the turbine. These include, the exact pressure distribution on the blade profile, the blade tip vortices and turbulence caused by the supports [23]. Hence the model used will have the following assumptions:

- The blades are only subjected to the aerodynamic lift (which is assumed to be a uniform force) and buckling under their own weight.
- The support struts are subjected to the weight of the blades and the aerodynamic

lift.

- The central columns are subjecting to the buckling load from the weights.

The stress model built will use the beam theory analysis for the formulation of the stress model. From the work of Ashuri [24] the safety factor can be assumed to be 1.5 which is commonly used for commercial VAWTS.

3.3.3. The Turbine Blades

The blades can essential be modelled as simple beams that have been hinged at both ends. There are two main forces acting on the blade. The shear force acting on the stagnation point on the leading edge and the lift force acting on the turbine blade. There are other forces acting on the blade such as down wash, shear from drag, vortices and effects aerodynamic effects of the blades on each other. However, due to the computational power required and the lack of data provided by the BEM model only the first two forces are going to be considered.

3.3.4. Area and Moment of Inertia

The area and the moment of inertia of the airfoil cross-section are directly proportional to the thickness and the maximum camber of the airfoil. The following couple of equations introduce area and bending inertia of airfoil sections [25].

$$A = K_A C t \tag{11}$$

$$I = K_I C t\left(t^2 + H^2\right) \tag{12}$$

Where: C is the maximum camber of the airfoil, t is the airfoil maximum thickness while K_A , K_I are fitting parameters.

4. RESULTS AND CONCLUSION

The optimization technique was done using multiple objective genetic algorithm. The multiple objective genetic algorithm is a genetic algorithm technique that uses multiple fitness functions in order to find the optimized parameters. This technique uses several fitness functions that may require different optimum values and therefore the technique balances between the fitness functions to produce the optimum parameters. Two sets of fitness functions were used. The first technique used the moment of inertia, the total mass and the length of the turbine blades.

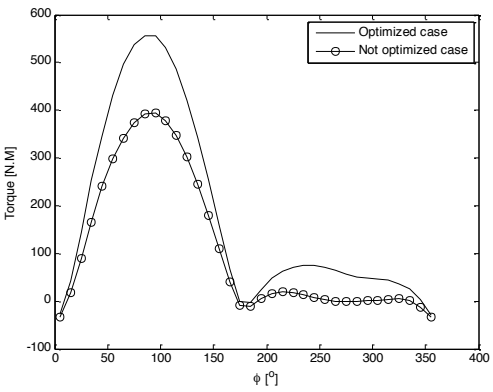


Figure 10. Torque through azimuth angle for optimized and non-optimized case

Figure 10 shows the difference between two cases, one is optimized while other case not optimized from blade torque point of view. The optimized case is the 8th case in the first optimization method as shown in (see Table 2), the data are 3.19 in TSR, 5.09m in diameter, 5.74m in blade length and turbine solidity is 0.29. On the other hand the non-optimized case has 0.27 in solidity, 4 in TSR, 25m^2 in swept area and 5m in diameter, both cases are done at 8m/s in wind speed, the blade torque for optimized case larger than non-optimized case and this is the anticipated result.

Trail	Moment of Inertia	Bending Moment	Blade Length	Tip Speed Ratio (λ)	Turbine Diameter	Swept Area
1	1255457	301641	10.06	3.32	3.04	30.58
2	1255243	301672	10.06	3.31	3.04	30.58
3	1255457	301641	10.06	3.32	3.04	30.58
4	6716081	5546699	2.8	2.71	9.35	26.16
5	3901140	3289094	3.43	2.76	7.85	26.93
6	7778516	6298789	2.63	2.66	9.76	25.7
7	8496348	6775841	2.54	2.58	10	25.43
8	1860528	917634	5.74	3.16	5.09	29.26
9	8515477	6764019	2.57	2.65	9.99	25.65
10	2839847	2051029	4.27	3.12	6.7	28.63

Table 2. First optimization attempt results

The second technique used 4 fitness functions the moment of inertia, the footprint, maximum torque and the length of the turbine blades.
Both techniques optimized two parameters the tip speed ratio which was bound between 2 – 4 and the turbine diameter which was bound between 3 – 18 meter.
The multiple objective produces a pare to front or a list of results each favouring a fitness function. The results are shown (see Table 2 and Table 3) below.

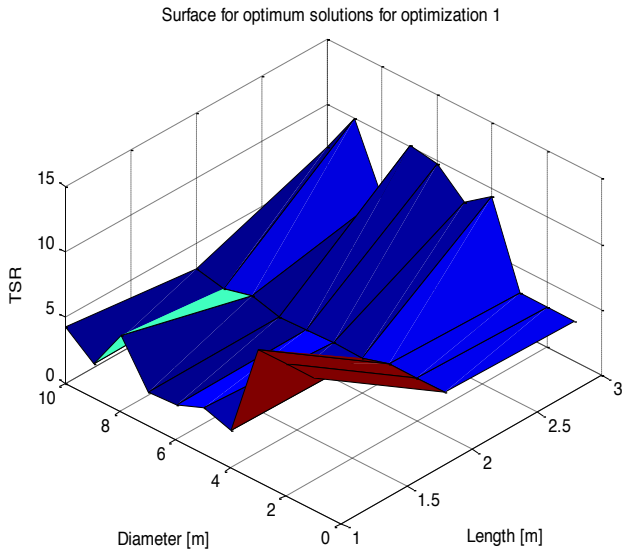


Figure 11. Surface plot for the first optimization approach

Trail	Moment of Inertia	Torque	Foot Print Area	Blade Length	Tip Speed Ratio (λ)	Turbine Diameter
1	324145	30	7	10.5	3	3
2	48599	689	113	2.2	2.5	12
3	879959	30	7	12.7	2.5	3.1
4	46675	680	113	2.2	2.6	12
5	878955	30	7	12.7	2.5	3.1
6	46623	676	113	2.2	2.6	12
7	54783	403	68	2.8	2.6	9.3
8	47443	598	101	2.3	2.6	11.4
9	58766	356	61	2.9	2.6	8.8
10	261693	28	7	10.2	3.3	3
11	157190	117	24	5.4	2.8	5.5
12	689062	34	8	11.5	2.6	3.2
13	47863	560	94	2.3	2.6	10.9
14	580385	49	11	9.7	2.5	3.8
15	150541	157	30	4.8	2.6	6.2
16	343000	73	16	7.5	2.6	4.5
17	110054	214	39	4	2.5	7
18	69108	282	50	3.3	2.7	8
19	46664	680	113	2.2	2.6	12

Table 3. Second optimization attempt results

The produced optimization can then be used to produce a surface plot using the three main design parameters (TSR, diameter and length) to include the optimum dimensions for a 5kw

low speed vertical axis wind turbine. The plots were drawn using MATLAB surface plots as shown in (see Figure 11 and Figure 12).

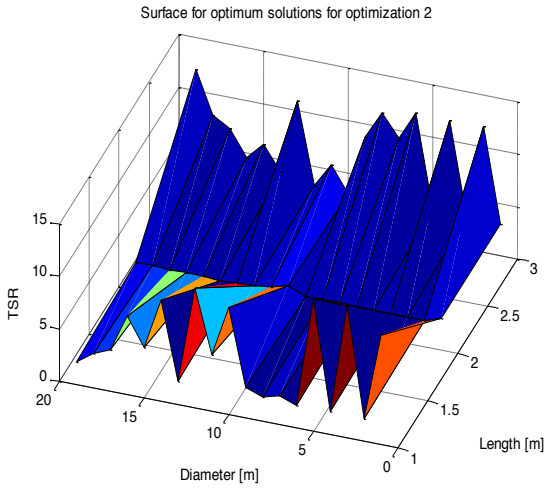


Figure 12. Surface plot for the second optimization method

4.1. Optimization Based on VAWT Number of Blades

The optimization process can further be expanded by attempting to further optimize the dimensions for the number of blades. It is known for a fact that adding the number of blades on the turbine act to stabilize the torque fluctuations due to the added mass as in the flywheel.

This is pretty advantageous and would make it a lot easier for the generator design and selection. In order to optimize for the number of blades where the main goal is torque stabilization the problem seems that the fitness function chosen would be the standard deviation for the torque. However, this would just favorably increase the number of blades regardless of the rest of the parameters.

Therefore, using multi-objective genetic algorithm the following fitness functions were selected: minimum torque standard deviation, maximum torque, minimum blade length and minimum area required to generate the 5kw.

The torque standard deviation was calculated by creating a vector that stores the torque value at every azimuth which was being calculated by the BEM model.

The model, however, was edited to include the extra blades by adding the torque inputs at different angles. The maximum torque fitness function simply ensure the maximum power coefficient and the choice of both the blade length and area is simply to ensure that the algorithm doesn't favour minimizing the length and maximizing the diameter.

The inputs, however, were changed for this problem. The main inputs were the diameter and the no. of blades. The diameter was varied between 3 – 7 and the number of blades from 3 – 5. The TSR and solidity where controlled so a chord input which was retrieved from the previous optimizations. It was noted that the optimum chord which can withstand

the stresses ranges from 0.4 – 0.6 meters. Therefore, this optimization was run twice. The first time was with a constant chord at 0.4 meter and the second run varied the chord from 0.4 – 0.6. The results are shown (see Table 3 and Table 4) below.

Trail	Torque (standard deviation)	Torque	Blade Length	Swept Area	Turbine Diameter	Number of Blades
1	13.4	55	8.3	12.6	4	3
2	18.8	166	4.2	38.4	7	4
3	10.7	177	4.3	38.4	7	5
4	3.3	75	8.7	19.3	5	5
5	13.2	46	13.4	13.4	4.1	5
6	31.4	89	6.7	21	5.2	4
7	9.1	53	11.8	14.4	4.3	5
8	10.7	177	4.3	38.4	7	5
9	6.5	64	10.1	16.9	4.7	5
10	26	60	9.1	14.6	4.3	4
11	21.6	122	5.9	29.2	6.1	5
12	17.4	150	4.4	33.3	6.5	4
13	29.4	76	7.4	17.9	4.8	4

Table 4. Third optimization attempt results

Trail	Torque (standard deviation)	Torque	Blade Length	Swept Area	Turbine Diameter	Number of Blades	Chord Length
1	13.4	54.5	8.3	12.6	4	3	0.4
2	4	88.8	8	22.2	5.3	5	0.43
3	21.3	187.2	4	38.5	7	4	0.49
4	13.2	46.3	14.4	12.6	4	4	0.57
5	11.6	44.1	15.5	12.6	4	4	0.58
6	24.1	189.5	4	38.5	7	4	0.51
7	11.5	44.1	15.5	12.6	4	4	0.58
8	25.3	148.6	4.7	31.2	6.3	4	0.46
9	56.2	154.6	5.3	33.2	6.5	4	0.6
10	22	188.7	4	38.5	7	4	0.5
11	11.4	112.6	5.1	25.3	5.7	4	0.45
12	21.7	61.4	11	14.7	4.3	4	0.56
13	37.6	90.7	7.2	21	5.2	4	0.48
14	33.6	79	8.1	18.5	4.9	4	0.49
15	51.1	166.8	4.9	35.2	6.7	4	0.58

Table 5. Fourth optimization attempt results

As with the previous stage of optimization a surface plot for the optimum solutions can be generated and using matlab as presented (see

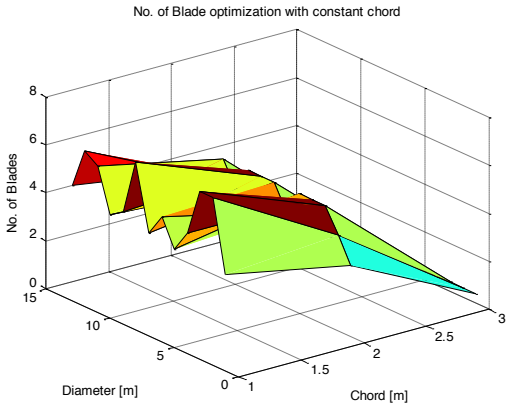


Figure 13 and

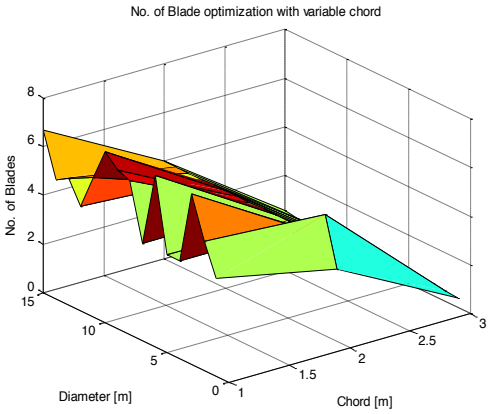


Figure 14) below.

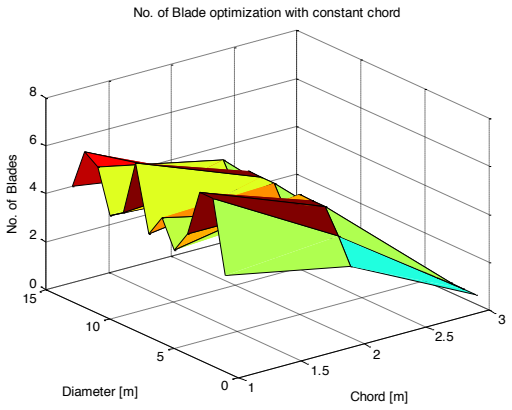


Figure 13. Number of blades optimization with constant chord

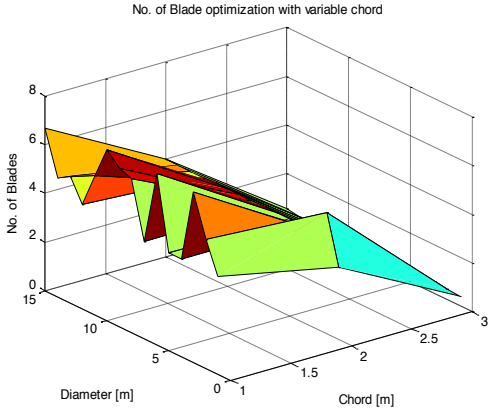


Figure 14. Number of blades optimization with variable chord

ACKNOWLEDGMENT

This research is supported by EU-Egypt Innovation fund through the RDI scheme 1 program, project No. ENP/2014/343 – 629.

REFERENCES

[1] R. Nobile, M. Vahdati, J. Barlow, A. Mewburn-Crook "Dynamic stall for a Vertical Axis Wind Turbine in a Two-Dimensional Study" *World Renewable Energy Congress,* Sweden, May 2011

[2] In Seong Hwang, Seung Yong Min, In Oh Jeong, Yun Han Lee and Seung Jo Kim, "Efficiency Improvement of a New Vertical Axis Wind Turbine by Individual Active Control of Blade Motion", *School of Mechanical & Aerospace Engineering*, Seoul National University San 56-1, Sillim-dong, Gwanak-gu, Seoul, 151-742, Korea.

[3] M. S. Selig, V. L. Coverstone-Carrol, "Application of a Genetic Algorithm to Wind Turbine Design", Department of Aeronautical and Astronautical Engineering, University of Illinois at Urbana-Champaign, Urbana, iL 61801-2935

[4] M. M. A. Bhutta, N. Hayat, A. U. Farooq, Z. Ali, S. R. Jamil & Z. Hussain, " Vertical axis wind turbine – Areview of various configurations and design techniqes", *Renewable and Sustainable Energy Reviews*, 16, 4 (2012) 1926 – 1939.

[5] I. S. Hwang, I. O. Jeong, Y. H. Lee and S. J. Kim "Aerodynamic Analysis and Rotor Control of a new Vertical Axis Wind Turbine by Individual Blade Control", *17th International Conference on Adaptive Structures and Technologies*, Oct. 16 – 19, 2006.

[6] M. R Castelli, A. Englaro, & E. Benini, " The Darrieus wind turbine: Proposal for a new performance prediction model based on CFD", *Energy,* 36 (2011) 4919 – 4939.

[7] M. Claessens, "The Design and Testing of Airfoils for Application in Small Vertical Axis Wind Turbines", *Delft University of Technology*, 2006

[8] B. K. Kirke, "Evaluation of Self Starting Wind Turbines for Stand-Alone Applications", *School of Engineering Griffith University Gold Coast Campus*, (1998).

[9] R. Howell, N. Qin, J. Edwards, & N. Durrani, " Wind tunnel and numerical study of a small vertical axis wind turbin", Renewable Energy, 35, 2 (2010) 412 – 422.

[10] D. C. Donha, & G. Risso, "Wind Turbine Controller Tuning by Using Genetic Algorithm, *ABCM Symposium Series in Mechatronics*, 1 (2004) 346-354.

[11] A. M. Biadgo, A. Simonovic, D. Komarov, S. Stupar, "Numerical and Analytical Investigation of Vertical Axis Wind Turbine" , Addis Ababa University, *FME Transactions* (2013) 41, 49 – 58

[12] H. Beri, Y. Yao "Double Multiple Stream Tube Model and Numerical Analysis of Vertical Axis Wind Turbine", *Energy and Power Engineering*, 3 (2011) 262 – 270.

[13] D. Marten, J. Wendler, "QBlade Guidelines", v0.6, January 18, 2013

[14] http://web.mit.edu/drela/Public/web/xfoil/

[15] M. Islam, David S. K. Ting, A. Fartaj, "Aerodynamic models for Darrieus-type straight-bladed vertical axis wind turbines", *Renewable and Sustainable Energy Reviews*, 12 (2008) 1087 – 1109.
[16] M. R. Castelli, S. D. Betta and E. Benini, "Effect of Blade Number on a Straight-Bladed Vertical-Axis Darreius Wind Turbine", *World Academy of Science, Engineering and Technology,* 61, 2012.
[17] R. DHRUV. "Performance Prediction and Dynamic Model Analysis of Vertical Axis Wind Turbine Blades with Aerodynamically Varied Blade Pitch", *North Carolina State University*, North Carolina, 2012.
[18] B. Stich, D. Bode, K. Freudenreich, "Simplified Load Assumptions for Wind Turbines with Vertical Axis", *Germanischer Lloyd Industrial Services GmbH, Brooktorkai 18*, 20457 Hamburg, Germany.
[19] S. M. Camporealea, V. Magib, "Streamtube model for analysis of vertical axis variable pitch turbine for marine currents energy conversion", *Energy Conversion & Management* 41 (2000) 1811 – 1827.
[20] Genetic Algortihm.(n.d.), Retrieved from Mathworks: http://www.mathworks.com/discovery/genetic-algorithm.html
[21] G. Bedon, M. R. Castelli, & E. Benini, "Optimization of a Darrieus vertical-axis wind turbine using blade element – momentum theory and avolutionary algorithm", *Renewable Energy*, 59 (2013) 184 – 192.
[22] M. DRELA, H. YOUNGREN, XFOIL 6.94 User Guide, *MIT Aero & Astro*, 2001.
[23] B. MONTGOMERIE, "Methods for Root Effects, Tip Effects and Extending the Angle of Attack Range to ±180 deg, with Application to Aerodynamics for Blades on Wind Turbines and Propellers", *FOI Swedish Defence Research Agency, Scientific Report*, FOI-R-1035-SE, 2004.
[24] B. Eng, "Structural Optimization of Multi-Megawatt Offshore Vertical Axis Wind Turbine Rotors", *Delft University of Technology*, 2013
[25] T. Ashuri, "Integrated Aeroservoelastic Design and Optimization of Large Offshore Wind Turbines", *Delft University of Technology*, 2012.
[26] Area and bending inertia of airfoil sections. (n.d.). Retrieved from MIT open source: http://ocw.mit.edu/courses/aeronautics-and-astronautics/16-01-unified-engineering-i-ii-iii-iv-fall-2005-spring-2006/systems-labs-06/spl10b.pdf

SYMCOMP 2017
Guimarães, 6-7 April 2017

SYMBOLIC COMPUTATION OF IDEAL COLUMN CRITICAL LOADS FOR THE LIMIT VALUES OF ELASTIC END RESTRAINTS

Miguel M. Neves[1,*] and Hugo Policarpo[1,2]

1: LAETA/IDMEC, Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mail: maneves@dem.ist.utl.pt

2: IPFN, Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mail: hugo.policarpo@tecnico.ulisboa.pt

Keywords: symbolic computation, buckling, critical loads, ideal column, elastic end restraints

Abstract *This publication presents a technique developed to be used in structural mechanic courses (and also in solid mechanics). The main purpose is to present a technique used to obtain critical loads of columns using symbolic computation. The knowledge of this stability requirement is important for structural slender members subjected to compressive loads because they may either fail due to yielding or to lateral deflection, i.e., buckling. The theory considers the Euler-Bernoulli beam with a compressive axial load, which differential equation has an analytical solution. The critical loads as well as the constants of the displacement solution are determinated by applying the elastic boundary conditions using longitudinal as well as torsional springs. For boundary conditions that are homogeneous, the system of equations is indeterminate and its characteristic polynomial allows to obtain the critical loads and its corresponding constants of the displacement (eigenmode) solution. The symbolic computation is used for 2D cases assuming that the beam section has two principal axis of symmetry and at each end two springs (one longitudinal and other torsional). The general expression obtained is then simplified and analyzed for the limit values of the elastic supports which allows to model also the cases of beams with rigid supports. The symbolic computation technique is presented and discussed. Several known cases are presented to illustrate the symbolic computation technique adopted. These examples show the potential of symbolic computation tools to deal with complex formulation where hand calculation would be very demanding and unnecessary nowadays.*

1. INTRODUCTION

The buckling analysis of ideal columns (or beams under compressive axial load) is a well-established procedure. Indeed, the differential equation can be obtained from an infinitesimal element using the equilibrium and constitutive equation of Euler beam theory.
To obtain the differential equation, it is assumed that the beam is geometrically perfect, built from an isotropic and homogeneous material, and subject to a compressive axial load perfectly aligned with the axis of the beam.
In a bifurcation analysis one is looking for which loads the fundamental equilibrium path (followed in the static equilibrium) presents a bifurcating path, i.e. the possibility of another equilibrium position for the same axial load.
In the case of ideal columns the bifurcating path is only imminent when the axial load achieves its critical value, known as Euler critical load Pcr. It means that the problem changes from determinate to indeterminate when it reaches this critical load, and its value is obtained via the corresponding characteristic determinant of the system of equations, i.e. typically by solving an eigenvalue problem. The equation system obtained is homogeneous.
Note that, if the axial load has an eccentricity then the boundary conditions have to include the produced moment and the problem becomes nonlinear as soon as the loading process begin. With eccentric axial load the beam has no Euler critical load Pcr, because the equation system obtained is non-homogeneous and that is out of scope of this text.
The solution of the homogeneous system of equations of the ideal columns can be done for the more generic case where the beam has elastically restrained ends. Also, the solution with elastically restrained ends must be able to give the limit solutions where the elastic springs have infinite or zero rigidity. These calculations can be done by hand calculation but it involves symbolic manipulations of expressions which take a considerable time and is considered hard work.
An automatic symbolic manipulation allows to treat this problem with less hard work by defining simple procedures.
The objective of this work is to illustrate the usefulness of the automatic symbolic manipulation in the buckling analysis of elastic beams. In particular it is presented an application study done with the algebraic manipulation system *Sympy* [1], a symbolic package of the *Octave* software [2]. The symbolic-manipulation systems is essentially an expert system incorporating knowledge in the field of mathematics [3] that is used to ease the hard work.
A few examples are presented considering different boundary conditions and presented in comparison with the respective analytical solution. However, the procedures presented can be generalized and applied to many other similar eigenvalue problems.
Nowadays, numerical methods are intensively used in the structural analysis instead of the classical analytical techniques. The main reason is the power of the numerical methods to solve complex problems and its generality of application, with the price of a less fundamental understanding of the involved phenomena (including the main assumptions and theory restrictions) when compared to the traditional analytical approach. This author considered in [4] to exist " …obvious advantages of adopting a balanced philosophy—that is both efficient and rational—to engineering calculations, in which both numerical and analytical options are

retained". That balance should be somehow explored at Computational Mechanics and Structural Mechanics courses. For it, a number of symbolic-manipulation codes have been developed. Some assessments of the use of automatic Symbolic Computation in structural mechanics can be found in [3].
The choice between the systems to use can be considered arbitrary, especially at a basic level. Here, it was used the sympy package [1] for GNU Octave software [2].

2. HAND CALCULATIONS FOR THE ELASTICALLY RESTRAINED COLUMN BUCKLING

Let us consider a beam subject to an axial compressive force N (or P) and with both ends elastically restrained by longitudinal and torsional springs, as illustrated by the Fig. 1. The beam has a length L and flexional rigidity EI. The springs at both extremities have the following longitudinal elastic stiffness $\overline{\alpha_0}$, $\overline{\alpha_L}$ and torsional elastic stiffness $\overline{\beta_0}$, $\overline{\beta_L}$, where the sub-index refers to the corresponding coordinate $x=0$ and $x=L$ along the beam axis (Ox).

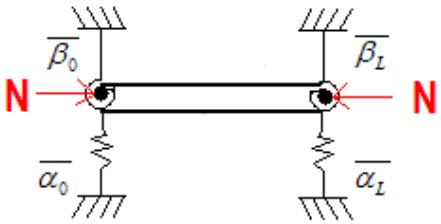


Figure 1. Beam subject to an axial compressive force N (or P) and with both ends elastically restrained by longitudinal and torsional springs.

The equilibrium in the conditions of the Euler Bernoulli beam theory is given by the following homogeneous differential equation,

$$EI\frac{\partial^4 w}{\partial x^4}+P\frac{\partial^2 w}{\partial x^2}=0 \Rightarrow \frac{\partial^4 w}{\partial x^4}+\lambda^2\frac{\partial^2 w}{\partial x^2}=0 \quad \lambda_i=\sqrt{\frac{P_i}{EI}} \tag{1}$$

which solution – transversal displacement mode w at critical load bifurcation (it is a eigenmode associated with its eigenvalue λ_i) – is given by

$$w(x)=C_1\cos(\lambda x)+C_2\sin(\lambda x)+C_3x+C_4 \tag{2}$$

where the constants C_k, are unknowns of the system of equations that are imposed by the

boundary conditions:

$$x=0 \quad \begin{cases} \dfrac{V}{EI}=\dfrac{k_0}{EI}w_0 \rightarrow -w_0'''-\lambda^2 w'=\dfrac{k_0}{EI}w_0=\alpha_0 w_0 \\ \dfrac{M}{EI}=-\dfrac{h_0}{EI}\theta_0 \rightarrow -w_0''=-\dfrac{h_0}{EI}w'_0=-\beta_0 w'_0 \end{cases} \tag{3}$$

and

$$x=L \quad \begin{cases} \dfrac{V}{EI}=\dfrac{k_L}{EI}w_L \rightarrow -w_L'''-\lambda^2 w_L'=-\dfrac{k_L}{EI}w_L=-\alpha_L w_L \\ \dfrac{M}{EI}=-\dfrac{h_L}{EI}\theta_L \rightarrow -w_L''=-\dfrac{h_L}{EI}w'_L=-\beta_L w'_L \end{cases} \tag{4}$$

$$\alpha_i=-\frac{\overline{\alpha_i}}{EI}=-\frac{k_i}{EI} \qquad \beta_i=-\frac{\overline{\beta_i}}{EI}=-\frac{h_i}{EI} \tag{5}$$

In these equations, V is the transversal (shear) effort, M is the bending moment, ()' means derivation with respect to x, and θ is the rotation of the transversal section (in this case θ=w').
The boundary conditions involve the following derivatives

$$\begin{aligned} w(x)&=C_1\cos(\lambda x)+C_2\sin(\lambda x)+C_3x+C_4 \\ w'(x)&=-\lambda C_1\sin(\lambda x)+\lambda C_2\cos(\lambda x)+C_3 \\ w''(x)&=-\lambda^2 C_1\cos(\lambda x)-\lambda^2 C_2\sin(\lambda x) \\ w'''(x)&=\lambda^3 C_1\sin(\lambda x)-\lambda^3 C_2\cos(\lambda x) \end{aligned} \tag{6}$$

Writing the four boundary conditions (3,4) with the derivatives (6), one obtains the following system of equations:

$$\begin{bmatrix} \alpha_0 & 0 & \lambda^2 & \alpha_0 \\ \lambda^2 & \beta_0\lambda & \beta_0 & 0 \\ \alpha_L\cos(\lambda L) & \alpha_L\sin(\lambda L) & \alpha_L L-\lambda^2 & \alpha_L \\ -\lambda^2\cos(\lambda L)-\beta_L\lambda\sin(\lambda L) & -\lambda^2\sin(\lambda L)+\beta_L\lambda\cos(\lambda L) & \beta_L & 0 \end{bmatrix} \begin{Bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix} \tag{7}$$

One can see that equation (7) is indeterminate, and its characteristic polynomial allows to obtain the following expression:

$$\begin{aligned} &\left\{-(\alpha_0+\alpha_L)\lambda^6+\left[\beta_0\beta_L(\alpha_0+\alpha_L)+\alpha_0\alpha_L L\right]\lambda^4+\alpha_0\alpha_L(\beta_0+\beta_L-\beta_0\beta_L L)\lambda^2\right\}\sin(\lambda L)+ \\ &+\left\{(\alpha_0+\alpha_L)(\beta_0+\beta_L)\lambda^5-\alpha_0\alpha_L L(\beta_0+\beta_L)\lambda^3-2\alpha_0\alpha_L\beta_0\beta_L\lambda\right\}\cos(\lambda L)+2\alpha_0\alpha_L\beta_0\beta_L\lambda=0 \end{aligned} \tag{8}$$

The critical load of the beam is obtained from the smallest value of λ for the specific boundary conditions (defined by the values of $\alpha_0, \beta_o, \alpha_L, \beta_L$). Considering a particular case of a clamped-free beam under compressive load, then the four boundary conditions are the following.

$$\begin{cases} w(0)=0 \rightarrow \alpha_0 = \infty \\ \dfrac{\partial w}{\partial x}(0)=0 \rightarrow \beta_0 = \infty \\ w(L)=\ldots \rightarrow \alpha_L = 0 \\ \dfrac{\partial^2 w}{\partial x^2}(L)=\ldots \rightarrow \beta_L = 0 \end{cases} \tag{9}$$

Remembering that in this case of clamped-free beam it has $\alpha_L = 0$, $\beta_L = 0$. This is introduced in (8). The meaning is that the displacement (transversal or rotation) at end of the beam is zero when the corresponding spring stiffness is infinity.

$$\begin{aligned} &\left\{-(\alpha_0+\alpha_L)\lambda^6+[\beta_0\beta_L(\alpha_0+\alpha_L)+\alpha_0\alpha_L L]\lambda^4+\alpha_0\alpha_L(\beta_0+\beta_L-\beta_0\beta_L L)\lambda^2\right\}\sin(\lambda L)+ \\ &+\left\{(\alpha_0+\alpha_L)(\beta_0+\beta_L)\lambda^5-\alpha_0\alpha_L L(\beta_0+\beta_L)\lambda^3-2\alpha_0\alpha_L\beta_0\beta_L\lambda\right\}\cos(\lambda L)+2\alpha_0\alpha_L\beta_0\beta_L\lambda=0 \end{aligned}$$

Finally, the obtained expression is divided by α_0, β_o, then the expression simplifies to:

$$\frac{1}{\alpha_0\beta_0}\left[-\alpha_0\lambda^6 \sin(\lambda L)+\alpha_0\beta_0\lambda^5 \cos(\lambda L)\right]=0$$

and remembering that β_o is infinite it simplifies to:

$$-\lambda^5 \cos(\lambda L)=0 \tag{10}$$

The equation (10) has the following solutions:

$$\lambda^5 \cos(\lambda L)=0 \Rightarrow \lambda L=(2*n-1)\frac{\pi}{2} \tag{11}$$

Assuming the following values E=73*10^9 N/m^2, I= 8.5547*10^{-8} m^4 and L=2 m, one obtains

$$P_n=\left((2n-1)\frac{\pi}{2}\right)^2\frac{EI}{L^2} \Rightarrow P_1=\pi^2\frac{EI}{(2L)^2}=3852.2N \tag{12}$$

3. SYMBOLIC COMPUTATION METHODOLOGY

In the following boxes is presented the application of the symbolic computation developed for the problem of the previous section. It can be copied from here to the command window of the symbolic tool.

At this point, it should be mentioned that the script is kept simple and to a beginner's level. For this reason, it is certainly possible to improve.

For the homogeneous isotropic beam previously presented, the solution is obtained using a symbolic computation procedure developed for a symbolic toolbox. Furthermore, to ease the presentation of the command lines, these steps are delimited in this text by a box. Here, we used the octave symbolic package *octsympy 2.5.0* [1] but others can be equally used.

```
clear all, clc
%%
% octave
%%
pkg load symbolic
%%
syms x L P EI lamb C1 C2 C3 C4 alpha0 alphaL beta0 betaL
%%
w=C1*cos(lamb*x)+C2*sin(lamb*x)+C3*x+C4
```

When these commands were executed one obtains at command windows:

w = (sym) C1*cos(lamb*x) + C2*sin(lamb*x) + C3*x + C4

Next, one compute the four equations which considers the four generic elastic (springs) boundary conditions.

```
eq1=simplify(subs(-diff(w,x,3)-lamb^2*diff(w,x,1)-alpha0*w,x,0))
eq2=simplify(subs(-diff(w,x,2)+beta0*diff(w,x,1),x,0))
eq3=simplify(subs(-diff(w,x,3)-lamb^2*diff(w,x,1)+alphaL*w,x,L))
eq4=simplify(subs(diff(w,x,2)+betaL*diff(w,x,1),x,L))
```

After these commands were executed one obtains:

eq1 = (sym) -C1*alpha0 - C3*lamb**2 - C4*alpha0

eq2 = (sym) C1*lamb**2 + beta0*(C2*lamb + C3)

eq3 = (sym) C1*alphaL*cos(L*lamb) + C2*alphaL*sin(L*lamb) + C3*L*alphaL - C3*lamb**2 + C4*alphaL

eq4 = (sym) betaL*(-C1*lamb*sin(L*lamb) + C2*lamb*cos(L*lamb) + C3) - lamb**2*(C1*cos(L*lamb) + C2*sin(L*lamb))

The equations are then separated in terms of their factors Ci to obtain the matrix of the system of equations. For it, is necessary to convert these equations to a matrix which is done by the command ***equationsToMatrix***.

```
[H, b] = equationsToMatrix([eq1==0, eq2==0,eq3==0, eq4==0 ], [C1 C2 C3 C4])
```

After this command, one obtains the following matrix H:

```
H = (sym 4x4 matrix)
 [              -alpha0                                  0                                                                  -lamb^2        -alpha0]
 [               lamb^2                               beta0*lamb                                                             beta0              0   ]
 [         -alphaL*cos(L*lamb)                      -alphaL*sin(L*lamb)                                  L*alphaL – lamb^2       -alphaL]
 [ lamb*(-betaL*sin(L*lamb) - lamb*cos(L*lamb))  lamb*(betaL*cos(L*lamb) - lamb* sin(L*lamb))        betaL          0   ]
```

3.1. Case of a Clamped-Free Beam

The determinant of this matrix gives the characteristic polynomial, which after introduction of springs values at x=L by alphaL, betaL as zero and the value of springs values at x=0, the critical load is obtained from the lowest value of lamb (higher than zero).

```
%=====================================================================
% Clamped (x=0)  *** Free(x=L)
% i.e.    alpha0=beta0=inf e  alphaL=betaL=0
%=====================================================================
solu_CL=subs(simplify(det(H)),[alphaL,betaL],[0,0])
solu_CL=expand(solu_CL/(alpha0*beta0))
solu_CL=limit(solu_CL,beta0, inf)
lamb=solve(solu_CL==0, lamb)
disp('Beam Clamped(x=0)  *** Free(x=L) axially compressed');
Pcr=(lamb(2))^2*EI
```

One can see that replacing alphaL and betaL by 0, results in the following expression:

```
- alpha0*beta0*lamb^5 *cos(L*lamb) + alpha0*lamb^6*sin(L*lamb)
```

Divinding by /(alpha0*beta0) and taking the limit when beta0 goes to infinity results in

```
  Solu_CL=  -lamb^5 *cos(L*lamb)
```

Finally solving the trigonometric equation for the first nonzero value of lamb one gets

```
lamb=solve(solu_CL==0, lamb)
lamb = (sym 3x1 matrix)
  [      0        ]
  [   pi /(2L)  ]
  [3*pi /(2L)  ]
```

```
Beam Clamped(x=0)  *** Free(x=L) axially compressed
Pcr = (sym)
    2
  pi  *EI
  ---------
       2
   4*L
```

One can observe that this results is the one obtained in (12).

3.2. Case of a Clamped–Clamped Beam

The same determinant of the matrix H can be used for other combinations of springs values (including values that are not 0 or infinity). To obtain the critical load for the case of clamped-clamped beam (e.g. clamped fixed – clamped rolling) one can use the following block of instructions.

```
%=====================================================================
% Clamped (x=0)  *** Clamped (x=L)
% i.e.    alpha0=beta0=inf e  alphaL=betaL=inf
%=====================================================================
solu_CL= simplify(det(H));
solu_CL=expand(solu_CL/(alpha0*beta0*alphaL*betaL));
solu_CL=limit(solu_CL,alpha0,inf);
solu_CL=limit(solu_CL,beta0,inf);
solu_CL=limit(solu_CL,alphaL,inf);
solu_CL=limit(solu_CL,betaL,inf)
lamb_=solve((solu_CL == 0),lamb)
%%
%% solu_CL = (sym) -lamb*(L*lamb*sin(L*lamb) + 2*cos(L*lamb) - 2)
%%
lamb_=solve((cos(L*lamb) == 1),lamb)  %%by observation lamb=(2*pi/L) is a solution
%%
disp('Beam Clamped(x=0)  *** Clamped rolling (x=L) axially compressed');
Pcr=(lamb_(2))^2*EI
```

The obtained result is

```
solu_CL = (sym) -lamb*(L*lamb*sin(L*lamb) + 2*cos(L*lamb) - 2)
Beam Clamped(x=0)  *** Clamped rolling (x=L) axially compressed
Pcr = (sym)
        2
  4*pi *EI
  --------
     2
```

Of course, depending on the solver or even on previous simplifications, in some cases we do not get all the computations automatically. But, the symbolic expression obtained is sufficiently simplified to allow, in this case, to overcome this last step, as illustrated in this case with observation that $cos(\lambda L)=1$.

5. CONCLUSIONS

In this paper authors present a simple script of an algebraic manipulation for the buckling problems of ideal column/beams under compression. The solution assumes elastically restrained ends of the column, which has been shown to take great advantage of symbolic manipulation in the extreme case of springs with infinity or even zero rigidity.

The main conclusion is that the proposed algebraic manipulation scripts allows that an intensive symbolic computation, like this buckling analysis, to be handled with a reasonable effort even for a beginner in symbolic computation.

However, one should be aware that in some cases one will not get all the computations automatically, depending on the solver or even on previous simplifications.

In the presented application we treated a problem with a closed form solution which is a good start for teaching symbolic computation programming in courses like solid mechanics, computational mechanics, structural mechanics and others. The authors also remember that the potential of these tools can be applied in many other areas inclusive in areas were the problems have no closed form solution but approximate solution that can be formulated and developed in most of it with symbolic computation (see other presentation in this symcomp 2017 from one of the authors [7], and also [8-10]).

REFERENCES

[1] *Octsympy*, Octave Forge new symbolic package. The development site is at: [Online] https://github.com/cbm755/octsympy. octsympy 2.5.0 [Accessed Feb. 2, 2017]

[2] GNU *Octave*, Scientific Programming Language, [Online] https://www.gnu.org/software/octave/ [Accessed Jan 29, 2017]

[3] Noor, K, Andersen, C. M. "Computerized symbolic manipulation in structural mechanics-Progress and potential." *Comput. Structures* 10, pp 95-118, 1979.

[4] Pavlovic, M.N. "Symbolic computation in structural engineering", *Computers and Structures* 81, pp. 2121–2136, 2003.

[5] Rizzi, NL; Tatone, A. "Symbolic manipulation in buckling and postbuckling analysis", *Computers & Structures* Vol. 21 (4), pp. 691-700. 1985.

[6] Rizzi, NL; Tatone, A. "Using Symbolic computation in analysis", *J Symbolic computation*, Vol. 1, pp. 317-321. 1985

[7] Neves, M.M. "On an application of symbolic computation to derive a double scale asymptotic technique for linear-buckling of periodic microstructures", in *proceedings of SYMCOMP 2017*, Guimarães, 6-7 April 2017, ©ECCOMAS, Ed. Amélia Loja and Stéphane Clain, Portugal, 2017.

[8] Policarpo, H.; Neves, M.M.; Maia, N.M.M. "On a hybrid analytical–experimental technique to assess the storage modulus of resilient materials using symbolic

computation", Journal of Symbolic Computation, Vol. 61–62, pp 31–52, (2014).

[9] Policarpo, H., Neves, M.M.; Using symbolic computation for teaching structural mechanics I: Frames, in Proceedings of CSEI2012 – Conferência Nacional sobre Computação Simbólica no Ensino e na Investigação, Lisboa, 2-3 April 2012, Ed.s Joaquim Infante, 2012.

[10] Neves, M.M.; Policarpo, H.; Matos, M.A.A.S. "Automatizing symbolic computations for the elasticity direct method to obtain the three dimensional displacement field of a bar under uniaxial tension", in proceedings of SYMCOMP 2013, 1st International Conference on Algebraic and Symbolic Computation, Lisbon Portugal, 9-10 September, ECCOMAS, Ed. Amélia Loja, Joaquim Infante Barbosa, and José Alberto Rodrigues, Portugal, 2013.

APPENDIX

The full script for the section 3.2. is listed here.

```
 clear all, clc
%%
 % octave
%%
pkg load symbolic

syms x L P EI lamb C1 C2 C3 C4 alpha0 alphaL beta0 betaL
w=C1*cos(lamb*x)+C2*sin(lamb*x)+C3*x+C4

eq1=simplify(subs(-diff(w,x,3)-lamb^2*diff(w,x,1)-alpha0*w,x,0))
eq2=simplify(subs(-diff(w,x,2)+beta0*diff(w,x,1),x,0))
eq3=simplify(subs(-diff(w,x,3)-lamb^2*diff(w,x,1)+alphaL*w,x,L))
eq4=simplify(subs(diff(w,x,2)+betaL*diff(w,x,1),x,L))

 [H, b] = equationsToMatrix([eq1==0, eq2==0,eq3==0, eq4==0 ], [C1 C2 C3 C4]);

%=====================================================================
%=====================================================================
% Clamped (x=0)  *** Clamped (x=L)
% i.e.    alpha0=beta0=inf e  alphaL=betaL=inf
%=====================================================================
solu_CL= simplify(det(H));
solu_CL=expand(solu_CL/(alpha0*beta0*alphaL*betaL));
solu_CL=limit(solu_CL,alpha0,inf);
solu_CL=limit(solu_CL,beta0,inf);
solu_CL=limit(solu_CL,alphaL,inf);
solu_CL=limit(solu_CL,betaL,inf)
lamb_=solve((solu_CL == 0),lamb)
%% solu_CL = (sym) -lamb*(L*lamb*sin(L*lamb) + 2*cos(L*lamb) - 2)
lamb_=solve((cos(L*lamb) == 1),lamb)  %%by observation
disp('Beam Clamped(x=0)  *** Clamped rolling (x=L) axially compressed');
Pcr=(lamb_(2))^2*EI
```

SYMCOMP 2017
Guimarães, 6-7 April 2017

CONDUCT RISK: DISTRIBUTION MODELS WITH VERY THIN TAILS

Peter Mitic[1, 2]

1: Santander UK
2, Triton Square, Regent's Place, London NW1 3AN
e-mail: peter.mitic@santandergcb.com
2: Department of Computer Science, University College London
Gower Street, London WC1E 6BT

Keywords: conduct risk, Mathematica, R, capital value, value-at-risk, VaR, regulation, fat-tailed, $\exp(-x^4/2)$, goodness-of-fit, TNA

Abstract

Regulatory requires dictate that financial institutions must calculate risk capital (funds that must be retained to cover future losses) at least annually. Procedures for doing this have been well-established for many years, but recent developments in the treatment of conduct risk (the risk of loss due to the relationship between a financial institution and its customers) have cast doubt on 'standard' procedures. Regulations require that operational risk losses should be aggregated by originating event. The effect is that a large number of small and medium-sized losses are aggregated into a small number of very large losses, such that a risk capital calculation produces a hugely inflated result. To solve this problem, a novel distribution based on a probability density with an $\exp(-x^4/(2s^2))$ component is proposed, where s is a parameter to be estimated. Symbolic computation is used to derive the necessary analytical expressions with which to formulate the problem, and is followed by numeric calculations in R. Goodness-of-fit and parameter estimation are both determined by using a novel method developed specifically for use with probability distribution functions. The results compare favourably with an existing model that used a LogGamma Mixture density, for which it was necessary to limit the frequency and severity of the losses. No such limits were needed using the $\exp(-x^4/2)$ density.

Disclaimer

The opinions, ideas and approaches expressed or presented are those of the author and do not necessarily reflect Santander's position. As a result, Santander cannot be held responsible for them. The values presented are just illustrations and do not represent Santander losses.

1. INTRODUCTION

Regulated financial institutions (hereinafter referred to as 'banks') are required annually to assess the amount of capital that must be retained to cover operational risk losses that might be suffered in the following year. The European Banking Authority (EBA) defines *Operational Risk* as "*the risk of losses stemming from inadequate or failed internal processes, people and systems or from external events. Operational risk includes legal risks but excludes reputational risk and is embedded in all banking products and activities.*" [1]. Informally, operational risk is the risk of "things going wrong". Such capital must be retained by the bank and cannot be used for lending. The amount retained should be enough to cover expected losses, which are typically a mixture of known liabilities and averages based on known losses for prior years. In addition, the amount retained should also include an estimate to cover 'unexpected' losses that cannot be anticipated as individual amounts. The amounts actually retained vary greatly depending on the size of the bank: from a few hundreds of millions of euros for a small retail bank to many billions for a large international investment bank. The requirement to retain funds, but also and to make funds available for lending and investment, always creates conflict: the two activities are contradictory. A balance must therefore be struck when calculating the amount of capital to retain. It should be enough to satisfy regulations, but should not so excessive as to hinder business activity. The calculation of the capital amount is therefore an important part of a bank's risk control function.

The Basel Committee on Banking Supervision (often known as the "Basel Committee") is the originating organisation for regulations that govern the management of operational risk, and for regulations on the calculation of operational risk capital [2]. At the time of writing three principle streams for calculating of operational risk capital are in operation. The *Basic Indicator Approach* is based on annual revenue, whereas the *Standardized Approach* uses annual revenue of business lines. The third stream is the Advanced Measurement Approach (AMA), in which each bank is permitted to develop its own risk model, provided they are consistent with the regulations in [2]. The model proposed in this paper falls into the AMA category. Those regulations give a very strong indication of what has to be done, and what might be described as a "standard model" has emerged. Such a "standard model" typically has the following components:

- A statistical model of severity (i.e. magnitude) of losses, based on a 'fat-tailed' distribution (i.e. one for which the probability that a very large loss occurs is relatively large – a precise definition will be given in the Methodology section)
- A statistical model of loss frequency: often Poisson
- A convolution model to combine the severity and frequency models, and extract the 99.9% value-at-risk (VaR). VaR can be thought of as 99.9 percentile loss, the figure being specified by the Basel regulations [2]. The precise method is given in the

Methodology section.

Conduct Risk is a significant component of operational risk. Some banks treat it as a separate entity, but a capital calculation for it is needed whatever its classification. The Basel regulations define a risk class taxonomy, and conduct risk is part of the 4th category in Table 1 below.

Basel ID	*Basel Category*	*Abbreviation*	*Components*
1	Internal Fraud	IF	misappropriation of assets, tax evasion, bribery
2	External Fraud	EF	hacking, third-party theft, forgery
3	Employment Practices and Workplace Safety	EPWS	discrimination, compensation, employee health and safety
4	Clients, Products, and Business Practice	CPBP	market manipulation, antitrust, improper trade, product defects, legal actions
5	Damage to Physical Assets	DPA	natural disasters, terrorism, vandalism
6	Business Disruption and Systems Failures	BDSF	disruptions, software and hardware failures
7	Execution, Delivery, and Process Management	EDPM	data entry errors, accounting errors, failed mandatory reporting, negligent loss

Table 1. Basel Risk Class taxonomy

Section 2 gives more details of the nature of conduct risk.

2. CONDUCT RISK

"Conduct Risk" (CR) may be regarded as "risk that arises as a result of how firms and employees conduct themselves, particularly in relation to clients and competitors" [3]. Its essential element is how a bank behaves with respect to its customers and other stakeholders, and encompasses items such as compensation resulting from complaints, regulatory fines and costs of mis-selling. Mis-selling of Payment Protection Insurance (PPI) is a significant part of CR losses in the UK. They, in particular, are the losses that drive the proposed model in this paper.

2.1. The Source of Conduct Risk Losses

Thomson Reuters [4] gives a somewhat alarming report on how a selection of firms view CR: "*81 percent of firms remain unclear about what conduct risk is and how to deal with it*." This report gives no indication of the sample size, but it does signal that problems exist. It does not say that there is a general problem in how losses should be classified within the taxonomy in Table 1. It is not always clear which category a particular loss should be allocated to. Such classification problems are mostly invisible to the statistical modeller, who sees the end result: a sequence of numbers with a timestamp and a Basel risk category for each. Thomson Reuters [5] lists, according to their respondents, the principal sources of conduct risk. In terms of frequency of response, the most significant five are (in decreasing order of frequency):

1. culture;
2. corporate governance;
3. conflicts of interest;
4. reputation;
5. sales practices.

Of these, all except "reputation" enter into the modelling process. The remaining categories map to the Basel sub-category "improper trade" in Table 1. Payment Protection Insurance is easy to identify as belonging to the CPBP category. It is a particular problem in the UK, where customer compensations have been huge. The Financial Conduct Authority (FCA) defines: "*Payment Protection Insurance (PPI) is designed to cover your debt repayments in circumstances where you aren't able to make them such as accident, redundancy, or illness.*" [6]. It was found that customers were sold PPI in conjunction with other financial products (loans, mortgages etc.) without telling them that PPI was an optional extra, and without explaining the product adequately. The Guardian [7] reports massive PPI payouts in Figure 2. Note that the amounts in the graphic sum to £39.7bn, to which £600m should be added (the latest payout at the time from Barclays Bank and the subject of the article, making a total of £40.3bn.

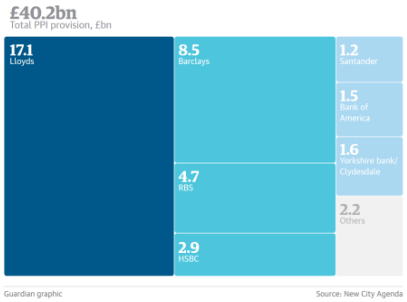


Figure 1. PPI payouts to October 2016

2.2. Characteristics of Conduct Risk Losses

Conduct risk losses comprise, typically, thousands of small- to medium-sized-payments (ranging from tens of pounds to a few thousands) made to individual customers. In addition there will be costs such as legal fees and regulatory fines. Both of those could be large: of the order of millions of pounds. As such, 'fat-tailed' distributions serve very well in capital models. They can model precisely the situation that exists: a very large number of small losses totalling a small proportion of the total loss, with a small number of large losses making up the balance of the total. It is not unknown for the largest 5% of losses to be worth 95% of the total loss. Compared with other categories of operational risk losses, conduct risk losses in the form of customer payments are an extreme. Their annual frequency is higher, their mean loss is smaller, and the calculated capital tends to be larger due to the high frequency.

The Basel regulations [2] direct that operational risk losses should be aggregated by originating event. Such aggregation is apparently inconsistent with using individual payments to customers. Originating event is not the only way to aggregate, although others may not be Basel-compliant. Consider the case where a decision to market a product is distributed regionally throughout a firm, and that the distribution is rolled out over an extended period covering several years. Each individual action within the roll-out scheme could be considered a root event for all the losses arising from it. The losses are therefore grouped by time and by region. The rollout could be further subdivided by branch, sales force or customer category. That adds another layer of potential root events. These simple considerations show that the term *root event* is not well defined. There is a fuller discussion of this topic in [8].

Against the argument in favour of aggregation is the view that an originating event should be the contract made between a bank and each individual customer. Without such a contract, there would be no breach of contract and no compensation payment due to inappropriate conduct. In the absence of any specific regulatory directive, most banks take the view that conduct risk losses should be aggregated, either by originating event, or by accounting period.

The result of aggregation of conduct risk losses is a very small number of very large 'losses', since common practice is to define very few originating events. Hundreds of thousands of small losses are thereby condensed into only ten, twenty or thirty 'losses'. This causes considerable capital modelling problems because existing 'fat-tailed' capital models do not cope well with a small number of large losses. Effectively the small number of large losses comprises the tail of a distribution (i.e. very high losses that occur with a very small probability), and 'fat-tailed' models estimate values in the tail of a tail, which are "super high"! They are effectively outliers of a set that does not exist.

In principle, conduct risk may be modelled in exactly the same way as any other operational risk, provided that there are not too few losses to find a reasonable distribution fit. The result of calculating capital using only the tail of a distribution is a grossly inflated value. To give an example, we recently modelled a set of about 190000 conduct risk payments which had been

aggregated into 30 aggregated losses. The capital calculated from the payments data was in the order of £260m, but the capital calculated from the aggregations was near to £3700m. If we summarise the aggregation process, we could say that the aggregation takes us from one extreme to the other: a large number of small losses to a small number of large losses.

3. MODELLING METHODOLOGY

This section describes the basis of the standard operational risk capital calculation. The method depends on modelling loss severity by an appropriate probability distribution. The advance presented in this paper is to find a probability distribution that models the severity of conduct risk losses in a more satisfactory way than 'fat-tailed' distributions. In all cases, regulatory capital is assessed in the same way, as described in the next section.

3.1. Data

This section illustrates some examples of aggregated conduct risk losses. Figure 2 shows three distinct examples. The histograms show single instances of very large losses covering a five year period. Compared with other operational risk losses, all are outliers in the sense that all are atypically large.

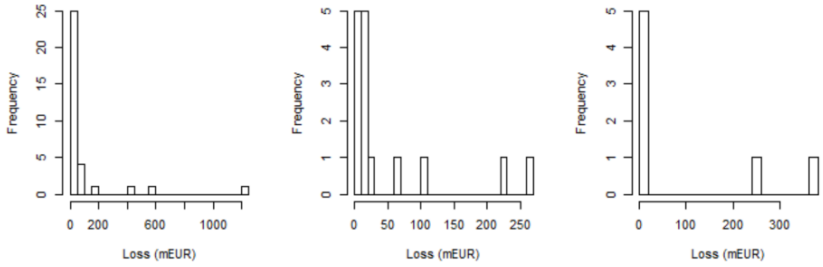


Figure 2. Typical conduct risk loss histograms

The minimum, maximum and mean losses (in mEUR) for the examples in Figure 2 are:

Left-hand data set:	(0.1, 1208.0, 92.7)
Middle data set:	(1.0, 264.2, 52.5)
Right-hand data set:	(0.6 362.0, 91.7)

Using these and similar data sets causes difficulties for determining parameters of a severity distribution. Iterative methods, such as *maximum likelihood*, do not always converge, and if they do, they sometimes converge to values that are clearly wrong. Usually such a failure is due to a flat feature space. Gradient search methods produce essentially random results because the result of each iteration is very similar to the result of the previous iteration.

3.2. The Capital Calculation

The capital calculation assumes that a suitable severity and a suitable frequency distribution has been fitted to empirical data. Any such pair of distributions should have passed an appropriate goodness-of-fit (GoF) test. The method described in this section is due to Frachot et al [9], and is known as the Loss Distribution Approach (LDA). It is a Monte Carlo process based on a convolution of the severity and frequency distributions. Other methods are available, notably Panjer recursion and the Fast Fourier Transform (FFT). See [10] for details of all three, and also of how they are used in the general context of operational risk.

Algorithm LDA is a summary of the LDA method. For a set of N losses covering a period y years, and let n be the number of Monte Carlo trials used. Further, let the fitted severity distribution to those losses be $F(l, \boldsymbol{p})$, where l is a loss and $\boldsymbol{p}$ is a vector of parameters obtained from the fit.

Algorithm LDA

a) Calculate an annual loss frequency, $\lambda = N/y$
b) Repeat n times
   i) Obtain a sample size z by drawing a random sample of size 1 from a Poisson(λ) distribution (*)
   ii) Draw a sample of size z, S_z, from the severity distribution, F.
   iii) Sum the losses in S_z to obtain $\sum S_z$, the estimate of annual loss.
c) End Repeat
d) Calculate the 99.9^{th} percentile of the retained sums $\sum S_z$ (99.9 is the percentage specified by the Basel regulations [2].

(*) A Poisson distribution is not the only way to do this. A negative binomial distribution is often used. For large λ, a normal approximation to the binomial is appropriate.

Algorithm LDA is easy to implement and is applicable for any severity distribution. In practice, n might exceed one million before stability of the result is established. In that case, the Monte Carlo process might take one or two hours. In most cases 100000 iterations suffice and the calculation takes minutes.

3.3. Fat-tailed distributions

An account of "fat-tailed" distributions may be found in [11, 12]. Equation (1) gives a characterisation for the density function $f(x)$ and the corresponding distribution function $F(x)$ for a random variable in terms of a parameter a. For large x, these functions are polynomial-like rather than exponential-like.

$$\left.\begin{array}{ll} f(x) \sim x^{-(1+a)} & a > 0;\ x \to \infty \\ F(x) \sim x^{-a} & a > 0;\ x \to \infty \end{array}\right\} \quad (1)$$

Typical examples of "fat-tailed" distributions are the LogNormal, with distribution function (in terms of the Normal distribution function Φ) $F(x) = \Phi\big((\ln(x) - \mu)/\sigma\big);\ \mu \in \mathbb{R}, \sigma > 0$, and the Weibull, with distribution function $F(x) = 1 - e^{\left(\frac{x}{\theta}\right)^{\tau}};\ 0 < \tau < 1, \theta > 0$. Normalisation factor for density. The Pareto distribution, $F(x) = 1 - \left(1 + \frac{x}{\theta}\right)^{-\alpha};\ \theta, \alpha > 0$ is particularly troublesome because it nearly always returns very high 99.9% VaR values.

The 99.9% VaR values obtained using "fat-tailed" distributions with data as illustrated in section 3.1 are always intuitively unacceptably high, which prompts a search for an alternative. In many cases the capital value is orders of magnitude greater than the expected value. *Algorithm LDA* works well using a "fat-tailed" distribution for data where there is a mixture of small- to mid-value losses with some very large losses. Aggregated losses effectively constitute the tail of a distribution that has a missing body.

3.4. Thin- and Very-thin-tailed distributions

In order to use *Algorithm LDA* with aggregated conduct risk losses, the required distribution is exactly the opposite of a "fat-tailed" distribution: namely a "thin-tailed" distribution. The distinguishing characteristic of such a distribution is that the probability of generating an extreme value in a random sample should be much less than the probability of generating an extreme value using an exponential-based distribution (the "low loss = high probability; high loss = low probability" criterion). A first approximation to a "thin-tailed" distribution is the Normal distribution. This proved to be easy to do but still resulted in a capital value which was unreasonably large. An alternative is distribution which resembles the Normal distribution but was capable of generating relatively fewer very high value losses in a random sample. In parallel there should also be a relatively high probability of generating lower valued losses. Such a distribution can be obtained, by replacing the << $e^{-\frac{x^2}{2}}$ >> term in the standard Normal density by << $e^{-\frac{x^4}{2}}$ >>, and by choosing a suitable domain for x.

The density plot for such a distribution resembles that of the Normal distribution, but decays much faster for losses that are a large distance from the mean loss. This is a big advantage but the disadvantage is that the probability of generating very small losses in a random sample is small. The actual density proposed introduces a scale parameter *s*, to be determined from the data, so that the non-normalised density contains << $e^{-\frac{x^4}{2s^4}}$ >>. The resulting density is termed the *Exp4* distribution, and its formal probability density function *f*(*x*,*s*) is given in equation (2).

$$f(x,s) = \frac{1}{s2^{\frac{1}{4}}\Gamma\left(\frac{5}{4}\right)} e^{-\frac{x^4}{2s^4}}; \quad s > 0, x > 0 \qquad (2)$$

The normalising factor $s2^{\frac{1}{4}}\Gamma\left(\frac{5}{4}\right)$ is produced by Mathematica, and the final expression for *f*(*x*,*s*) requires inclusion of appropriate assumptions to make the result useable. Appendix A shows the details, and an illustrative graph. The density definition shown is the most convenient way to formulate *f*(*x*,*s*). The fourth power in the exponential term ensures that the probability of generating very high values in a random sample is low, and the domain $x>0$ ensures that the probability of generating low values in a random sample is high (as the density graph in Appendix A shows).

The corresponding cumulative distribution function, *F*(*x*,*s*) is given in equation (3).

$$F(x,s) = \int_0^x f(t,s)dt = 1 - \frac{1}{\Gamma\left(\frac{1}{4}\right)}\Gamma\left(\frac{1}{4}, \frac{x^4}{2s^4}\right); \quad s > 0, x > 0 \qquad (3)$$

The integral of *f*(*x*,*s*) in equation (3) is again provided by Mathematica, which requires explicit assumptions $s > 0$ and $x > 0$ to avoid generating an unwieldy expression involving an exponential integral (function ExpIntegral[] in Mathematica) See Appendix A for the appropriate expressions. In equation (3), the two-parameter form of the Gamma function is the *upper incomplete gamma* function, $\Gamma(a,x) = \int_x^\infty t^{a-1}\, e^t dt$. See [13] for details. The shape of the *Exp4* density and distribution functions is unlike those of "fat-tailed" distributions. Most notably, "fat-tailed" distributions typically have near vertical slopes for small losses. The *Exp4* distribution is set up to use scaled losses: specifically each loss divided by the mean loss. This scaling ensures that the arguments *x* in (2) and (3) are small positive real numbers which are consistent with the domains indicated.

In the R statistical language, which was used for the numerical calculations in this paper, the *upper incomplete gamma* function is implemented in the *gsl* package by the *gamma_inc*() function. It is customary in R to define four functions per probability distribution: one for the density, one for the distribution, one for the inverse distribution, and one for random number generation. The names of these functions are, by convention, prefixed by *d*, *p*, *q* and *r* respectively. The R implementation for the *Exp4* set is given in Appendix B.

3.5. Parameter Estimation and Goodness-of-fit

Given the *Exp4* distribution and data as described in the previous sections, the usual way to proceed is to estimate the distribution parameters, and use them in a GoF test to assess the quality of the fit. Attempting fit the *Exp4* distribution to aggregated loss data using standard maximum likelihood methods proved to be difficult in R. Parameter values that were clearly incorrect were often returned. This was attributed to using a gradient search in a parameter space for which even large changes in parameter values had little effect on an objective function. This situation can be likened to an attempt to climb hills in an almost flat landscape. A search for an optimal solution proceeds in an almost random direction, and any 'optimal' solution found is only locally optimal.

As an alternative, parameter optimisation was combined with a GoF test using the author's *TN* ("Transformed Normal") method [14]. There are three *TN* tests, and the simplest conceptually and in practical terms is the *TN-A* test. All were originally intended as dedicated GoF tests for cumulative distribution functions obtained from fitting "fat-tailed" distributions to operational risk data. The *TN-A* test has the advantages the following advantages over 'traditional' GoF tests such as Anderson-Darling (AD) or Kolmogorov-Smirnov (KS).

- It satisfies the criterion "*if a fit looks good, the test should say so*". The AD and KS tests were found to reject distributions that that were intuitive good fits, especially for large data sets.
- It is independent of the number of data points.
- It is entirely deterministic.
- The value of the *TN-A* statistic (i.e. the test's objective function) is a direct measure of the goodness-of-fit. The smaller then *TN-A* value, the better the fit. In contrast, the AD and KS statistic values only indicate whether or not a null hypothesis should be rejected.

The way in which the *TN-A* statistic is formulated implies that it can also be used a data fitting methodology. That is the approach used to fit the *Exp4* distribution to conduct risk loss data, and is the first time the *TN* formulation has been used in this way. The following section therefore explains the basic concepts of the *TN* formulation.

3.5.1. The *TN-A* method: goodness-of-fit

Given a set of n real numbers $X = \{ x_1, x_2, \ldots, x_n \}$ (which are the aggregated conduct risk losses), assign a probability y_i to each loss x_i, where $y_i = (i - 0.5)/n$. The probability y_i is this a cumulative

probability. The set $D = \{x_i, y_i\}_{i=1}^{n}$ then defines the cumulative empirical distribution of X. the requirement is to fit a probability distribution to D.

Suppose that, in general, a fit is proposed using a distribution function $F(x, \boldsymbol{p})$, where $\boldsymbol{p}$ is a vector of parameters *that has already been* determined. If the fit is reasonable, a plot of the set D superimposed on a plot of $F(x, \boldsymbol{p})$ should resemble the left-hand part of Figure 3. Notice the relative positions of the points of D to the curve defined by $F(x, \boldsymbol{p})$. There are consecutive chains of points that are either below or above the $F(x, \boldsymbol{p})$ curve. If the points in these chains are joined by straight lines, very few of those joining lines cross the $F(x, \boldsymbol{p})$ curve. That is an accurate assessment of what happens with actual data. The combination $\boldsymbol{L} = \{D, F(x, \boldsymbol{p})\}$ was referred to as *Loss-Space* in [14], because actual losses are involved. The right-hand part of Figure 3 will be explained below. It is obtained by applying a transformation T to the points in $\boldsymbol{L}$, thereby deriving *Probability-Space* (so-called because the transformation is based on a probability measure.

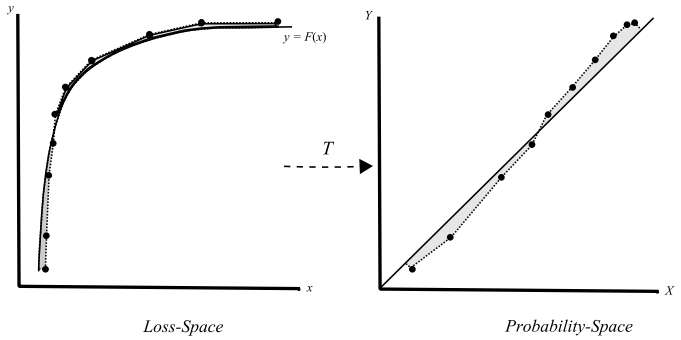


Figure 3. *Loss-Space* transformed to *Probability-Space*

Consider now the effect of the mapping T, defined in equation (4), below.

$$\left.\begin{array}{c} T: \{\mathbb{R}^{+} \otimes (0,1)\} \rightarrow (0,1)^2 \\ T(x,y) \rightarrow (F(x,\boldsymbol{p}), y) = (X,Y) \end{array}\right\} \tag{4}$$

Under T, the elements y_i in D are unchanged, other than to be renamed Y_i. The elements x_i in D are mapped by the distribution function, and are labelled X_i. The result is that *Probability-Space* is the unit square. the $F(x, \boldsymbol{p})$ curve always maps to the line $Y = X$. This makes it much easier to analyse the "Enclosed Area" in *Probability-Space*, which is the area enclosed between the line $Y = X$ and the mapped points $T(x_i, y_i)$ once it is joined by straight line segments. This is the shaded area in the right-hand part of Figure 3. Calculating that "Enclosed Area" in *Probability-Space* is very easy. The smaller it is, the better the fit. That is the essence of the

TN group of GoF tests.

The significance level for a data fit using the *TN-A* test can be calculated very easily. An elaborate topological argument in [14] leads eventually to a simple quadratic expression for the significance level (equation 5). In (5), p is a probability and $A(p)$ is the "Enclosed Area".

$$A(p) = 2\sqrt{2}p\left(1 - \sqrt{2}p\right) \tag{5}$$

To obtain a p% 2-tailed significance level for the *TN-A* statistic, the value $p/2$ (expressed as a probability rather than a percentage) should be used in (5). For example, the 5% significance level is calculated using $p = 5/(2\times100) = 0.025$, for which $A(0.025) = 0.0682$. If the calculated "Enclosed Area", $\tilde{A}$, is less than 0.0682, the fit "passes the significance test at 5%". A more rigourous restatement of that phrase is given in Appendix C.

3.5.2. The *TN-A* method: parameter estimation

Focussing now on the best fit itself, by supplying values for the *Exp4* distribution parameter s, an array of *TN-A* values can be built, and the minimum selected. The value of s corresponding to the minimum *TN-A*, $\hat{s}$, is then the required parameter value. In practice $\hat{s}$ is found by starting with an initial estimate, followed by an ordered search. In practice the ordered search is likely to be embedded in a function call. In R this is the function optimize(), and its equivalent in Mathematica is FindMinimum[].

Note, however that using the *TN-A* test to provide a best fit only works well in the current context because only one parameter needs to be estimated. Possibly, searching a bivariate parameter space would also work reasonably well, but would take longer. To search a parameter space with more than two parameters would probably require a sophisticated algorithm in order to search efficiently.

Using the *TN-A* method for parameter estimation as well as GoF automatically produces an optimal fit with an assessment of how good that fit is.

3.6. Extension to higher powers of *x/s*

Use of a density containing an << x^4 >> term prompts the question of whether or not a higher power of x would be appropriate. One has to be careful not to over-fit in these circumstances. It may be possible to find a distribution that either fits the data better or produces a lower capital value, but that distribution type may not work so well with other data sets. To see the effect of an extension if the *Exp4* distribution to higher powers of x, the same procedures as were used in section 3.4 were applied to higher even powers of x.

Equations (6) and (7) define the density *fn*() and distribution *Fn*() functions respectively for the required generalisation: the *ExpN* distribution. As with the *Exp4* distribution, Mathematica was used to derive them.

$$fn(x,s) = \frac{1}{s2^{\frac{1}{n}}\Gamma\left(1+\frac{1}{n}\right)} e^{-\frac{x^n}{2s^n}}; \quad s>0, x>0, n\epsilon\mathbb{Z}^+ \tag{6}$$

$$Fn(x,s,n) = 1 - \frac{1}{\Gamma\left(\frac{1}{n}\right)}\Gamma\left(\frac{1}{n}, \frac{x^n}{2s^n}\right); \quad s>0, x>0, n\epsilon\mathbb{Z}^+ \tag{7}$$

A quick check shows that when $n = 4$, *ExpN* reduces to *Exp4*. A useful test of the applicability of the *ExpN* distribution is to consider the 99.9 percentile for selected values of *n*. Figure 4 shows a comparison of these percentiles for even values of *n* between 6 and 20, together with the corresponding *Exp4* percentile. It is clear from Figure 4 that there is a reduction in the 99.9 percentile value as *n* increases, but that the rate of decreases diminishes. Very little further decrease is observed for $n \geq 12$. Therefore the case $n = 12$, is a potential rival to the *Exp4* distribution, but there is a warning against using it in Section 5.

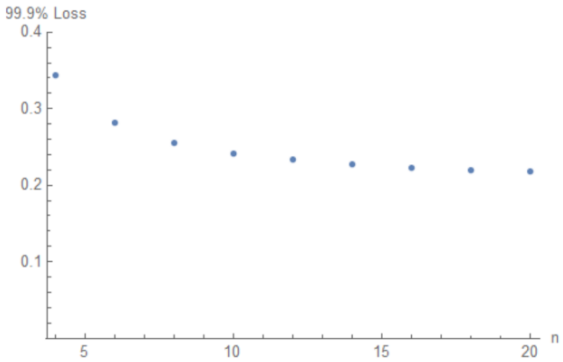


Figure 4. Variation of *ExpN* 99.9 percentiles with *n*

4. RESULTS

Capital values for each of twelve sets of aggregated conduct risk losses were calculated using Algorithm LDA and the *Exp4* distribution. Histograms of three of them are shown in Figure 2, section 3.1. All twelve have the same characteristics of having only a few large losses, with no small or medium sized losses. Simple statistics (number and sum of losses) are indicated in Table 2, below. Table 2 also shows the capital values (99.9% VaR) for each of them with their *TNA* test values. In the previous section the *ExpN* distribution with parameter value $n = 12$ was introduced as a potential alternative to *Exp4*. Instead of restricting *n* to a modestly low value, consider a much higher value, $n = 100$, as a proxy for infinity.

Therefore, as a comparison, the Table 2 also shows capital values obtained by fitting a Normal distribution and the *ExpN* distribution with $n = 100$ (i.e $n \to \infty$) to each data set. The sum and capitals shown are in millions of euros.

Data Set	Count	Sum	Capital	TNA	Normal Capital	*ExpN* ($n \to \infty$) Capital
1	29	1876.8	539.2	0.110	1825.7	516.2
2	21	2110.9	395.3	0.074	2673.4	365.7
3	33	3059.4	503.8	0.098	3045.6	471.2
4	17	2287.2	1670	0.031	1728.5	1557.1
5	29	3437.9	1390.1	0.036	2411.5	1295.3
6	14	2039.1	1307.5	0.074	1958.4	1141.7
7	40	2680.9	968.6	0.118	1827.5	908.7
8	13	2329.3	1617.7	0.040	2099.2	1495.9
9	15	787.3	170.6	0.094	871.9	147.8
10	7	642.2	1276.7	0.240	1100.1	1179.9
11	10	1385.7	993.9	0.073	1508.0	891.1
12	11	391.4	235.3	0.101	450.6	262.8

Table 2. Capital Values using the Exp4 and Normal distributions

Given the low number of data points, it was anticipated that goodness of fit would, in general, be poor. In reality, a more optimistic result emerged. The significance levels for a two-tailed test were all below the 10% level (i.e. 5% in each tail), and many of them were within the 5% limit (2.5% in each tail). These results indicate that The *Exp4* distribution is, indeed, a good fit to the data. There is one exception: data set 10. This has only 7 aggregated losses, and the *Exp4* distribution is biased towards the lower values. In the *LDA* random sampling, very high values still figure significantly, and these inflate the capital value. Clearly *Exp4* is inadequate for this particular data set. Figure 5 shows the densities for best and worst *Exp4* fits (left-hand is best, right-hand is worst). The blue profiles are the *Exp4* densities and the red profiles are empirical densities.

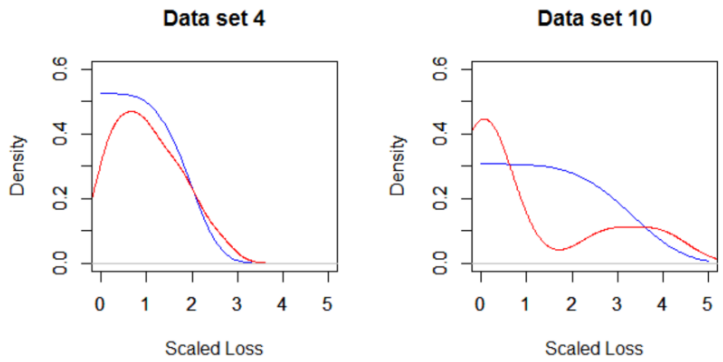


Figure 5. Comparison of *Exp4* and Normal densities for best and worst fits:
(LHS=Best, RHS=Worst; blue = *Exp4*, red = empirical)

In Figure 5, the 'worst' fit fails so badly because the empirical distribution is bimodal. A possible better fit is to use an *Exp4* mixture with density of the form $(1\text{-}p)\times Exp4(s_1) + p\times Exp4(s_2)$, where p is a real-valued weight in the range (0,1), and s_1 and s_2 are the parameters of two *Exp4* distributions. The idea of a mixture distribution was used in [15], but with Normal distributions. In that case each loss was replaced by random samples from Normal distributions. The results were only just acceptable, and replacing the Normal distribution by *Exp4* distributions is a possible way forward. The 'best' fit works well because the bulk of the probability mass falls in the part of the domain of the *Exp4* distribution that can generate the highest probability in random sampling. The 'best' fit is a good illustration of the main criteria for a distribution to work well with aggregated losses: a high probability of generating lower valued losses with a low probability of generating higher valued losses.

The extended distribution *ExpN* with $n \to \infty$ is an acceptable fit in most cases. A general trend with the *ExpN* distribution is for the capital values to decrease and the *TN-A* measures to increase as n increases, but both at a decreasing rate of change. If $n > 12$ the gain in lower capital becomes comparable with the stochastic error inherent in the *LDA* process.

All capitals *ExpN* in Table 2 are less than their corresponding *Exp4* capitals except for Data Set 12. It is likely that sufficient medium-sized losses are sampled in the *ExpN* case ($n\to \infty$) to inflate the *ExpN* capital. The mean reduction, excluding Data Set 12, is 8.1%.

5. DISCUSSION

This research was prompted by the unsuitability of 'traditional' "fat-tailed" distributions for modelling conduct risk losses. Regulations [1, 2] require operational risk losses to be aggregated by root event, and for conduct risk losses the result is a very small number of huge losses. Effectively the loss distribution has lost its body and is left with only a tail. The resulting capital values obtained using the LDA process (*Algorithm LDA*, section 3.2) are always much greater than is merited by anticipated future losses.

A suggested solution is to use exactly the extreme opposite of a "fat-tailed" distribution, namely a "very-thin-tailed" distribution. Such a class of distribution decays faster than the Normal distribution for large losses. *Exp4* is a simple extension of the Normal density, and symbolic computation is useful in generating the required density, and integrating it to derive its distribution. If the domain is restricted to positive real numbers only, *Exp4* satisfies the "low loss = high probability; high loss = low probability" criterion stated in section 3.4.

The numerical results are encouraging. They are consistent with expectations based on the previous results which did not involve aggregated losses. Furthermore they are consistent with projected conduct risk losses for the following year.

Using the proposed *ExpN* distribution ($n > 4$) is not recommended. It must be noted that a density with an $exp(-x^n)$ term when n is large (and in this context $n > 4$ is large) appears to be very contrived! The particular case $exp(-x^{100})$ illustrates the point well. It represents extreme overfitting. It is better, in principle, to use a simpler distribution and accept a higher capital value as a more prudent guard against future losses. If the Regulator thinks that capital is too low due to a contrived calculation method, an 'add-on' can be imposed, and the effort in finding a distribution that results in the lowest possible capital is wasted. Therefore the *Exp4* distribution suffices; it works well enough.

REFERENCES

[1] European Banking Authority. Regulation and Policy: Operational Risk. https://www.eba.europa.eu/regulation-and-policy/operational-risk. 2017

[2] The Basel Committee on Banking Supervision. Supervisory Guidelines for the Advanced Measurement Approaches. http://www.bis.org/publ/bcbs196.pdf, June 2011

[3] Risk.net. Conduct Risk: a useful way to carve through chaos, http://www.risk.net/operational-risk-and-regulation/opinion/2446853/conduct-risk-a-useful-way-to-carve-through-chaos, 2015

[4] Thomson Reuters, Conduct Risk Report 2014–5, https://risk.thomsonreuters.com/, 2015

[5] Thomson Reuters “Conduct Risk Report 2013”, https://risk.thomsonreuters.com/, 2013
[6] Financial Conduct Authority. Payment protection insurance (PPI) explained. https://www.fca.org.uk/consumers/payment-protection-insurance-ppi-explained, 2016
[7] The Guardian. Bill for PPI mis-selling scandal tops £40bn. https://www.theguardian.com/money/2016/oct/27/ppi-mis-selling-scandal-bill-tops-40bn-pounds. 27 October 2016
[8] Mitic, P. The problems with conduct risk loss aggregation, Risk.net, http://www.risk.net/risk-management/3848486/the-problems-with-conduct-risk-loss-aggregation, 2 Feb 2017
[9] Frachot, A., Georges,P. and Roncalli, T. Loss Distribution Approach for operational risk. Working paper, Groupe de Recherche Operationnelle, Credit Lyonnais, France. URL: http://ssrn.com/abstract=1032523, 2001
[10] Shevchenko,P.V. Implementing Loss Distribution Approach for Operational Risk, *Applied Stochastic Models in Business and Industry*, 26(3), pp: 277-307, 2010.
[11] Goldie,C.M. and Klüppelberg,C. Subexponential distributions. *A practical guide to heavy tails (eds. Adler,R., Feldman,R.E. and Taqqa,M.S.),* pp.435-459, Birkhauser Boston MA, USA, 1998
[12] Bocker, K. and Kluppelberg, C. Operational VaR: a Closed-Form Approximation *Risk*, pp. 90-93, Dec 2005.
[13] Weisstein,E.W. Incomplete Gamma Function. MathWorld, http://mathworld.wolfram.com/IncompleteGammaFunction.html, 2016
[14] Mitic, P. Improved Goodness-of-fit measures, *Jnl. Operational Risk*, 10(1), 2015
[15] Mitic, P. Chapter 9 in “Risk Appetite Setting and Modelling Conduct Risk”, in “Conduct Risk – the Definitive Guide”, Ed. Peter Haines, Incisive Media. 2016

APPENDIX A

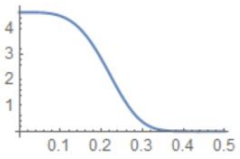
Mathematica scripts to calculate the *Exp4* density

```
f1[x_, s_] := E^(-(x^4/(2*s^4)))
k1 = Integrate[f1[x, s], {x, 0, Infinity}, Assumptions -> {Re[s^4] > 0}];
k = FullSimplify[k1, Assumptions→{Im[s]==0, Re[s]>0}]
```

$$2^{1/4}\, s\, \text{Gamma}\left[\frac{5}{4}\right]$$

```
f[x_, s_] := f1[x, s] / (2^(1/4) s Gamma[5/4])
```

```
Plot[f[x, 0.2], {x, 0, 0.5}]
```



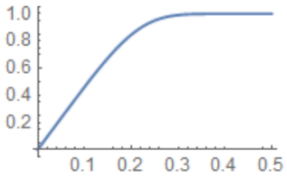
Mathematica scripts to calculate the *Exp4* distribution

```
F[x_, s_] := FullSimplify[Integrate[f[t, s] , {t, 0, x}] ,
    Assumptions → {Im[s] == 0, Re[s] > 0, Im[x] == 0, Re[x] > 0}]
```

```
F[x, s]
```

$$1 - \frac{\text{Gamma}\left[\frac{1}{4}, \frac{x^4}{2\,s^4}\right]}{\text{Gamma}\left[\frac{1}{4}\right]}$$

```
Plot[F[x, 0.2], {x, 0, 0.5}]
```



APPENDIX B

R functions to implement the *Exp4* probability distribution.

Functions *dNormExp4()* and *pNormExp4()* correspond to Equations (2) and (3) respectively. Function *qNormExp4()* is the inverse of *pNormExp4()* and uses a search method to do the inversion. This makes it slow, and a direct method would be preferable. Function *rNormExp4()* generates *Exp4*-distributed random numbers by inverting the distribution function.

```
dNormExp4 <- function(x,s)
{ (1/(s*2^0.25*gamma(5/4)))*exp(-0.5*(x/s)^4)}

pNormExp4 <- function(x,s)
{
  p <- 1 - (1/(gamma(1/4)))*gsl::gamma_inc(1/4, x^4/(2*s^4))
  return(p)
}

qNormExp4 = function(p, s, eps = 1e-10)
{
  lim <- 20*s
  x <- rootSolve::uniroot.all(function(z) { pNormExp4(z, s) - p }, interval = c(eps, lim - eps) )
  if (length(x)>1)  {return(x[1])}
  else  {return(x)}
}

rNormExp4 = function(n, s)
{
  nums <- runif(n)
  rn <- unlist(lapply(1:n, function(z_) {qNormExp4(nums[z_], s)}))
  return(rn)
}
```

APPENDIX C

Formal hypothesis test for use with the *TN-A* test.

Let $X = \{ x_1, x_2, \ldots, x_n\}$be a sample of size n drawn from a random variable V that has some probability distribution Δ (it is implied that this distribution has well-defined density and distribution functions f and F respectively. Formulate null and alternative hypotheses for a 2-tailed test as follows.

Null hypothesis (H_0): $V \sim \Delta$ (i.e. V has a Δ-distribution)
Alternative hypothesis (H_1): $V \not\sim \Delta$ (i.e. V has any other distribution)

Given a calculated value, t, of the *TN-A* statistic and a p% critical value, t_p, of the *TN-A* statistic:

Reject H_0 at p% if $t > t_p$
Reject H_1 at p% if $t \leq t_p$.

Some useful 2-tail *TN-A* critical values are: 0.014 (1%), 0.068 (5%) and 0.131 (10%). Others may be found by inverting equation (5).

SYMCOMP 2017
Guimarães, 6-7 April 2017

STABILIZATION OF A STEADY STATE VISCOELASTIC COMPUTATIONAL CODE USING THE FINITE VOLUME METHOD

Célio Fernandes[1]*, Silvino Araújo[2] , Luís Ferrás[1] and João Nóbrega[1]

1: Institute for Polymers and Composites/i3N, University of Minho
2: Instituto de Ciências Exatas e Naturais, Faculdade de Matemática, Universidade Federal do Pará, Brazil

e-mails: cbpf@dep.uminho.pt ; silvino@ufpa.br ; luis.ferras@dep.uminho.pt ; mnobrega@dep.uminho.pt

Keywords: OpenFOAM, Upper-convected Maxwell model, Sudden contraction flow, Flow around a cylinder

Abstract *In this work we propose a methodology to improve the stability of the viscoelasticFluidFoam solver available in OpenFOAM®, aiming to increase the numerical stability when dealing with viscoelastic fluid flows.*

The improvements are based on the modification of the discrete elastic-viscous stress splitting formulation that avoid the decoupling between velocity and stress fields. The developments are assessed with two benchmark 2D case studies of an upper-convected Maxwell (UCM) fluid, namely the 4:1 planar contraction flow (41PC) and the flow around a confined cylinder (FACC), through the comparison with results available in the literature. In both cases, the simulations were performed at Re = 0.01, which correspond to creeping flow conditions, and Deborah number in range [0, 5] and [0, 0.8], respectively for the 41PC and FACC case studies.

The results obtained in both test cases were accurately predicted in the sense that the vortex length size (41PC) and drag coefficient (FACC) were predicted with an accuracy of less than 0.6% and 0.08%, respectively, obtained by comparing the finest mesh result and the extrapolated value. In both case studies the method preserved the second order accuracy.

SYMCOMP 2017
Guimarães, 6-7 April 2017

HOW NON-INTEGER ORDER DERIVATIVES CAN BE USEFUL TO RHEOLOGY

L.L. Ferrás[1,2*], Neville J. Ford[2], Maria Luísa Morgado[3], Magda Rebelo[4], Gareth H. McKinley[5] and João M. Nóbrega[1]

1: Institute for Polymers and Composites/i3N
University of Minho
Campus de Azurém 4800-058 Guimarães, Portugal
e-mail: luis.ferras@dep.uminho.pt

2: Department of Mathematics
University of Chester
CH1 4BJ, UK

3: Centro de Matemática, Polo CMAT-UTAD
Departamento de Matemática
Universidade de Trás-os-Montes e Alto Douro, UTAD
Quinta de Prados 5001-801, Vila Real, Portugal

4: Centro de Matemática e Aplicações (CMA) and Departamento de Matemática
Faculdade de Ciências e Tecnologia
Universidade NOVA de Lisboa
Quinta da Torre, 2829-516 Caparica, Portugal

5: Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
e-mail: {njford@chester.ac.uk, luisam@utad.pt, msjr@fct.unl.pt, gareth@mit.edu, mnobrega@dep.uminho.pt}

Keywords: Fractional Viscoelastic Models, Caputo Fractional Derivative, Numerical Solution, Graded Meshes, Relaxation Tests, Rheology

Abstract. *In this work we extend a numerical method developed by the group for the solution of fractional differential equations governing the flow of complex fluids. The method is robust and can now deal with graded meshes in time. The grading can be performed in a semi-automatic way, taking into account the evolution in time of the gradient of stress. We also explore the ability of fractional viscoleastic models to fit rheological data obtained from small-amplitude oscillatory shear experiments with blood.*

1 INTRODUCTION

Viscoelastic materials are abundant in nature and present in our daily lives. Examples are paints, blood, polymers, biomaterials, muscle-tendon units, etc. Therefore, it is important to understand this viscoleastic behavior.

In 1867 James Clerk Maxwell while studying the theory of gases proposed a viscoelastic model, that was latter adapted to viscoelastic fluids. The model is simple, consisting of a spring (elasticity) and a dashpot (viscosity) in series, as shown in Fig. 1.

The differential equation describing such behavior for a linear viscoelastic fluid is given by $\boldsymbol{\sigma}+\tau\frac{d\boldsymbol{\sigma}}{dt}=\eta\dot{\boldsymbol{\gamma}}$ [1], with $\boldsymbol{\sigma}$ the stress tensor, $\dot{\boldsymbol{\gamma}}=(\nabla\boldsymbol{u}+(\nabla\boldsymbol{u})^T)$ the rate of deformation tensor, $\mathbf{u}$ the velocity vector, τ the relaxation time of the fluid and η the zero shear rate viscosity.

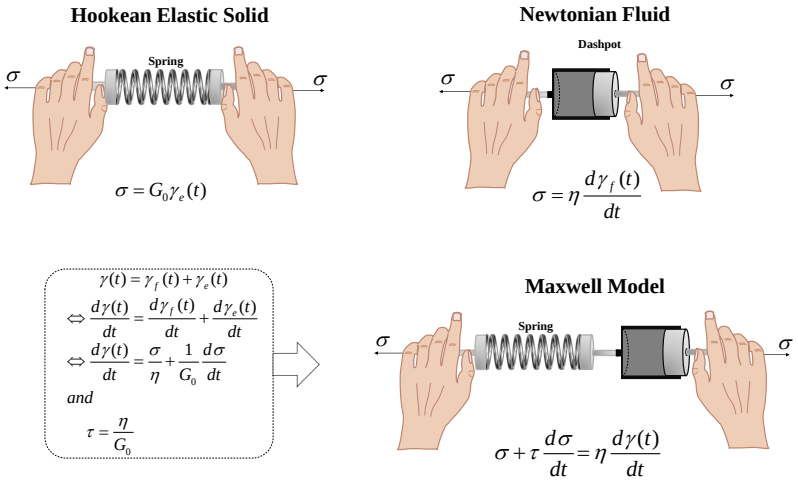


Figure 1: Schematic of a Hookean elastic solid, a Newtonian fluid and the Maxwell model. The hands are stretching the spring, the dashpot and a combination of both.

From figure 1 we see that for Newtonian fluids the stress, σ, is proportional to the rate at which the fluid deforms $\frac{d\gamma_f(t)}{dt}$, while for elastic solids, we follow Hooke's law (the force needed to extend or compress a spring by some distance is proportional to that distance), meaning that the stress, σ, is proportional to the deformation of the spring, γ_e. The combination of the spring and the dashpot leads to the Maxwell model (we assume that the total deformation is given by the sum of the deformations obtained from the spring and the dashpot, and that the stress is the same in both elements).

The Maxwell model can alternatively be written in integral form as:

$$\boldsymbol{\sigma}(t)=\int_0^t G_0 e^{-\frac{t-t'}{\tau}}\frac{d\boldsymbol{\gamma}}{dt'}dt' \tag{1}$$

where $G(t) = G_0 e^{\frac{-t}{\tau}}$ is the relaxation modulus (the response of the stress to a jump in deformation), $\boldsymbol{\gamma}$ is the deformation tensor, and it is assumed that the fluid is at rest for $t < 0$ (if we differentiate Eq. 1 in order to time, the differential model is easily obtained). A closer look into Eq. 1 allows us to interpret the integral as follows: the stress at time t, $\sigma(t)$, is given by the sum of the infinitesimal weighted $\dot{\gamma}(t')\, dt'$ with $t' \leq t$ (the weights are given by the negative exponential function). Note that as t' approaches the current time, t, the weights become bigger and bigger, because recent deformations are expected to influence more the present time, when compared to past deformations.

One way to experimentally characterize viscoelastic fluids is to perform a relaxation test (step-strain), that is, we impose a step displacement and we measure the response of the stress to this deformation. This experiment is illustrated in figure 2 for a viscoelastic fluid. At the instant $t = a$ we impose a constant deformation, and we observe that the stress relaxes until it becomes zero (for fluids). Note the reduction in magnitude (see inset with the schematic of the spring-dashpot relaxation) of the stress shown in figure 2. For $t = t_2$ we see that the spring is almost relaxed, and that the dashpot is more stretched (a response to the jump in deformation).

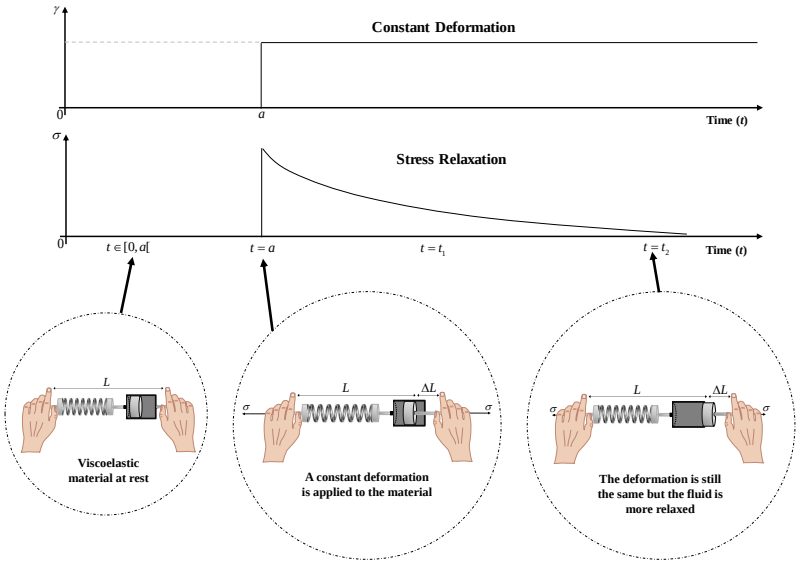


Figure 2: Stress relaxation of a viscoelastic material following a Maxwell model after a step displacement.

The result shown in figure 2 can be easily verified for the Maxwell model. A step in deformation given by $\gamma = \gamma_0 H(t-a)$ with $H(t)$ the Heaviside function, leads to the folowing expression for the stress relaxation,

$$\sigma(t) = G_0 \gamma_0 e^{-\frac{t-a}{\tau}}. \tag{2}$$

with $\tau = \eta/G_0$. As expected, this type of relaxation does not describe all materials, and therefore, more relaxation functions should be explored besides the negative exponential one.

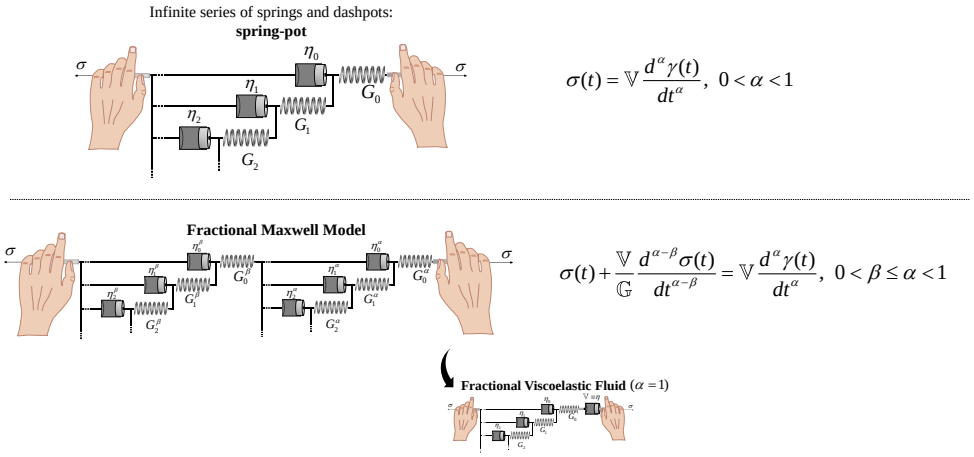


Figure 3: (top) Mechanical equivalent of a springpot, an infinite combination of springs and dashpots. (bottom) Mechanical equivalent of the FMM, a combination of two springpots with parameters $(\mathbb{V}, \alpha)$ and $(\mathbb{G}, \beta)$, respectively. For $\alpha = 1$ we obtain the Fractional Viscoelastic Fluid (FVF).

For example, several classes of complex fluids (biopolymers, bread dough, etc) show a different type of fading memory [3, 4], in which the relaxation modulus is given by,

$$G(t) = St^{-\alpha}, \tag{3}$$

where S is the gel strength and $0 < \alpha < 1$ [4, 5, 6].

For such materials, if we write the relaxation modulus, $G(t)$, in the form $G(t-t') = \frac{\mathbb{V}}{\Gamma(1-\alpha)}(t-t')^{-\alpha}$ then Eq. (1) can be written as (the meaning of $\mathbb{V}$ will be explained later),

$$\boldsymbol{\sigma}(t) = \mathbb{V}\frac{d^\alpha \boldsymbol{\gamma}(t)}{dt^\alpha} = \frac{1}{\Gamma(1-\alpha)}\int_0^t \mathbb{V}(t-t')^{-\alpha}\frac{d\boldsymbol{\gamma}}{dt'}dt', \tag{4}$$

where $\frac{d^\alpha \boldsymbol{\gamma}(t)}{dt^\alpha}$ represents the Caputo fractional derivative [2] of order $0 < \alpha < 1$. This derivative generalises the classical derivatives to non-integer order.

A connection between the relaxation modulus given in Eq. 3 and Eq. 4 is obtained by setting $S = \frac{\mathbb{V}}{\Gamma(1-\alpha)}$. This way we have a fractional derivative operator in our viscoelastic

model, and, we can take advantage of the vast literature on fractional calculus to better understand this model.
The parameter $\mathbb{V}$ (with dimensions $[Pa.s^{\alpha}]$) is a constant for a fixed α and is known as a quasiproperty, a numerical measure of a dynamical process [7, 8].

The first equality in Eq. 4 is known as a springpot [9] and, if for $\alpha = 1$ we have $\mathbb{V} = \eta$, and, for $\alpha = 0$ we have $\mathbb{V} = G_0$, we obtain a general constitutive relationship to represent a viscous fluid ($\alpha = 1$), an elastic solid ($\alpha = 0$) and a mix of both states ($0 < \alpha < 1$).

For a mechanical interpretation in terms of springs and dashpots see figure 3 top. This interpretation (infinite combination of springs and dashpots) was derived by Schiessel and Blumen [10].

The Maxwell model makes use of two elements, a spring and a dashpot, and therefore, the fractional version of the Maxwell model (Fractional Maxwell Model - FMM) also uses two elements (two springpots $\boldsymbol{\sigma}(t) = \mathbb{V}\frac{d^{\alpha}\boldsymbol{\gamma}(t)}{dt^{\alpha}}$ and $\boldsymbol{\sigma}(t) = \mathbb{G}\frac{d^{\beta}\boldsymbol{\gamma}(t)}{dt^{\beta}}$ with $0 < \beta < \alpha < 1$) combined in series as shown in figure 3 bottom [11, 12]. The model is given by:

$$\boldsymbol{\sigma}(t) + \frac{\mathbb{V}}{\mathbb{G}}\frac{d^{\alpha-\beta}\boldsymbol{\sigma}(t)}{dt^{\alpha-\beta}} = \mathbb{V}\frac{d^{\alpha}\boldsymbol{\gamma}(t)}{dt^{\alpha}}, \tag{5}$$

with $0 < \beta < \alpha < 1$. The FMM shows unbounded stress growth following start-up of steady shear at $\dot{\gamma} = \dot{\gamma}_0 H(t)$ when $\alpha \neq 1$, that is, $\eta^{+}(t) = \lim_{t\to\infty} \sigma(t)/\dot{\gamma}_0$ diverges as $t^{1-\alpha}$. Therefore, it is not suitable to describe the steady flow of viscoelastic fluids. On the other hand, if we assume $\alpha = 1$ (combination of a springpot and a dashpot in series as shown in figure 3 bottom), the stress growth becomes bounded, which makes the model suitable for describing the transient response of fluids. From now on, the FMM with $\alpha = 1$ will therefore be referred to as the Fractional Viscoelastic Fluid model (FVF).

Is is interesting to see that if we consider a relaxation experiment such as the one shown in figure 2 (a step in deformation given by $\gamma = \gamma_0 H(t)$) the stress obtained from the FMM shows a relaxation behavior that can be represented by the Mittag-Leffler function, a generalisation of the exponential function [13]. The analytical expression [11, 12] is given by:

$$\sigma(t) = \mathbb{G}\gamma_0 t^{-\beta} E_{\alpha-\beta,1-\beta}\left(-\frac{\mathbb{G}}{\mathbb{V}}t^{\alpha-\beta}\right). \tag{6}$$

Note that for $a = b = 1$ we have that $E_{a,b}(z) = e^{z}$. By looking at Eq. 6 we see that to obtain an exponential function we must have $\alpha - \beta = 1$ and $1 - \beta = 1$, that is, $\alpha = 1$ and $\beta = 0$. This leads to the relaxation equation obtained for the Maxwell model, Eq. 2, with a relaxation time of $\mathbb{V}/\mathbb{G}$, as expected. More particular cases of the FMM are described in detail in [14].

The characteristic relaxation time is defined as the the inverse of the frequency at which the storage (G') and loss modulus (G'') intersect. For the FMM we do not always obtain an intersection between G' and G'' (this is only possible when $0 \leq \beta < 0.5 < \alpha \leq 1$ [7, 14]), and therefore the characteristic relaxation time is defined as $\tau = (\mathbb{V}/\mathbb{G})^{1/(\alpha-\beta)}$ [7].

Since this model generalizes the Maxwell model, and based on the fact that the multi-mode Maxwell model (combination of Maxwell elements in parallel) usually provides a better fit to rheological data, we are expecting to obtain better fits to rheological data with the FMM when compared with the Maxwell model (more on this will be presented in the next sections).

The objective of this work is to develop a numerical method for the solution of time-dependent unidirectional flows, where the material under study can be characterized by a FMM. The numerical method is based on a previous work by the group [14], that is generalized and explained in detail.

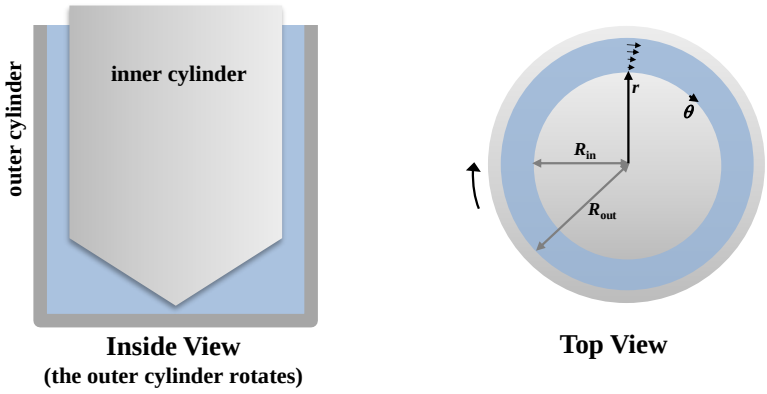


Figure 4: Schematic of the geometry under study. The complex material is contained between two cylinders

The numerical code will be tested in a step-strain test, similar to the one shown in figure 2. The experimental setup is given by two cylinders forming an annular geometry, as shown in figure 4. The material to be studied is initially at rest and contained between the two cylinders. Suddenly the outer cylinder rotates for a period of time and then stops after turning through an angle $\Delta\theta$, and we measure how the stress relaxes. The deformation for this case is defined as $\gamma = R_{out}\Delta\theta/(R_{out} - R_{in})$. The numerical method presented in the next Section will mimic this experimental procedure.

It should be remarked that the FMM is not frame-invariant and therefore it can only be used for linear viscoelastic flows. An extension of this model able to deal with non-linear

flows is proposed in [7], and a numerical method for these non-linear flows is proposed in [14].

The work is organised as follows: The numerical method is presented in Section 2. In Section 3 we present the numerical results obtained for the step-strain test, and the paper ends with the conclusions in Section 4.

2 NUMERICAL METHOD

The equations governing the 1D transient flow shown schematically in figure 4 (right) are the momentum equation,

$$\rho \frac{\partial u_\theta(r,t)}{\partial t} = \eta \left(\frac{2}{r} + \frac{\partial}{\partial r} \right) \sigma_{r\theta}(r,t), \tag{7}$$

and the shear component of the constitutive equation (the FMM) that is given by,

$$\left(1 + \frac{\mathbb{V}}{\mathbb{G}} \frac{\partial^{\alpha-\beta}}{\partial t^{\alpha-\beta}} \right) \sigma_{r\theta}(r,t) = \mathbb{V} \frac{\partial^{\alpha} \gamma(r,t)}{\partial t^{\alpha}}, \tag{8}$$

or in integral form,

$$\sigma_{r\theta}(r,t) + \frac{\mathbb{V}}{\mathbb{G}} \frac{1}{\Gamma(1-(\alpha-\beta))} \int_0^t (t-t')^{-(\alpha-\beta)} \frac{\partial \sigma_{r\theta}(r,t')}{\partial t'} dt' = \mathbb{V} \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-t')^{-\alpha} \frac{\partial \gamma(r,t')}{\partial t'} dt', \tag{9}$$

where u_θ is the tangential velocity, and $\sigma_{r\theta}$ is the shear stress [14].

For this particular flow, the rate of deformation, $\dot{\gamma}(r,t')$, is given by $\left(\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r} \right)$ for each t' and r. When $\alpha \neq 1$ we need to consider all the past deformations, $\gamma(r,t')$, by computing the integral on the right-hand-side of Eq. 9. If $\alpha = 1$ we obtain a classical derivative of the deformation at the current time, and there is no need to consider all the previous deformation history.

If we apply the operation $\left(\frac{2}{r} + \frac{\partial}{\partial r} \right)$ to Eq. 8 and $\left(1 + \frac{\mathbb{V}}{\mathbb{G}} \frac{\partial^{\alpha-\beta}}{\partial t^{\alpha-\beta}} \right)$ to Eq. 7, we obtain the following system of integro-differential equations:

$$\rho \left(1 + \frac{\mathbb{V}}{\mathbb{G}} \frac{\partial^{\alpha-\beta}}{\partial t^{\alpha-\beta}} \right) \frac{\partial u_\theta(r,t)}{\partial t} = \frac{\mathbb{V}}{\Gamma(1-\alpha)} \int_0^t (t-t')^{-\alpha} \left(\frac{\partial^2 u_\theta(r,t')}{\partial r^2} + \frac{1}{r} \frac{\partial u_\theta(r,t')}{\partial r} - \frac{u_\theta(r,t')}{r} \right) dt', \tag{10}$$

$$\left(1 + \frac{\mathbb{V}}{\mathbb{G}} \frac{\partial^{\alpha-\beta}}{\partial t^{\alpha-\beta}} \right) \sigma_{r\theta}(r,t) = \frac{\mathbb{V}}{\Gamma(1-\alpha)} \int_0^t (t-t')^{-\alpha} \left(\frac{\partial u_\theta(r,t')}{\partial r} - \frac{u_\theta(r,t')}{r} \right) dt'. \tag{11}$$

For the boundary conditions we assume the outer cylinder starts rotating at $t = 0$ with a tangential velocity given by, $u_\theta(R_{out}, t) = \dot{\theta}(t) R_{out}$, where

$$\dot{\theta}(t) = \frac{\Delta\theta}{\psi\sqrt{\pi}} e^{-\frac{(t-t_d)^2}{\psi^2}}, \tag{12}$$

while the inner cylinder is fixed ($u_\theta(R_{in}, t) = 0$). This way we mimic the experiment shown in figure 2, by imposing a step deformation (through a finite rotation $\Delta\theta = \int \dot{\theta} dt$ of the outer cylinder).

Note that as $\psi \to 0$, $u_\theta(R_{out}, t)$ converges to the Dirac delta function multiplied by the angle turned, $\Delta\theta\delta(t)$ (assuming $t_d = 0$). The need for the delay time, t_d, comes from the initial condition $\frac{du_\theta(R_{out},0)}{dt} = 0$ imposed by the governing equations and the experiment itself.

This system of equations (Eqs. 10 and 11) provides a 1-way coupling between stress and velocity, meaning that we can solve for velocity first, and then calculate the shear stress with the previously calculated velocity field.

In order to solve this system of equations numerically, we need to obtain an approximation for all the operators (time and spatial derivatives). For that, we consider a uniform space mesh on the interval $[R_{in}, R_{out}]$ defined by the gridpoints $r_i = R_{in} + i\Delta r$, $i = 0, \ldots, N$, where $\Delta r = \frac{R_{out}-R_{in}}{N}$. For the discretisation in time we consider graded meshes with time gridpoints $t_s = \sum_{k=1}^{s} \Delta t[k]$, $s = 0, 1, ..., S$, where $\Delta t[k] = t_k - t_{k-1}$ is the adaptive time step. The numerical method that will be used to solve Eq. 10 is inspired on the method presented by Sun and Wu [15] for the fractional diffusion-wave equation.

The fractional differential equation governing the evolution of velocity in time and space (Eq. 10) can be re-written as:

$$\frac{\rho}{\mathbb{V}} \frac{\partial u_\theta(r,t)}{\partial t} + \frac{\rho}{\mathbb{G}} D_t^{1+\alpha-\beta} u_\theta(r,t) = D_t^\alpha \left(\frac{\partial^2 u_\theta(r,t)}{\partial r^2} + \frac{1}{r}\frac{\partial u_\theta(r,t)}{\partial r} - \frac{u_\theta(r,t)}{r^2} \right), \tag{13}$$

where we have written the fractional derivative in a more compact way $D_t^\alpha f$.
Denoting by u_i^s an approximation of $u_\theta(r_i, t_s)$, $i = 1, ..., N-1$, $\quad s = 1, ..., S$ with $\Delta t[s] = t_s - t_{s-1}$ we can define the following:

$$u_i^{s-1/2} \equiv \frac{1}{2}\left(u_i^s + u_i^{s-1}\right), \qquad \delta_t u_i^{s-1/2} \equiv \frac{1}{\Delta t[s]}\left(u_i^s - u_i^{s-1}\right) \tag{14}$$

$$\delta_r u_{i-1/2}^s \equiv \frac{1}{\Delta r}\left(u_i^s - u_{i-1}^s\right), \qquad \delta_r^2 u_i^s \equiv \frac{1}{\Delta r}\left(\delta_r u_{i+1/2}^s - \delta_r u_{i-1/2}^s\right) \tag{15}$$

Each term in Eq. 13 is then substituted by its respective finite difference approximation at the mesh point $(r, t) = (r_i, t_{s-1/2})$, that is, $\frac{\partial u_\theta(r,t)}{\partial t} \approx \delta_t u_i^{s-1/2}$, $\frac{\partial^2 u_\theta(r,t)}{\partial r^2} \approx \delta_r^2 u_i^{s-1/2}$, $\frac{1}{r}\frac{\partial u_\theta(r,t)}{\partial r} \approx$

$\frac{1}{4r_i}\left(\delta_r u_i^s + \delta_r u_i^{s-1}\right)$, $\frac{u_\theta(r,t)}{r^2} \approx \frac{u_i^{s-1/2}}{r_i^2}$. Note that these approximations come from an average in time (a well known procedure for increasing the accuracy of the method [16]).

The fractional derivative, D_t^ϵ, with $\epsilon = 1 + \alpha - \beta$, is approximated by:

$$\begin{aligned}
&\frac{1}{\Gamma(2-\epsilon)} \int_0^{t_s} \left(t_s - t'\right)^{1-\epsilon} \frac{d^2 u_\theta(t')}{dt'^2} dt' \approx \\
&\quad \approx \frac{1}{\Gamma(1-(\alpha-\beta))} \left[\frac{a_{s,s}}{\Delta t[s]} \left(\delta_t u_i^{s-1/2} - \delta_t u_i^{(s-1)-1/2} \right) \right] \\
&\quad - \frac{1}{\Gamma(1-(\alpha-\beta))} \left[\sum_{j=0}^{s-1} \left(\frac{a_{s,j+1}}{\Delta t[j+1]} \left(\delta_t u_i^{(j+1)-1/2} - \delta_t u_i^{(j)-1/2} \right) \right) \right]
\end{aligned} \tag{16}$$

with

$$a_{s,k} = \frac{1}{2-\epsilon} \left[(t_s - t_{k-1})^{2-\epsilon} - (t_s - t_k)^{2-\epsilon} \right]. \tag{17}$$

This approximation is of order $O(\Delta t^{3-\epsilon})$ when using a uniform mesh in time [15]. The case of graded meshes is being studied by the group.

The discretized momentum equation is then given by

$$\begin{aligned}
&\frac{\rho}{\mathbb{V}} \delta_t u_i^{s-1/2} + \frac{\rho}{\Gamma(1-(\alpha-\beta))\,\mathbb{G}} \left[\frac{a_{s,s}}{\Delta t[s]} \left(\delta_t u_i^{s-1/2} - \delta_t u_i^{(s-1)-1/2} \right) \right] \\
&\quad - \frac{\rho}{\Gamma(1-(\alpha-\beta))\,\mathbb{G}} \left[\sum_{j=0}^{s-1} \left(\frac{a_{s,j+1}}{\Delta t[j+1]} \left(\delta_t u_i^{(j+1)-1/2} - \delta_t u_i^{(j)-1/2} \right) \right) \right] \\
&\quad - \left\{ \frac{\Delta t[s]^{1-\alpha}}{\Gamma(2-\alpha)} \right\} \left(\delta_r^2 u_i^{s-1/2} + \frac{1}{2r_i} \left(\frac{\delta_r u_i^s + \delta_r u_i^{s-1}}{2} \right) - \frac{u_i^{s-1/2}}{r_i^2} \right) \\
&\quad = \frac{1}{\Gamma(2-\alpha)} \sum_{j=0}^{s-2} \left[\left(\delta_r^2 u_i^{(j+1)-1/2} + \frac{1}{2r_i} \left(\frac{\delta_r u_i^{(j+1)} + \delta_r u_i^{(j+1)-1}}{2} \right) - \frac{u_i^{(j+1)-1/2}}{r_i^2} \right) d_{sj} \right]
\end{aligned} \tag{18}$$

for $i = 1, ..., N-1$ and $s = 1, ..., S$, with $d_{sj} = \left[(t_s - t_j)^{1-\alpha} - (t_s - t_{j+1})^{1-\alpha} \right]$. At the boundaries we have $u_0^s = \phi_0(t_s)$ and $u_N^s = \phi_N(t_s)$, with $\phi_i(t_s)$ a general function of time. This will generate (at each time step) a linear system of $(N-1)\times(N-1)$ algebraic equations for $(N-1)\times(N-1)$ unknowns.

The numerical method used to solve the discretized stress equation is similar to the one used for the momentum equation. The main difference is that the order of the fractional derivative on stress is $\alpha - \beta$, and therefore the approximation obtained for the velocity fractional derivative (Eq. 16) needs to be slightly adapted by considering $\alpha - \beta$ in the place of $1 + \alpha - \beta$. The discretized stress equation is then given by:

$$\sigma_i^s + \frac{\mathbb{V}}{\Gamma(1-(\alpha-\beta))\,\mathbb{G}}\frac{1}{\Delta t}\left[b_{l,l}\left(\frac{\sigma_i^s-\sigma_i^{s-1}}{\Delta t[s]}\right) - \sum_{j=0}^{s-1} b_{l,j+1}\left(\frac{\sigma_i^{j+1}-\sigma_i^j}{\Delta t[j+1]}\right)\right] =$$
$$= \left\{\frac{1}{\Gamma(2-\alpha)}\sum_{j=0}^{s-2}\left[\left(\frac{u_{i+1}^{j+1}-u_{i-1}^{j+1}}{2\Delta r} - \frac{u_i^{j+1}}{r_i}\right)d_{sj}\right]\right\} \tag{19}$$
$$+\frac{\Delta t[s]^{1-\alpha}}{\Gamma(2-\alpha)}\mathbb{V}\left(\frac{u_{i+1}^s-u_{i-1}^s}{2\Delta r} - \frac{u_i^s}{r_i}\right)$$

for $i = 1, ..., N-1$ and $s = 1, ..., S$ with,

$$b_{s,k} = \frac{1}{1-\Theta}\left[(t_s - t_{k-1})^{1-\Theta} - (t_s - t_k)^{1-\Theta}\right], \tag{20}$$

where $\Theta = \alpha - \beta$.

2.1 Time Graded Meshes

For Newtonian fluids the relaxation is instantaneous, but for viscoelastic fluids we can have characteristic relaxation times that are several orders of magnitude higher than the time of step (a Hookean solid has an infinite relaxation time). Therefore, it is extremely important to use graded meshes in time.

In this work we propose two approaches. One that is based on the behavior of the shear stress, where we propose a certain function to distribute the time-mesh points, and an automatic approach based on the local gradients (in time) of the stress, where the time-mesh points are created as we evolve in time.

For the *first approach* we performed several numerical experiments and we observed that a good mesh grading could be given by:

$$t_s = \begin{cases} 2t_d - 2t_d\left(1-\frac{s}{N_1}\right)^{r_1} & s < N_1 \\ 2t_d + T\left(\frac{s-N_1}{N_2}\right)^{r_2} & s \geqslant N_1 \end{cases} \tag{21}$$

where T/t_d is the normalized duration of the experiment, N_1 is the number of grid points to be used in the *first part* of the numerical experiment, N_2 is the number of points in the *second part* of the numerical experiment and r_1, r_2, control the nonlinear expansion/contraction of the mesh in time. We consider the first part of the experiment goes from $t = 0$ until $2t_d$. This way we can capture the higher gradients in a more controlled way, as shown in the schematic of figure 5, where we see a typical evolution of the velocity of the outer cylinder and also the typical evolution of the shear stress.

The *second approach* is based on the local gradient of stress in time. We still divide the domain into two parts, but the process is more automatic. The idea is to have the new time

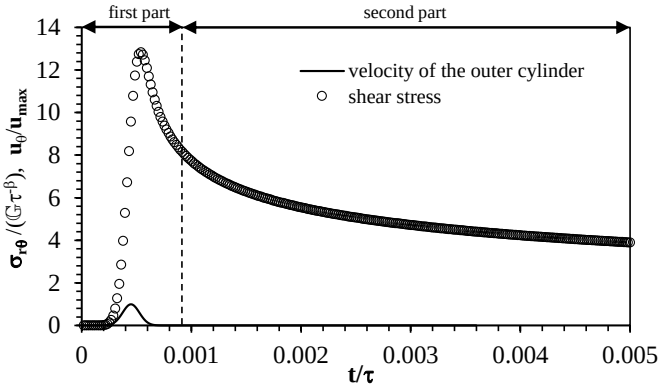


Figure 5: Typical behavior of velocity and shear stress in a shear experiment. Note the first and second parts. The normalized stress relaxation and pulse in velocity were obtained for a deformation of 100%, $\mathbb{V} = 24.96$ Pa.s$^{\alpha}$, $\mathbb{G} = 1.56$ Pa.s$^{\beta}$, $\alpha = 1$, $\beta = 0.31$, $t_d/\tau = 4.5 \times 10^{-4}$, $\psi/\tau = 1.0 \times 10^{-4}$.

step, Δt_{new}, as a function of the variation of the stress, $d\sigma_{r\theta}/dt$, and the minimum and maximum time steps, Δt_{min} and Δt_{max}, provided by the user. Note that Δt_{min} must take into account the time-step used in the *first part*. For this case we propose a grading given by:

$$t_s = \begin{cases} 2t_d - 2t_d \left(1 - \frac{s}{N_1}\right)^{r_1} & s \leq N_1 \\ t_{s-1} + \frac{2\Delta t_{max}}{1+\left(\frac{d\sigma_{r\theta}(t_{s-1})}{dt}+1\right)^{r_2}} & s > N_1 \end{cases} \tag{22}$$

The need for a maximum time step comes from the observation of numerical oscillations at long times in the first grading method proposed.

3 NUMERICAL RESULTS

To perform the numerical simulation of a step-strain we need to obtain first the model parameters for the FMM. In this work we have considered a fit to the storage (G') and loss (G'') moduli data of blood, obtained by passive microrheology, that is based on the diffusivity of tracer particles dispersed in a solution. This reohlogical data is presented in the work of Campo-Deaño et al. [17].

Human blood is a dense suspension of platelets, leucocytes, and mainly erythrocytes in plasma (aqueous polymer solution), as shown in figure 6. Depending on the flow conditions erythrocytes may aggregate or disperse and they may also deform, leading to a non-Newtonian behavior [17]. This non-Newtonian behavior is problematic, especially in small vessels, being closely related to incident cardiovascular events [17]. Therefore it is important to have a good model for the blood behavior, so that more information on these fluids can be obtained or extrapolated.

Note that fractional derivatives allow the modeling of anomalous diffusion, and, passive microrheology is based on the diffusion of particles, therefore, in the future it would be interesting

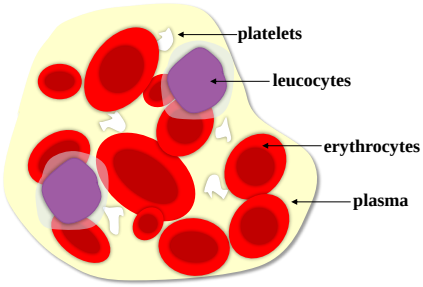


Figure 6: Schematic of a blood sample.

to relate fractional viscoelastic models with the complex modulus obtained from an anomalous diffusion of particles (see [17]).

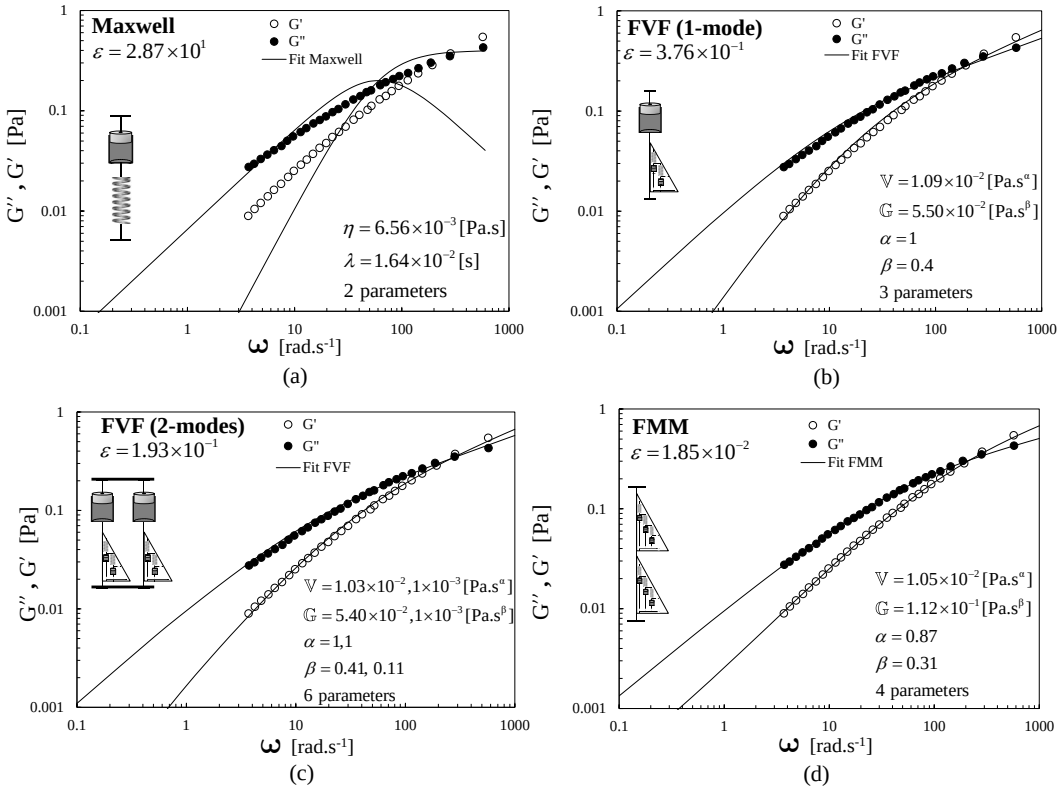


Figure 7: fit to the storage (G') and loss (G'') moduli rheological data of blood. (a) Classical Maxwell model; (b) 1-mode FVF; (c) 2-modes FVF; (d) FMM.

In figure 7 we show the fit to the linear viscoelastic moduli $G'(\omega)$ and $G''(\omega)$, by considering

the classical Maxwell model, the FVF (FMM with $\alpha = 1$) with one and two modes and the FMM. To compare the different fits we created an error function that is given by:

$$\varepsilon = \sum_i \left[\log G'_i - \log G'_{fit}(\omega_i)\right]^2 + \sum_i \left[\log G''_i - \log G''_{fit}(\omega_i)\right]^2. \tag{23}$$

As expected, a better fit is obtained with the fractional models when compared to the classical one. Note that with the 3 parameter FVF we already obtain a good fit. The use of more than one mode for the FVF does not seem to change much the quality of the results. This happens because α dictates the behavior of the G' and G'' for small frequencies, and $\alpha = 1$ is fixed. Therefore, since we could obtain a good fit with one mode, the second mode could not improve much the results. An almost perfect fit is obtained with the 4 parameter FMM. In conclusion we can say that these fractional models are powerful and can be used to model complex materials, at least in the linear regime studied here.

For the numerical simulations we consider a portion of fluid between two cylinders of radii $R_{in} = 22.65\,mm$ and $R_{out} = 25\,mm$, leading to a gap of $2.35\,mm$ [3]. The FMM parameters are the ones given in figure 7(d) with a relaxation time of $\tau = (\mathbb{V}(\mathbb{G})^{1/(\alpha-\beta)} = 1.459 \times 10^{-2}$ seconds. The parameters used in the outer cylinder boundary condition are, $\psi/\tau = 2.06 \times 10^{-1}$, $t_d/\tau = 1.71 \times 10^{0}$, and we impose a deformation of 100%, $\gamma_0 = 1$ (meaning that the distance traveled by the outer cylinder is the same as the gap between the two cylinders). The mesh size used in the simulations is $\Delta r/(R_o - R_i) = 1.61 \times 10^{-3}$ and $\Delta t_{min}/\tau = 6.86 \times 10^{-2}$.

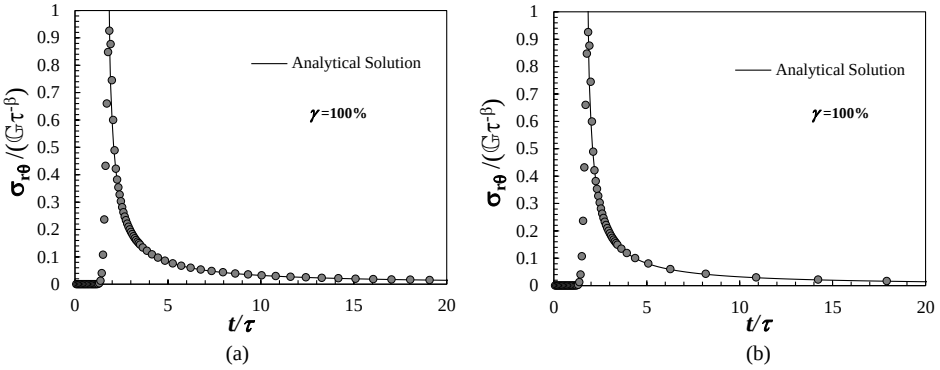


Figure 8: Normalized shear stress relaxation obtained for the step-strain test shown in figure 2 considering the FMM, $\gamma = 100\%$, $\psi/\tau = 2.06 \times 10^{-1}$, $t_d/\tau = 1.71 \times 10^{0}$, $\beta = 0.31$, $\alpha = 0.87$, $\mathbb{V} = 1.05 \times 10^{-2}$, $\mathbb{G} = 1.12 \times 10^{-1}$. In this simulation we have considered $\Delta t_{min}/\tau = 6.86 \times 10^{-2}$, $\Delta r/\tau = 1.61 \times 10^{-3}$. (a) *first approach*; (b) *second approach*.

The numerical simulations were performed with the software Mathematica (V. 11) in a computer with a processor Intel(R) Core(TM)i7-4650U CPU @ 1.70GHz 2.30GHz and 8Gb of RAM.

For the simulation with the *first approach* we have considered $r_1 = 1$, $r_2 = 1.687$, $T/t_d = 40$ ($t_d = 0.025$ seconds), $T/\tau = 6.86 \times 10^{1}$, $N_1 = 50$ and $N_2 = 60$, leading to a maximum time step

of $\Delta t_{max}/\tau = 1.92 \times 10^0$. We tried different r_2 values, but we observed numerical oscillations when $r_2 > 1.7$. More tests need to be performed in the future together with the derivation of the theoretical order of convergence.

To simulate one second of real life relaxation of blood this simulation took 2 minutes and 17 seconds. We also simulated 10 seconds of relaxation and it took approximately 40 minutes ($N_2 = 250$ and $r_2 = 1.668$). This huge increase in simulation time comes from the number of previous deformations that we have to store that are needed at each time step to compute velocity and stress, a consequence of the nonlocal nature of the fractional differential operator.

We also performed numerical simulations with the *second approach* by considering $N_1 = 50$ and $\Delta t_{max}/\tau = 4.11$. For the same 2 minutes and 17 seconds of computation we could run 3 seconds of stress relaxation. No numerical oscillations were observed.

This second method is more automatic than the first one, and provided faster results. Although, we can not say it is always faster because it depends on different parameters. What we can say is that the grading is automatic and evolves with the numerical solution, and, we can control the maximum step size. In the first approach the time step is always increasing.

The results obtained with both methods are shown in figure 8, and both provided a good agreement with the analytical solution, Eq. 6. From the mesh distribution it is obvious in both cases where are the first and second parts of the mesh. For the *first approach* the mesh is gradually increasing, and for the *second approach* the increase in the mesh size is more abrupt, but then evolves to a constant grading.

We may conclude that both methods have their own advantages and disadvantages, and that to obtain a good mesh distribution some numerical experiments need to be performed a priori. In the future we will explore with more detail the range of applicability of both grading methods.

4 CONCLUSIONS

We have explained in detail a generalization of a method previously developed by the group for the numerical simulation of a step strain (time-dependent unidirectional flow), with the fluid contained between two cylinders (the inner cylinder is fixed and the outer cylinder rotates with a pulse velocity profile). The method can solve fast transients (the outer cylinder rotation velocity approaches a Dirac delta function multiplied by the angle turned) and we have implemented a semi-automatic mesh grading in time that allows the numerical solution of a step strain for longer relaxation times with a feasible computation time.

We also explain in detail the relaxation of complex materials and present a fit to the storage and loss moduli rheological data obtained for blood in small amplitude oscillatory shear.

The results obtained in a case study comprising a step-deformation test of blood between concentric cylinders, showed that the time mesh grading scheme gives accurate predictions for the stress evolution.

5 ACKNOWLEDGMENTS

L.L. Ferrás and J.M. Nóbrega would like to thank the financial funding by FEDER through the COMPETE 2020 Programme, the National Funds through FCT - Portuguese Foundation for Science and Technology under the project UID/CTM/50025/2013. L.L. Ferrás would also like to thank the funding by FCT through the scholarship SFRH/BPD/100353/2014. M.L. Morgado

would like to thank the funding by FCT through Project UID/MAT/00013/2013 and M. Rebelo would also like to thank the funding by FCT through Project UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

REFERENCES

[1] Bird R.B., Armstrong R.C., Hassager O., *Dynamics of Polymeric Liquids, Fluid Mechanics, second ed., vol. I*, Wiley, 1987.

[2] Caputo, M., "Linear Models of Dissipation whose Q is almost Frequency Independent-II", *Geophysical Journal International*, Vol. **13**, pp. 529-539, 1967.

[3] Keshavarz, B., Divoux, T., Manneville, S., McKinley, G.H., "Nonlinear viscoelasticity and generalized failure criterion for biopolymer gels", submitted for publication in *Physical Review Letters* 2016.

[4] Ng, T.S.-K., McKinley G.H., Padmanabhan, M., "Linear to non-linear rheology of wheat flour dough", *Appl Rheol*, Vol. **16**, pp. 265-274, 2006.

[5] Chambon, F., Winter H.H., "Stopping of crosslinking reaction in a PDMS polymer at the gel point", *Polymer Bulletin*, Vol. **13**, pp. 499-503, 1985.

[6] Chambon, F., Winter H.H., "Linear viscoelasticity at the gel point of a crosslinking PDMS with imbalanced stoichiometry", *Journal of Rheology*, Vol. **31**, pp. 683-697, 1987.

[7] Jaishankar, A., McKinley, G.H., "Power-law rheology in the bulk and at the interface: quasi-properties and fractional constitutive equations", *Proc. R. Soc. A. The Royal Society*, 2012 DOI: 10.1098/rspa.2012.0284.

[8] Scott-Blair, G.W., "The role of psychophysics in rheology", *J. Colloid Science*, Vol. **2**, pp. 21-32, 1947.

[9] Koeller, R.C., "Applications of fractional calculus to the theory of viscoelasticity", *Journal of Applied Mechanics*, Vol. **51**, pp. 299-307, 1984.

[10] Schiessel, H., Blumen, A., "Hierarchical analogues to fractional relaxation equations", *Journal of Physics A: Mathematical and General*, Vol. **26**, pp. 5057-5069, 1993.

[11] Schiessel, H., Metzler, R., Blumen, A., Nonnenmacher, T.F., "Generalized viscoelastic models: their fractional equations with solutions", *Journal of physics A: Mathematical and General*, Vol. **28**, pp. 6567-6584, 1995.

[12] Friedrich, C., "Relaxation and retardation functions of the Maxwell model with fractional derivatives", *Rheologica Acta*, Vol. **30**, pp. 151-158, 1991.

[13] Podlubny, I., *Fractional differential equations: an introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*, Academic press, 1998.

[14] Ferrás, L.L., Ford, N.J., Morgado, M.L., Rebelo, M., McKinley, G.H., Nóbrega, J.M., "Fractional Viscoelastic Flows: theoretical and numerical studies for simple geometries", submitted to Applied Mathematical Modeling, 2017.

[15] Sun, Z.Z., Wu, X., "A fully discrete difference scheme for a diffusion-wave system", *Applied Numerical Mathematics*, Vol. **56**, pp. 193-209, 2006.

[16] Crank, J., Nicolson, P., "A practical method for numerical evaluation of solutions of partial differential equations of the heat conduction type", *Proc. Camb. Phil. Soc.*, Vol. **43**, pp. 5067, 1947.

[17] Campo-Deaño, L., Dullens, R.P., Aarts, D.G., Pinho, F.T., Oliveira, M.S., "Viscoelasticity of blood and viscoelastic blood analogues for use in polydymethylsiloxane in vitro models of the circulatory system", *Biomicrofluidics*, Vol. **7**, pp. 034102, 2013.

SYMCOMP 2017
Guimarães, 6-7 April 2017

INQUIRING ABOUT PEDIATRIC HYPERTENSION

M. Filomena Teodoro[1,2*], Carla Simão[3,4] and Andreia Romana[3]

1: CEMAT, Center for Computational and Stochastic Mathematics
Instituto Superior Técnico
Lisbon University
Avenida Rovisco Pais, 1, 1048-001 Lisboa, Portugal

2: CINAV, Naval Research Center
Portuguese Naval Academy
Portuguese Navy
Base Naval de Lisboa
Alfeite, 1910-001 Almada, Portugal
e-mail: maria.alves.teodoro@marinha.pt

3: Medicine Faculty
Lisbon University
Av. Professor Egas Moniz, 1600-190 Lisboa

4: Pediatric Department
Santa Maria's Hospital
Centro Hospitalar Lisboa Norte
Av. Professor Egas Moniz, 1600-190 Lisboa
e-mail: carla.mail@netcabo.pt

Keywords: Pediatric hypertension, caregiver, knowledge, generalized linear model, multivariate techniques

Abstract. *Arterial hypertension in pediatric age is an important public health problem, whose prevalence has increased significantly over time. Pediatric arterial hypertension (PAH) is a highly prevalent, silent, and perhaps because of this, under-diagnosed in most cases, with multiple repercussions on the health of children and adults. Health professionals and family members must know the PAH existence, the negative consequences associated with it, the risk factors and it's prevention. In [1] can be found a statistical data analysis using a simpler questionnaire. In [2]such work was completed. The present article presents a statistical approach to an extension of such questionnaire applied to a distinct population and filled online. This work considers a simple questionnaire with 15 questions which was built under the aim of easy and quick answers (Yes/No). Were estimated models by general linear models with enough explication of the data. It was explored the binary outcome issue. We also developed multivariate techniques. The results are promising.*

Acknowledgements: This work was supported by Portuguese funds through the *Center of Naval Research* (CINAV), Naval Academy, Portuguese Navy, Portugal and the *Center for Computational and Stochastic Mathematics* (CEMAT), *The Portuguese Foundation for Science and Technology* (FCT), University of Lisbon, Portugal, project UID/Multi/046-21/2013.

REFERENCES

[1] Teodoro, M. Filomena and Simão, C. , "Perception about Pediatric Hypertension", *Journal of Computational and Applied Mathematics*, Elsevier Vol. **312**, pp. 209-215, 2017.

[2] Teodoro, M. Filomena and Simão, C. , "Completing the Analysis of a Questionnaire About Pediatric Blood Pressure", *Transactions on Biology and Biomedicine*, World Scientific and Engineering Academy and Society Vol. **14**, pp. 56-64, 2017.

Author Index