Aplimat

volume IV

2011

Journal of Applied Mathematics

ISSN: 1337-6365

APLIMAT - JOURNAL OF APPLIED MATHEMATICS

VOLUME 4 (2011), NUMBER 1

APLIMAT - JOURNAL OF APPLIED MATHEMATICS

VOLUME 4 (2011), NUMBER 1

Edited by: Slovak University of Technology in Bratislava

Editorial Office:
Institute of natural sciences, humanities and social sciences
Faculty of Mechanical Engineering
Slovak University of Technology in Bratislava
Námestie slobody 17
812 31 Bratislava

Correspodence concerning subscriptions, claims and distribution:
F.X. spol s.r.o
Dúbravská cesta 9
845 03 Bratislava 45
journal@aplimat.com

Frequency: One volume per year consisting of four issues at price of 120 EUR, per volume, including surface mail shipment abroad.
Registration number EV 2540/08

Information and instructions for authors are available on the address:
http://www.journal.aplimat.com/

Printed by: FX spol s.r.o, Azalková 21, 821 00 Bratislava

APLIMAT - JOURNAL OF APPLIED MATHEMATICS
VOLUME 4 (2011), NUMBER 1

DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS

APLIMAT - JOURNAL OF APPLIED MATHEMATICS

VOLUME 4 (2011), NUMBER 1

MODELING AND SIMULATION

APLIMAT - JOURNAL OF APPLIED MATHEMATICS
VOLUME 4 (2011), NUMBER 1

NEW TRENDS IN MATHEMATICS EDUCATION

LIST OF REVIEWERS

Abderramán Marrero Jesús C., Professor	U.P.M. Madrid Tech. University, Madrid, Spain
Ajevskis Viktors, Dr. Math	European Central Bank,Frankfurt, Germany
Andrade Marina, Professor Auxiliar	IBS - IUL, Lisboa, Portugal
Ansary M. A., PhD	University of Rajshahi, Rajshahi, Bangladesh
Bácsó Sándor, CSc.	University of Debrecen, Debrecen, Hungary
Bartošová Jitka, RNDr., PhD.	University of economics in Prague, Jindřichův Hradec, Czech Republic
Bečvář Jindřich, doc. RNDr., CSc.	Univerzita Karlova, Praha, Czech Republic
Bělašková Silvie, Mgr. Bc.	Tomas Bata University, Zlín, Czech Republic
Beránek Jaroslav, doc. RNDr., CSc.	Masaryk University, Brno, |Czech Republic
Bezrucko Aleksandrs, Mgr.	Riga Technical University, Riga, Latvia
Biswas Md. Haider Ali, M Phil	Khulna University, Khulna, Bangladesh
Bujok Petr, Mgr.	University of Ostrava, Ostrava, Czech Republic
Carkova Viktorija, Dr. Math.	The Latvian University, Riga, Latvia
Carkovs Jevgenijs, Dr. hab. math.	The Latvian University, Riga, Latvia
Cerny Jindrich, Ing.	University of Economics, Praha, Czech Republic
Čada Roman	University of West Bohemia, Plzen, |Czech Republic
Čadil Jan, PhD.	Economics and Management University, Prague, Czech Republic
Dobrucky Branislav, Prof.	University of Zilina, Zilina, Slovak Republic
Doležalová Jarmila, doc., RNDr., CSc.	VŠB-TU, Ostrava – Poruba, Czech Republic
Dragomirescu Florica Ioana, Lecturer	University "Politehnica" of Timisoara, Timisoara, Romania
Fabbri Franco, Dr.	University of Turin, Turin, Italy

Ferreira Manuel Alberto M., Professor Catedrático	ISCTE – IUL, Lisboa, Portugal
Filipe José António, Professor Auxiliar	ISCTE – IUL, Lisboa, Portugal
Gavalec Martin, professor RNDr CSc	University of Hradec Králové, Hradec Králové, Czech Republic
Habiballa Hashim, RNDr. PaedDr., PhD.	University of Ostrava, Ostrava, Czech Republic
Hanzel Pavol, Prof., RNDr., CSc.	Matej Bel University, Banská Bystrica, |Slovak Republic
Henzler Jiří, doc. RNDr., CSc.	University of Economics, Prague, Czech Republic
Hinterleitner Irena, PhD.	Brno University of Technology, Brno, Czech Republic
Hola Bohdana, master	University of Economy, Prague, Czech Republic
Hošková-Mayerová Šárka, Assoc. Prof., RNDr., PhD.	University of Defence, Brno, Czech Republic
Hrbáč Lubomír, doc. Dr. Ing.	Technical University Ostrava, Ostrava, Czech Republic
Hušek Miroslav, profesor	MFF,UK, Praha, Czech Republic
Chvalina Jan, Prof. RNDr., DrSc.	Brno University of Technology, Brno, Czech Republic
Ioan Rus A., Professor	Babes-Bolyai University, Cluj-Napoca, Romania
Jančařík Antonín, dr.	Charles University, Prague, Czech Republic
Jánošíková Ľudmila, doc., Ing., PhD.	,University of Zilina, Zilina, Slovak Republic
Jukl Marek, RNDr., PhD.	Palacký University, Olomouc, Czech Republic
Klazar Martin, doc. RNDr. Dr.	MFF,UK, Praha, Czech Republic
Klůfa Jindřich, Prof.RNDr., CSc.	University of Economics, Prague, Czech Republic
Klufová Renata, RNDr., PhD.	University of South Bohemia, České Budějovice, Czech Republic
Košťál Igor, Ing.,PhD.	University of Trencin, Trencin, Slovak Republic
Kováčová Monika, Mgr., PhD.	Slovak University of Technology, Bratislava, Slovak Republic
Kovár Martin, doc., RNDr., PhD.	University of Technology, Brno, Czech Republic
Kovarik Vladimir	Czech Republic
Kreml Pavel, doc. RNDr., CSc.	VŠB-TU Ostrava, Ostrava - Poruba, Czech Republic

Kuková Mária, Mgr.	Matej Bel Univerzity, Banská Bystrica, Slovak Republic
Kures Miroslav	Brno University of Technology, Brno, Czech Republic
Lacina Karel, Prof., PhD., DrSc.	University of Finances and Public Administration, Prague, Czech Republic
Liviu Cadariu, Lecturer, PhD	University of Timisoara, Timisoara, Romania
Lopes Ana Paula, PhD.	ISCAP,Polytechnic Institute of Oporto – IPP, Porto, Portugal
Lungu Nicolaie, Prof.	Technical university of Cluj-Napoca, Cluj-Napoca, Romania
Malacká Zuzana, RNDr., PhD.	University of Zilina, Zilina, Slovak Republic
Malek Josef, Prof. RNDr. DSc., CSc.	Charles University, Prague, Czech Republic
Maroš Bohumil, docent, RNDr., CSc.	Vysiké učení technické, Brno, Czech Republic
Martincova Penka, doc., Ing., PhD.	University of Zilina, Žilina, Slovak Republic
Matvejevs Andrejs, Dr.sc. Ing.	Riga Technical university, Riga, Latvia
Matvejevs Aleksandrs, Dr.math.	Riga Technical university, Riga, Latvia
Menzio Maria Rosa, Dr. math	Torino, Italy
Mikeš Josef, Prof. RNDr. DrSc.	Palacky University, Olomouc, Czech Republic
Miskolczi Martina, Ing. Mgr.	Vysoká škola ekonomická, Prague, Czech Republic
Mišútová Mária, doc. RNDr., PhD.	Slovak University of Technology, Trnava, Slovak Republic
Moučka Jiří, doc. RNDr., PhD.	University of Defence, Brno, Czech Republic
Muresan Anton S., Professor	Babes-Bolyai University, Cluj-Napoca, Romania
Neuman František, Prof. RNDr., DrSc.	Czech Academy of Sciences, Brno, Czech Republic
Nosková Barbora, Ing.	The University of Economics, Prague, Czech Republic
Okrajek Petr, Mgr.	Přírodovědecká fakulta, Brno, Czech Republic
Oplatkova Zuzana, Ing., PhD.	Tomas Bata University in Zlin, Zlin, Czech Republic
Pokorný Milan, PaedDr., PhD.	Trnava University, Trnava, Slovak Republic
Pokorny Michal, Prof., Ing., PhD.	Univerzity of Zilina, Zilina, Slovak Republic

Wimmer Gejza, Professor	Matej Bel University, Banská Bystrica, Slovak Republic
Winitzky de Spinadel Vera Martha, Dr in Mathematical Science	Ciudad Universitaria - Pabelleon III, Buenos Aires, Argentina
Witkovský Viktor, doc. RNDr., CSc.	Academy of Sciences, Bratislava, Slovak Republic
Zeithamer Tomáš, Ing., PhD.	University of Economics, Prague, Czech Republic
Zuzana Chvátalová, RNDr., PhD.	Brno University of Technology, Brno, Czech Republic
Žáček Martin, Mgr.	University of Ostrava, Ostrava, Czech Republic
Žváček Jiří, doc., CSc.	Charles Umiversity, Prague, Czech Republic

Journal of Applied Mathematics

volume IV (2011), number I

INTEGRAL AND INTEGRO-DIFFERENTIAL EQUATIONS, MODIFICATIONS OF ADOMIAN DECOMPOSITION METHOD AND HOMOTOPY PERTURBATION METHOD.

ARCHALOUSOVÁ Olga , (CZ) ŠMARDA Zdeněk , (CZ)

Abstract. In the paper analytic methods, the homotopy perturbation method and Adomian decomposition method are applied for solving integral equations and high-order integro-differential equations. To illustrate the ability and reliability of these methods numerical examples are given revealing their effectiveness and simplicity.

Key words and phrases. Integral and integro-differential equation, Homotopy perturbation method, Adomian decomposition method.

Mathematics Subject Classification. Primary 45J05; Secondary 34A08.

1 Introduction

In the paper, we intend to carry out a comparative study between He's homotopy perturbation method [4-5] and the Adomian decomposition metod [1-2]. He's homotopy perturbation method has a constructive attraction that provides the exact solution by computing only a few iterations of the solution series. In addition, He's technique may give the exact solution for nonlinear equations without any need for so-called He's polynomials. In both methods a solution is considered as the summation of an infinite series which usually converges rapidly to the exact solution. Using these methods we determine the exact solutions and approximate solutions for certain classes of integral and integro-differential equations.

1.1 Homotopy perturbation method

To illustrate the homotopy perturbation method, He [4] considered the following nonlinear differential equation:

$$A(u) = f(r), \quad r \in \Omega, \tag{1}$$

with boundary conditions

$$B(u, \partial u/\partial n) = 0, \quad r \in \Gamma,$$

where a is a general differential operator, B is a boundary operator, $f(r)$ is a known analytic function, Γ is the boundary of the domain Ω. Suppose the operator A can be divided into two parts: M and N. Then (1) can be rewritten as follows:

$$M(u) + N(u) = f(r). \tag{2}$$

He in [5] constructed a convex homotopy $H(v,p)\colon \Omega \times [0,1] \to R$ which satisfies

$$H(v,p) = (1-p)[M(v) - M(y_0)] + p[A(v) - f(r)] = 0, \tag{3}$$

or

$$H(v,p) = M(v) - M(y_0) + pM(y_0) + p[N(v) - f(r)] = 0, \tag{4}$$

where $v \in \Omega$ and $p \in [0,1]$ is an imbedding parameter, y_0 is an approximation of (1). It is obvious that

$$H(v,0) = M(v) - M(y_0) = 0, \qquad H(v,1) = A(v) - f(r) = 0,$$

and changing the variation of p from 0 to 1 is the same as changing $H(v,p)$ from $M(v) - M(y_0)$ to $A(v) - f(r)$. In topology, this is called deformation, $M(v) - M(y_0)$ and $A(v) - f(r)$ are called homotopic. Owing to the fact that $0 \le p \le 1$ can be considered as a small parameter we can assume the solution (3) and (4) can be expressed as a series in p, as follows:

$$v = v_0 + pv_1 + pv_2 + \dots \tag{5}$$

when $p \to 1$ equations (3) and (4) correspond to equation (2) and (5) becomes the approximate solution of equation (2), i.e.

$$u(x) = \lim_{p \to 1} v = v_0 + v_1 + v_2 + \dots . \tag{6}$$

The series (6) is convergent for most of the cases and also the rate of convergence depends on how we choose $A(v)$ (see [6]).

1.2 Adomian decomposition method

We recall basic principles of the Adomian decomposition method. Consider the general equation $Tu = g$, where T represents a general nonlinear differential operator involving both linear and nonlinear terms. The linear term is decomposed into $L + R$ where L is easily invertible and R

is the reminder of the linear operator. For convenience, L may be taken as the highest order derivation. Thus the equation may be written as

$$Lu + Ru + Nu = g, \tag{7}$$

where Nu represents the nonlinear terms. From (7) we have

$$Lu = g - Ru - Nu. \tag{8}$$

Since L is invertible the equivalent expression is

$$u = L^{-1}g - L^{-1}Ru - L^{-1}Nu. \tag{9}$$

A solution u can be expressed as following series

$$u = \sum_{n=0}^{\infty} u_n, \tag{10}$$

with reasonable u_0 which may be identified with respect to the definition of L^{-1} , g and $u_n,\ n > 0$ is to be determined. The nonlinear term Nu will be decompsed by the infinite series of Adomian polynomials

$$Nu = \sum_{n=0}^{\infty} A_n, \tag{11}$$

where polynomials A_n are given by

$$v(\lambda) = \sum_{n=0}^{\infty} \lambda^n u_n \tag{12}$$

$$N(v(\lambda)) = \sum_{n=0}^{\infty} \lambda^n A_n. \tag{13}$$

Here λ is a parameter introcuded for convenience. From (12) and (13) we have

$$A_n = \frac{1}{n!}\frac{d^n}{d\lambda^n}\left[N\sum_{k=0}^{\infty}\lambda^k u_k\right]_{\lambda=0}, \quad n \geq 0. \tag{14}$$

The Adomian polynomials A_n are given as

$$\begin{aligned}
A_0 &= F(u_0) \\
A_1 &= u_1F'(u_0) \\
A_2 &= u_2F'(u_0) + \frac{u_1^2}{2!}F''(u_0) \\
A_3 &= u_3F'(u_0) + u_1u_2F''(u_o) + \frac{u_1^3}{3!}F'''(u_0) \\
&\vdots
\end{aligned} \tag{15}$$

Now substituting (10) and (11) into (9) we get

$$\sum_{n=0}^{\infty} u_n = u_0 + L^{-1}R\left(\sum_{n=0}^{\infty} u_n\right) - L^{-1}\sum_{n=0}^{\infty} A_n. \tag{16}$$

Consequently, with a suitable u_0 we can write

$u_1 = -L^{-1}Ru_0 - L^{-1}A_0$

$\vdots$

$u_{n+1} = -L^{-1}Ru_n - L^{-1}A_n$

$\vdots$

2 Integro-differential equations

To illustrate the basic ideas of the homotopy perturbation method, let us consider the following integro-differential equation

$$y''(x) - xy(x) = g(x) + \int_0^x x^2 e^t y(t)dt, \quad y(0) = 1,\ y'(0) = 0, \tag{17}$$

where

$$g(x) = -(1+x)\cos x - \frac{x^2}{2}\big(e^x(\cos x + \sin x) - 1\big).$$

For solving of (17) we consider the following homotopy

$$(1-p)(Y''(x) - y_0(x)) + p\Big(Y''(x) - xY(x) - g(x) - \int_0^x x^2 e^t Y(t)dt\Big) = 0. \tag{18}$$

By integration of equation (18) we obtain

$$\begin{aligned} Y(x) &= Y(0) + xY'(0) + \int_0^x \int_0^\tau y_0(s)dsd\tau \\ &- p\int_0^x \int_0^\tau \left(y_0(s) - sY(s) - g(s) - \int_0^s s^2 e^t Y(t)dt\right) dsd\tau. \end{aligned} \tag{19}$$

Suppose that the solution of (19) is in the form

$$Y(x) = Y_0(x) + pY_1(x) + p^2 Y_2(x) + \dots\ . \tag{20}$$

Substituting (20) into (19) and equating coefficients of p with the same power lead to

$$\begin{aligned} p^0 &: \quad Y_0(x) = 1 + \int_0^x \int_0^\tau y_0(s)dsd\tau, \\ p^1 &: \quad Y_1(x) = -\int_0^x \int_0^\tau \left(y_0(s) - sY_0(s) - g(s) - \int_0^s s^2 e^t Y_0(t)dt\right) dsd\tau, \\ &\vdots \\ p^j &: \quad Y_1(x) = -\int_0^x \int_0^\tau \left(-sY_{j-1}(s) + \int_0^s s^2 e^t Y_{j-1}(t)dt\right) dsd\tau, \quad j = 2, 3, \dots. \end{aligned}$$

Now assume that $y_0(x) = \sum_{n=0}^{\infty} a_n x^n,\ Y(0) = 1$ and the Taylor series of $Y_1(x)$ equal to zero. Then we get

$$\begin{aligned} &- (\frac{1}{2} + \frac{a_0}{2})x^2 - \frac{a_1}{6}x^3 + (\frac{1}{24} - \frac{a_2}{12})x^4 + (\frac{1}{40} + \frac{a_0}{40} - \frac{a_3}{20})x^5 + (-\frac{1}{720} - \frac{a_4}{30} + \frac{a_1}{180})x^6 \\ &+ (\frac{1}{336} + \frac{a_0}{252} - \frac{a_5}{42} + \frac{a_2}{504})x^7 + (\frac{13}{5760} + \frac{a_0}{448} + \frac{a_3}{1120} - \frac{a_6}{56} + \frac{a_1}{1344})x^8 + \cdots = 0. \end{aligned}$$

From here we have

$$a_0 = -1,\ a_1 = 0,\ a_2 = \frac{1}{2},\ a_3 = 0, a_4 = -\frac{1}{24},\ a_5 = 0, a_6 = \frac{1}{720},\ a_7 = 0, \ldots$$

Hence the exact solution of equation (17) has the form

$$\begin{aligned} y(x) &= Y_0(x) = 1 + \frac{a_0}{2}x^2 + \frac{a_1}{6}x^3 + \frac{a_2}{12}x^4 + \frac{a_3}{20}x^5 + \frac{a_4}{30}x^6 + \ldots \\ &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n}}{(2n)!} = \cos x. \end{aligned}$$

3 Fredholm integral equation

Consider the following Fredholm integral equation

$$u(x) = \cos\frac{\pi}{2}x + 2 - \int_0^1 \left(1 + \frac{\pi}{2}\left[\cos\frac{\pi}{2}t \sin^2\frac{\pi}{2}t + u^3(t)\right]\right)dt \tag{21}$$

and we solve it by using He's homotopy perturbation method which can be written in the following form:

$$Y(x) - \cos\frac{\pi}{2}x - 2p + p\int_0^1 \left(1 + \frac{\pi}{2}\left[\cos\frac{\pi}{2}t \sin^2\frac{\pi}{2}t + Y^3(t)\right]\right)dt = 0. \tag{22}$$

Substituting (20) into (22) and equating the coefficients of like terms with the identical powers of p we obtain

$$\begin{aligned} p^0 &: \quad Y_0(x) - \cos\frac{\pi}{2}x = 0 \Rightarrow Y_0(x) = \cos\frac{\pi}{2}x, \\ p^1 &: \quad Y_1(x) - 2 + \int_0^1 \left(1 + \frac{\pi}{2}\left[\cos\frac{\pi}{2}t \sin^2\frac{\pi}{2}t + Y_0^3(t)\right]\right)dt = 0 \Rightarrow Y_1(x) = 0, \\ p^2 &: \quad Y_2(x) + \int_0^1 \frac{\pi}{2}3Y_0^2(t)Y_1(t)dt = 0 \Rightarrow Y_2(x) = 0, \\ &\vdots \end{aligned} \tag{23}$$

It is obvious that $Y_k(x) = 0$ for $k \geq 1$. hence the exact solution of (21) has the form $Y(x) = \cos\frac{\pi}{2}x$.

Now we will solve equation (21) using the Adomian decomposition method. Substituting the decomposition series we get

$$\sum_{k=0}^{\infty} u_k(x) = \cos\frac{\pi}{2}x + 2 - \int_0^1 \Big(1 + \frac{\pi}{2}\Big[\cos\frac{\pi}{2}t\sin^2\frac{\pi}{2}t + \Big(\sum_{k=0}^{\infty} u_k(t)\Big)^3\Big]\Big)dt.$$

Thus

$$\sum_{k=0}^{\infty} u_k(x) = \cos\frac{\pi}{2}x + \frac{2}{3} - \frac{\pi}{2}\int_0^1 \Big(\sum_{k=0}^{\infty} u_k(t)\Big)^3 dt. \tag{24}$$

Put the zero component $u_0(x) = \cos\frac{\pi}{2}x + \frac{2}{3}$ then the remaining components $u_n(x),\ n \geq 1$ can be determined using the recurrence relations

$$u_n(x) = -\frac{\pi}{2}\int_0^1 A_n(t)dt,$$

where A_n are the Adomian polynomials. Thus

$$\begin{aligned}
u_1(x) &= -2 - \frac{35}{54}\pi,\\
u_2(x) &= 8 + \frac{293}{54}\pi + \frac{595}{648}\pi^2,\\
u_3(x) &= -35 - \frac{928}{27}\pi - \frac{7921}{648}\pi^2 - \frac{100835}{69984}\pi^3,\\
&\vdots
\end{aligned} \tag{25}$$

From here we obtain the solution $u(x)$ in a series form

$$u(x) = \cos\frac{\pi}{2}x - 29 - \frac{799}{27}\pi - \frac{407}{36}\pi^2 - \frac{100835}{69984}\pi^3 + \dots.$$

Hence the Adomian decomposition method cannot lead to the closed form of the solution as in the case of He's perturbation method. But obtaining of the closed form of the solution, using He's perturbation method, depends on selection of a homotopy equation.

Acknowledgement

This research has been supported by the Czech Ministry of Education in the frames of MSM002160503 Research Intention MIKROSYN New Trends in Microelectronic Systems and Nanotechnologies and MSM0021630529 Research Intention Inteligent Systems in Automatization.

References

[1] ADOMIAN, G.: *A Review of the Decomposition Method in Applied Mathematics*, J. Math. Anal. Appl. **135**(2-3) (1988), 501-544.

[2] ADOMIAN, G.: *Solving Frontier Problems of Physics: The decomposition method,* Kluwer Academic Publishers, Boston, (1994).

[3] GHORBANI, A.: *Beyond Adomian polynomials: He polynomials,* Chaos Solitons Fractals **39** (2009), 1486-1492.

[4] HE, J.H.: *Homotopy perturbation technique,* Comput. Methods Appl. Mech. Engrg. **178** (1999), 257-262.

[5] HE, J.H.: *A coupling method of a homotopy technique and a perturbation technique for non-linear problems,* J. Nonlinear Mech. **35** (2000), 37-45.

[6] NADJAFI, J.S., GHORBANI, A.: *He's homotopy perturbation method: An effective tool for solving nonlinear integral and integro-differential equations,* Computers and Mathematics with Applications **58** (2009), 2379-2390.

Current address

Ing. Olga Archalousová
Department of Mathematics
Faculty of Electrical Engineering and Communication
Brno University of Technology, Technická 8, 616 00 BRNO
e-mail: xfilip03@stud.feec.vutbr.cz

Doc. RNDr. Zdeněk Šmarda, CSc.
Department of Mathematics
Faculty of Electrical Engineering and Communication
Brno University of Technology, Technická 8, 616 00 BRNO
e-mail: smarda@feec.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

OSCILLATION OF SOLUTION OF A LINEAR THIRD-ORDER DISCRETE DELAED EQUATION

BAŠTINEC Jaromír, (CZ), DIBLÍK Josef, (CZ), BAŠTINCOVÁ (RYVOLOVÁ) Alena, (CZ)

Abstract. A linear third-order discrete delayed equation $\Delta x(n) = -p(n)x(n-2)$ with a positive coefficient p is considered for $n \to \infty$. This equation is known to have a positive solution if p fulfils an inequality. The goal of the paper is to show that, in the case of the opposite inequality for p, all solutions of the equation considered are oscillating for $n \to \infty$.

Key words and phrases. Discrete delayed equation, oscillating solution, positive solution, asymptotic behavior.

Mathematics Subject Classification. Primary 39A10, Secondary 39A11.

1 Introduction

The existence of a positive solution of difference equations is often encountered when analysing mathematical models describing various processes. This is a motivation for an intensive study of the conditions for the existence of positive solutions of discrete or continuous equations. Such analysis is related to an investigation of the case of all solutions being oscillating (for relevant investigation in both directions we refer, e.g., to [1]–[15] and to the references therein). In this paper, sharp conditions are derived for all the solutions being oscillating for a class of linear second-order delayed discrete equations.

We consider the delayed third-order linear discrete equation

$$\Delta x(n) = -p(n)x(n-2) \tag{1}$$

where $n \in \mathbb{Z}_a^\infty := \{a, a+1, \dots\}$, $a \in \mathbb{N}$ is fixed, $\Delta x(n) = x(n+1) - x(n)$, $p : \mathbb{Z}_a^\infty \to \mathbb{R}^+ := (0, \infty)$. A solution $x = x(n) : \mathbb{Z}_a^\infty \to \mathbb{R}$ of (1) is positive (negative) on $\mathbb{Z}_a^\infty$ if $x(n) > 0$ $(x(n) < 0)$ for every $n \in \mathbb{Z}_a^\infty$. A solution $x = x(n) : \mathbb{Z}_a^\infty \to \mathbb{R}$ of (1) is oscillating on $\mathbb{Z}_a^\infty$ if it is not positive or negative on $\mathbb{Z}_{a_1}^\infty$ for arbitrary $a_1 \in \mathbb{Z}_a^\infty$.

Definition 1.1 *Let us define the expression* $\ln_q t$, $q \ge 1$, *by* $\ln_q t = \ln(\ln_{q-1} t)$, $\ln_0 t \equiv t$ *where* $t > \exp_{q-2} 1$ *and* $\exp_s t = \exp\left(\exp_{s-1} t\right)$, $s \ge 1$, $\exp_0 t \equiv t$ *and* $\exp_{-1} t \equiv 0$ (*instead of* $\ln_0 t$, $\ln_1 t$, *we will only write* t *and* $\ln t$).

In [2] a delayed linear difference equation of higher order is considered and the following result related to equation (1) on the existence of a positive solution solution is proved.

Theorem 1.2 *Let* $a \in \mathbb{N}$ *be sufficiently large and* $q \in \mathbb{N}$. *If the function* $p : \mathbb{Z}_a^\infty \to \mathbb{R}^+$ *satisfies*

$$\begin{aligned} p(n) \le \frac{4}{27} + \frac{1}{9n^2} + \frac{1}{9(n\ln n)^2} + \frac{1}{9(n\ln n\ln_2)^2} + \frac{1}{9(n\ln n\ln_2 n\ln_3 n)^2} \\ + \dots + \frac{1}{9(n\ln n\ln_2 n\dots\ln_q n)^2}. \end{aligned} \tag{2}$$

for every $n \in \mathbb{Z}_a^\infty$ *then there exists a positive integer* $a_1 \ge a$ *and a solution* $x = x(n)$, $n \in \mathbb{Z}_{a_1}^\infty$ *of equation* (1) *such that* $x(n) > 0$ *holds for every* $n \in \mathbb{Z}_{a_1}^\infty$.

It is the open question whether all solutions of (1) are oscillating if inequality (2) is replaced by the opposite inequality

$$\begin{aligned} p(n) \ge \frac{4}{27} + \frac{1}{9n^2} + \frac{1}{9(n\ln n)^2} + \frac{1}{9(n\ln n\ln_2 n)^2} + \frac{1}{9(n\ln n\ln_2 n\ln_3 n)^2} \\ + \dots + \frac{1}{9(n\ln n\ln_2 n\dots\ln_{q-1} n)^2} + \frac{\kappa}{9(n\ln n\ln_2 n\dots\ln_q n)^2} \end{aligned} \tag{3}$$

assuming $\kappa > 1$ and n is sufficiently large.

Below we give a partial answer related to this open question. Namely, we prove that if inequality

$$p(n) \ge \frac{4}{27} + \frac{1}{9n^2} + \frac{\kappa}{9(n\ln n)^2} \tag{4}$$

holds where $\kappa > 1$ and n is sufficiently large, then all solutions of (1) are oscillatory. The proof of our main result will use a consequence of one of Y. Domshlak's results [8, Theorem 4]:

Lemma 1.3 *Let* q *and* r *be fixed natural numbers such that* $r - q > 2$. *Let* $\{\varphi(n)\}_1^\infty$ *be a bounded sequence of positive numbers and* ν_0 *be a positive number such that there exists a number* $\nu \in (0, \nu_0)$ *satisfying*

$$\sum_{q+1}^{s} \varphi(n) \le \frac{\pi}{\nu}, \quad s = q+1, \dots, r, \qquad \frac{\pi}{\nu} \le \sum_{q+1}^{r+s} \varphi(n) \le \frac{2\pi}{\nu}, \quad s = 1, 2. \tag{5}$$

Then, if for $n \in \mathbb{Z}_{q+3}^r$

$$p(n) \ge \mathcal{R} := \prod_{l=n-2}^{n} \left[\frac{\sin\left(\nu \sum\limits_{i=l+1}^{l+2} \varphi(i)\right)}{\sin\left(\nu \sum\limits_{i=l}^{l+2} \varphi(i)\right)} \right] \cdot \frac{\sin \nu\varphi(n-2)}{\sin \nu \sum\limits_{i=n-1}^{n} \varphi(i)} \tag{6}$$

holds, then any solution of the equation

$$x(n+1) - x(n) + p(n)x(n-2) = 0$$

has at least one change of sign on $\mathbb{Z}^r_{q-2}$.

Moreover, we will use an auxiliary result giving the asymptotic decomposition of the iterative logarithm [7]. The symbols "o" and "O" used below (for $n \to \infty$) stand for the Landau order symbols.

Lemma 1.4 *For fixed* $r, \sigma \in \mathbb{R} \setminus \{0\}$ *and fixed integer* $s \geq 1$ *the asymptotic representation*

$$\begin{aligned}\frac{\ln_s^\sigma(n-r)}{\ln_s^\sigma n} =& 1 - \frac{r\sigma}{n \ln n \dots \ln_s n} - \frac{r^2\sigma}{2n^2 \ln n \dots \ln_s n} \\ &- \frac{r^2\sigma}{2(n\ln n)^2 \ln_2 n \dots \ln_s n} - \dots - \frac{r^2\sigma}{2(n \ln n \dots \ln_{s-1} n)^2 \ln_s n} \\ &+ \frac{r^2\sigma(\sigma-1)}{2(n \ln n \dots \ln_s n)^2} - \frac{r^3\sigma(1+o(1))}{3n^3 \ln n \dots \ln_s n}\end{aligned}$$

holds for $n \to \infty$.

2 Main Result

In this part we give sufficient conditions for all solutions of equation (1) to be oscillatory as $n \to \infty$.

Theorem 2.1 *Let* $a \in \mathbb{N}$ *be sufficiently large,* $q \in \mathbb{N}$ *and* $\kappa > 1$. *Assuming that the function* $p \colon \mathbb{Z}_a^\infty \to \mathbb{R}^+$ *satisfies inequality* (4) *for every* $n \in \mathbb{Z}_a^\infty$, *all solutions of* (1) *are oscillating as* $n \to \infty$.

Proof. We set

$$\varphi(n) := \frac{1}{n \ln n}$$

and consider the asymptotic decomposition of $\varphi(n-1)$ when n is sufficiently large. Applying Lemma 1.4 (for $\sigma = -1$, $r = 1$ and $s = 1$), we get

$$\begin{aligned}\varphi(n-1) &= \frac{1}{(n-1)\ln(n-1)} = \frac{1}{n\,(1-1/n)\ln(n-1)} = \varphi(n) \cdot \frac{1}{1-1/n} \cdot \frac{\ln n}{\ln(n-1)} \\ &= \varphi(n) \left(1 + \frac{1}{n} + \frac{1}{n^2} + O\left(\frac{1}{n^3}\right)\right)\left(1 + \frac{1}{n \ln n} + \frac{1}{2n^2 \ln n} + \frac{1}{(n\ln n)^2} + O\left(\frac{1}{n^3}\right)\right).\end{aligned}$$

Finally, we obtain

$$\varphi(n-1) = \varphi(n)\left(1 + \frac{1}{n} + \frac{1}{n\ln n} + \frac{1}{n^2} + \frac{3}{2n^2\ln n} + \frac{1}{(n\ln n)^2} + O\left(\frac{1}{n^3}\right)\right). \tag{7}$$

Applying Lemma 1.4 (for $\sigma = -1$, $r = 2$ and $s = 1$), we get

$$\begin{aligned}\varphi(n-2) &= \frac{1}{(n-2)\ln(n-2)} = \frac{1}{n\,(1-2/n)\ln(n-2)} = \varphi(n)\cdot\frac{1}{1-2/n}\cdot\frac{\ln n}{\ln(n-2)}\\ &= \varphi(n)\left(1 + \frac{2}{n} + \frac{4}{n^2} + O\left(\frac{1}{n^3}\right)\right)\left(1 + \frac{2}{n\ln n} + \frac{4}{2n^2\ln n} + \frac{4}{(n\ln n)^2} + O\left(\frac{1}{n^3}\right)\right).\end{aligned}$$

Finally, we obtain

$$\varphi(n-2) = \varphi(n)\left(1 + \frac{2}{n} + \frac{2}{n\ln n} + \frac{4}{n^2} + \frac{6}{n^2\ln n} + \frac{4}{(n\ln n)^2} + O\left(\frac{1}{n^3}\right)\right). \tag{8}$$

Similarly, applying Lemma 1.4 (for $\sigma = -1$, $r = -1$ and $s = 1$), we get

$$\begin{aligned}\varphi(n+1) &= \frac{1}{(n+1)\ln(n+1)} = \frac{1}{n\,(1+1/n)\ln(n+1)} = \varphi(n)\cdot\frac{1}{1+1/n}\cdot\frac{\ln n}{\ln(n+1)}\\ &= \varphi(n)\left(1 - \frac{1}{n} + \frac{1}{n^2} + O\left(\frac{1}{n^3}\right)\right)\left(1 - \frac{1}{n\ln n} + \frac{1}{2n^2\ln n} + \frac{1}{(n\ln n)^2} + O\left(\frac{1}{n^3}\right)\right)\\ &= \varphi(n)\left(1 - \frac{1}{n} - \frac{1}{n\ln n} + \frac{1}{n^2} + \frac{3}{2n^2\ln n} + \frac{1}{(n\ln n)^2} + O\left(\frac{1}{n^3}\right)\right). \qquad (9)\end{aligned}$$

And finally, applying Lemma 1.4 (for $\sigma = -1$, $r = -2$ and $s = 1$), we get

$$\begin{aligned}\varphi(n+2) &= \frac{1}{(n+2)\ln(n+2)} = \frac{1}{n\,(1+2/n)\ln(n+2)} = \varphi(n)\cdot\frac{1}{1+2/n}\cdot\frac{\ln n}{\ln(n+2)}\\ &= \varphi(n)\left(1 - \frac{2}{n} + \frac{4}{n^2} + O\left(\frac{1}{n^3}\right)\right)\left(1 - \frac{2}{n\ln n} + \frac{2}{n^2\ln n} + \frac{4}{(n\ln n)^2} + O\left(\frac{1}{n^3}\right)\right)\\ &= \varphi(n)\left(1 - \frac{2}{n} - \frac{2}{n\ln n} + \frac{4}{n^2} + \frac{6}{n^2\ln n} + \frac{4}{(n\ln n)^2} + O\left(\frac{1}{n^3}\right)\right). \qquad (10)\end{aligned}$$

The right-hand side of inequality (6) we can rewrite in the form

$$\prod_{l=n-2}^{n}\left[\frac{\sin\left(\nu\sum\limits_{i=l+1}^{l+2}\varphi(i)\right)}{\sin\left(\nu\sum\limits_{i=l}^{l+2}\varphi(i)\right)}\right]\cdot\frac{\sin\nu\varphi(n-2)}{\sin\nu\sum\limits_{i=n-1}^{n}\varphi(i)}$$

$$= \frac{\sin\nu\varphi(n-2)}{\sin\nu\sum\limits_{i=n-1}^{n}\varphi(i)}\prod_{l=n-2}^{n}\left[\frac{\sin\left(\nu\sum\limits_{i=l+1}^{l+2}\varphi(i)\right)}{\sin\left(\nu\sum\limits_{i=l}^{l+2}\varphi(i)\right)}\right]$$

$$= \frac{\sin \nu\varphi(n-2)}{\sin\left(\nu\left(\varphi(n-1)+\varphi(n)\right)\right)} \cdot \frac{\sin\left(\nu\left(\varphi(n-1)+\varphi(n)\right)\right)}{\sin\left(\nu\left(\varphi(n-2)+\varphi(n-1)+\varphi(n)\right)\right)}$$
$$\times \frac{\sin\left(\nu\left(\varphi(n)+\varphi(n+1)\right)\right)}{\sin\left(\nu\left(\varphi(n-1)+\varphi(n)+\varphi(n+1)\right)\right)} \cdot \frac{\sin\left(\nu\left(\varphi(n+1)+\varphi(n+2)\right)\right)}{\sin\left(\nu\left(\varphi(n)+\varphi(n+1)+\varphi(n+2)\right)\right)}$$
$$= \frac{\sin \nu\varphi(n-2)}{\sin\left(\nu\left(\varphi(n-2)+\varphi(n-1)+\varphi(n)\right)\right)} \times \frac{\sin\left(\nu\left(\varphi(n)+\varphi(n+1)\right)\right)}{\sin\left(\nu\left(\varphi(n-1)+\varphi(n)+\varphi(n+1)\right)\right)}$$
$$\times \frac{\sin\left(\nu\left(\varphi(n+1)+\varphi(n+2)\right)\right)}{\sin\left(\nu\left(\varphi(n)+\varphi(n+1)+\varphi(n+2)\right)\right)} = (*).$$

Recalling the asymptotical decomposition of $\sin x$ when $x \to 0$: $\sin x = x + O(x^3)$, we get (since $\lim_{n\to\infty} \varphi(n+j) = 0$, $\varphi(n+j) = O(\varphi(n))$, $\varphi(n) = O(\varphi(n+j))$, $j = 0, \pm 1, \pm 2$ and (7)–(10) hold)

$$(*) = \frac{\nu\varphi(n-2) + O\left((\nu\varphi(n))^3\right)}{\left(\nu\left(\varphi(n-2)+\varphi(n-1)+\varphi(n)\right)\right) + O\left((\nu\varphi(n))^3\right)}$$
$$\times \frac{\left(\nu\left(\varphi(n)+\varphi(n+1)\right)\right) + O\left((\nu\varphi(n)))^3\right)}{\left(\nu\left(\varphi(n-1)+\varphi(n)+\varphi(n+1)\right)\right) + O\left((\nu\varphi(n))^3\right)}$$
$$\times \frac{\left(\nu\left(\varphi(n+1)+\varphi(n+2)\right)\right) + O\left((\nu\varphi(n))^3\right)}{\left(\nu\left(\varphi(n)+\varphi(n+1)+\varphi(n+2)\right)\right) + O\left((\nu\varphi(n))^3\right)}$$
$$= \frac{\varphi(n-2)}{\varphi(n-2)+\varphi(n-1)+\varphi(n)} \cdot \frac{\varphi(n)+\varphi(n+1)}{\varphi(n-1)+\varphi(n)+\varphi(n+1)}$$
$$\times \frac{\varphi(n+1)+\varphi(n+2)}{\varphi(n)+\varphi(n+1)+\varphi(n+2)} \cdot \left(1 + O\left((\nu\varphi(n))^2\right)\right) = (**)$$

Using previous the decompositions, we have

$$\frac{\varphi(n-2)}{\varphi(n-2)+\varphi(n-1)+\varphi(n)}$$
$$= \frac{1}{3}\left(1 + \frac{1}{n} + \frac{1}{n\ln n} + \frac{4}{3n^2} + \frac{3}{2n^2\ln n} + \frac{4}{3(n\ln n)^2}\right) + O\left(\frac{1}{n^3}\right),$$
$$\frac{\varphi(n)+\varphi(n+1)}{\varphi(n-1)+\varphi(n)+\varphi(n+1)}$$
$$= \frac{2}{3}\left(1 - \frac{1}{2n} - \frac{1}{2n\ln n} - \frac{1}{6n^2} - \frac{1}{4n^2\ln n} - \frac{1}{6(n\ln n)^2}\right) + O\left(\frac{1}{n^3}\right)$$

and

$$\frac{\varphi(n+1)+\varphi(n+2)}{\varphi(n)+\varphi(n+1)+\varphi(n+2)}$$
$$= \frac{2}{3}\left(1 - \frac{1}{2n} - \frac{1}{2n\ln n} + \frac{1}{3n^2} + \frac{1}{4n^2\ln n} + \frac{1}{3(n\ln n)^2}\right) + O\left(\frac{1}{n^3}\right).$$

Then

$$(**) = \frac{4}{27}\left[\left(1 + \frac{1}{n} + \frac{1}{n\ln n} + \frac{4}{3n^2} + \frac{3}{2n^2\ln n} + \frac{4}{3(n\ln n)^2}\right) + O\left(\frac{1}{n^3}\right)\right]$$

$$\times\left[\left(1-\frac{1}{2n}-\frac{1}{2n\ln n}-\frac{1}{6n^2}-\frac{1}{4n^2\ln n}-\frac{1}{6(n\ln n)^2}\right)+O\left(\frac{1}{n^3}\right)\right]$$

$$\times\left[\left(1-\frac{1}{2n}-\frac{1}{2n\ln n}+\frac{1}{3n^2}+\frac{1}{4n^2\ln n}+\frac{1}{3(n\ln n)^2}\right)+O\left(\frac{1}{n^3}\right)\right]\cdot\left(1+O\left((\nu\varphi(n))^2\right)\right)$$

$$=\frac{4}{27}\left(1+\frac{3}{4n^2}+\frac{3}{4(n\ln n)^2}\right)\left(1+O\left(\frac{1}{n^3}\right)\right)\left(1+O\left((\nu\varphi(n))^2\right)\right)$$

$$=\frac{4}{27}+\frac{1}{9n^2}+\frac{1}{9(n\ln n)^2}+O\left(\frac{\nu^2}{(n\ln n)^2}\right).$$

Finalizing our decompositions, we see that

$$\mathcal{R}=\frac{4}{27}+\frac{1}{9n^2}+\frac{1}{9(n\ln n)^2}+O\left(\frac{\nu^2}{(n\ln n)^2}\right).$$

It is easy to see that inequality (6) becomes

$$p(n)\geq\mathcal{R}=\frac{4}{27}+\frac{1}{9n^2}+\frac{1}{9(n\ln n)^2}+O\left(\frac{\nu^2}{(n\ln n)^2}\right). \tag{11}$$

and will be valid if (see (4))

$$\frac{4}{27}+\frac{1}{9n^2}+\frac{\kappa}{9(n\ln n)^2}\geq\frac{4}{27}+\frac{1}{9n^2}+\frac{1}{9(n\ln n)^2}+O\left(\frac{\nu^2}{(n\ln n)^2}\right)$$

or

$$\kappa\geq 1+O(\nu^2) \tag{12}$$

for $n\to\infty$. If $n\geq n_0$ where n_0 is sufficiently large, then (12) holds for sufficiently small $\nu\in(0,\nu_0]$ with ν_0 fixed because $\kappa>1$. Consequently (11) is satisfied and the assumption (6) of Lemma 1.3 holds for $n\in\mathbb{Z}_{n_0}^{\infty}$. Let $q\geq n_0$ in Lemma 1.3 be fixed and $r>q+2$ be so large that inequalities (5) hold. This is always possible since the series $\sum_{n=q+1}^{\infty}\varphi(n)$ is divergent. Then Lemma 1.3 holds and any solution of equation (1) has at least one change of sign on $\mathbb{Z}_{q-2}^{r}$. Obviously, inequalities (5) can be satisfied for another couple of (q,r), say (q_1,r_1) with $q_1>r$ and $r_1>q_1+2$ sufficiently large and by Lemma 1.3 any solution of equation (1) has at least one change of sign on $\mathbb{Z}_{q_1-2}^{r_1}$. Continuing this process we get a sequence of intervals (q_n,r_n) with $\lim_{n\to\infty}q_n=\infty$ such that any solution of equation (1) has at least one change of sign on $\mathbb{Z}_{q_n-2}^{r_n}$. This fact concludes the proof.

Acknowledgement

The paper was supported by grants P201/11/0768, P201/10/1032 of the Czech Grant Agency (Prague), by the Council of Czech Government MSM 0021630529, MSM 00216 30503 and by Grant FEKT-S-11-2-921 of Faculty of Electrical Engineering and Communication, BUT.

References

[1] AGARVAL, R.P., ZAFER, A.: Oscillation criteria for second-order forced dynamic equations with mixed nonlinearities, *Adv. Difference Equ.*, 2009, Article ID 938706, 20 p. (2009).

[2] BAŠTINEC, J., DIBLÍK, J., ŠMARDA, Z.: Existence of positive solutions of discrete linear equations with a single delay, *J. Difference Equ. Appl.*, Vol. 16, Issue 9 (2010), 1047 - 1056.

[3] BEREZANSKY, L., BRAVERMAN, E.: On existence of positive solutions for linear difference equations with several delays, *Adv. Dyn. Syst. Appl.*, **1** (2006), 29–47.

[4] BEREZANSKY, L., BRAVERMAN, E.: Oscillation of a logistic difference equation with several delays, *Adv. Difference Equ.*, 2006, Article ID 82143, 12 p. (2006).

[5] BOHNER, M., KAROUS, B., ÖCALAN, Ö.: Iterated oscillation criteria for delay dynamic equations of first order, *Adv. Difference Equ.*, 2008, Article ID 458687, 12 p. (2008).

[6] CHATZARAKIS, G.E., KOPLATADZE, R., STAVROULAKIS, I.P.: On existence of positive solutions for linear difference equations with several delays, *Adv. Dyn. Syst. Appl.*, **68** (2008), 994–1005.

[7] DIBLÍK, J., KOKSCH, N.: Positive solutions of the equation $\dot{x}(t) = -c(t)x(t-\tau)$ in the critical case, *J. Math. Anal. Appl.*, **250** (2000), 635–659.

[8] DOMSHLAK, Y.: Oscillation properties of discrete difference inequalities and equations: The new approach, *Funct. Differ. Equ.*, **1**, (1993), 60–82.

[9] DOMSHLAK, Y.,STAVROULAKIS, I.P.: Oscillation of first-order delay differential equations in a critical case, *Appl. Anal.*, **61** (1996), 359–371.

[10] DOROCIAKOVÁ, B., OLACH, R.: Existence of positive solutions of delay differential equations, *Tatra Mt. Math. Publ.*, **43** (2009), 63–70.

[11] GYÖRI, I., LADAS, G.: *Oscillation Theory of Delay Differential Equations,* Clarendon Press (1991).

[12] HANUŠTIAKOVÁ, L., OLACH, R.: Nonoscillatory bounded solutions of neutral differential systems, *Nonlinear Anal.*, **68** (2008), 1816–1824.

[13] KIKINA, L.K., STAVROULAKIS, I.P.: A survey on the oscillation of solutions of first order delay difference equations, *Cubo*, **7**, No 2, (2005), 223–236.

[14] MEDINA, R., PITUK, M.: Nonoscillatory solutions of a second-order difference equation of Poincaré type, *Appl. Math. Lett.*, **22**, No 5, (2009), 679–683.

[15] STAVROULAKIS, I.P.: Oscillation criteria for first order delay difference equations, *Mediterr. J. Math.*, **1**, No 2, (2004), 231–240.

Current address

BAŠTINEC Jaromír, doc., RNDr., CSc.
Department of Mathematics, Faculty of Electrical Engineering and Communication,
Brno University of Technology,
Technická 8, 616 00 Brno, Czech Republic
telefon: +420 541 143 222
email: bastinec@feec.vutbr.cz

DIBLÍK Josef, prof., RNDr., DrSc.
Brno University of Technology,

Technická 8, 616 00 Brno, Czech Republic
telefon: +420 541 143 155
email: diblik@feec.vutbr.cz, diblik.j@fce.vutbr.cz

BAŠTINCOVÁ (RYVOLOVÁ) Alena, Mgr.
Department of Mathematics, Faculty of Electrical Engineering and Communication,
Brno University of Technology,
Technická 8, 616 00 Brno, Czech Republic
email:xbasti00@stud.feec.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

CONTROLLABILITY OF STATIONARY LINEAR SYSTEMS WITH DELAY

BAŠTINEC Jaromír, (CZ), PIDDUBNA Ganna, (CZ)

Abstract. This paper presents results of controllability researches and construction of solution and control for linear stationary systems with delay. Analysis of controllability of linear stationary dynamic system with constant aftereffect in the case of commutative matrices is conducted. Necessary and sufficient conditions for controllability are defined, general solution of the system is shown and control is build.

Key words and phrases. Differential equation with delay, linaer stationary system, commutative matrices, controllability.

Mathematics Subject Classification. Primary 34K20, 34K25; Secondary 34K12.

1 Controllability of linear stationary systems with one delay

Let X be the state space of dynamic system; U be the set of the controlled effects (controls). Let $x = x(x_0, u, t)$ be the vector that characterizes the state of the dynamic system in the moment of time t by the initial condition x_0, $x_0 \in X$, $(x_0 = x|_{t=t_0})$ and by the control function u, $u \in U$.

Definition 1.1 *The state x_0 is called a controllable state in the class U (controlled state), if there are exist such control $u, u = u_{x_0} \in U$ and the number T, $t_0 \leq T = T_{x_0} < \infty$ that $x(x_0, u, T) = (0, ..., 0)^T$.*

Definition 1.2 *If every state $x_0, x_0 \in X$ of the dynamic system is controllable, then we say that the system is controllable (controlled system).*

Let us have the following Cauchy's problem:

$$\begin{aligned} &\dot{x}(t) = A_0 x(t) + A_1 x(t-\tau) + Bu(t),\ \ t \in [0,T]\,,\ T < \infty, \\ &x(0) = x_0,\ \ \ x(t) = \varphi(t),\ \ \ -\tau \le t < 0, \end{aligned} \tag{1}$$

where $x = \{x_1, ..., x_n\}$ is the vector of phase coordinates, $x \in X_,\ u(t) = \{u_1(t), ..., u_r(t)\}$ is the control function, $u \in U$, U is the set of piecewise-continuous functions; A_0, A_1, B are constant matrices of dimensions $(n \times n)$, $(n \times n)$, $(n \times r)$ respectively, τ is the constant delay.
The state space Z of this system is the set of n-dimensional functions.

$$\{x(\theta),\ t - \tau \le \theta \le t\} \tag{2}$$

The space of the n-dimensional vectors x (phase space X) is a subspace of Z. The initial state z_0 of the system (1) is determined by conditions

$$z_0 = \{x_0(\theta),\ x_0(\theta) = \varphi(\theta),\ -\tau \le \theta < 0,\ x(0) = x_0\}. \tag{3}$$

The state $z = z(z_0, u, t)$ of the system (1) in the space Z in the moment of time t is defined by trajectory segment (2) of phase space X.
Further consider that the movement system (1) goes ($t \ge 0$) in the space of continuous function.
Determining initial state of (3) of the function $\varphi(\theta)$ is piecewise-continuous.
In accordance with specified definitions of state (3) of the system (1) is controllable if there exist such control u, $u \in U$ that $x(t) \equiv 0,\ T - \tau \le t \le T$ when $T < \infty$.
The state (3) of the system (1) is relatively controllable if there exist such control u, $u \in U$, that $x(T) = 0$ when $T < \infty$.

Definition 1.3 *The matrix function which has the form of a polynomial of degree k in intervals $(k-1)\tau \le t \le 0$, "glued" in knots $t = k\tau, k = 0, 1, 2, ...$*

$$e_\tau^{At} = \left\{ \begin{array}{cc} \Theta, & -\infty < t < -\tau \\ I, & -\tau \le t < 0 \\ I + A\frac{t}{1!} + A^2\frac{(t-\tau)^2}{2!} + ... + A^k\frac{(t-(k-1)\tau)^k}{k!}, & (k-1)\tau \le t < k\tau \end{array} \right. .$$

where Θ is the zero matrix, I is the identity matrix, is called the delayed matrix exponential.

Delayed matrix exponentials (continuous or discrete) are used to solve boundary value problems and controllability problems cf. e.g. papers [1] – [4].

2 Representation of solution of the Cauchy problem for homogeneous system

Let us have the system of differential equations with constant delay

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-\tau), \quad x(t) \in R^n, \quad t \ge 0, \tag{4}$$

where $x(t) \equiv \varphi(t)$, if $-\tau \le t \le 0$ is the initial conditions, A_0, A_1 are square matrices with constant elements, $\varphi(t)$ is an arbitrary continuously differentiable initial vector-function.

Lemma 2.1 *Let matrices A_0 and A_1 be commutative, (i.e. $A_0A_1 = A_1A_0$). Then*

$$e^{A_0t}A_1 = A_1e^{A_0t}, \quad t \geq 0.$$

Proof. Follows from the definition (1.3).
Using all the above statements we obtain the explicit form of the fundamental matrix of the system (4) for commutative matrices A_0, A_1.

Theorem 2.2 *Let matrices A_0, A_1 of system (4) be commutative and let there exist A_0^{-1}. Then the matrix*

$$\mathrm{X}_0(t) = Ce^{A_0t}e_\tau^{D(t-\tau)}, \tag{5}$$

where $D = e^{-A_0\tau}A_1, C = (I + A_1A_0^{-1}), t \geq 0$ is the solution of the system (4) satisfying the initial conditions

$$\mathrm{X}_0(t) \equiv I, \quad -\tau \leq t \leq 0. \tag{6}$$

Proof. Performance matrix $\mathrm{X}_0(t)$ of condition (6) follows from the definitions for exponents e^{A_0t} and $e_\tau^{D(t-\tau)}$. We show that when $t \geq 0$ matrix $\mathrm{X}_0(t)$ is a solution of the system (4). After differentiation of (5), we get

$$\begin{aligned}
&(Ce^{A_0t}e_\tau^{D(t-\tau)})_t' = C(A_0(e^{A_0t}e_\tau^{D(t-\tau)}) + e^{A_0t}De_\tau^{D(t-2\tau)}) = \\
&= C(A_0(e^{A_0t}e_\tau^{D(t-\tau)}) + e^{A_0t}e^{-A_0\tau}A_1e_\tau^{D(t-2\tau)}) = C(A_0(e^{A_0t}e_\tau^{D(t-\tau)}) + A_1(e^{A_0(t-\tau)}e_\tau^{D(t-2\tau)})).
\end{aligned}$$

Using the notation (5), we get $\dot{\mathrm{X}}_0(t) = A_0\mathrm{X}_0(t) + A_1\mathrm{X}(t-\tau)$, thus we obtain the statement of theorem.

Theorem 2.3 *Let matrices A_0, A_1 of system (4) be commutative. Then the solution of the Cauchy problem for system (4) with initial conditions $x(t) \equiv \varphi(t), -\tau \leq t \leq 0$ has the form*

$$x(t) = \mathrm{X}_0(t)\varphi(-\tau) + \int_{-\tau}^{0} \mathrm{X}_0(t-\tau-s)\varphi'(s)ds$$

Proof. Solution of the system (4) which satisfies the initial conditions $x(t) \equiv \varphi(t), -\tau \leq t \leq 0$ can be written as

$$x(t) = \mathrm{X}_0(t)c + \int_{-\tau}^{0} \mathrm{X}_0(t-\tau-s)y'(s)ds, \tag{7}$$

where c is the vector of unknown constants, $y(t)$ is an unknown continuously differentiable vector-function and $\mathrm{X}_0(t)$ is the matrix defined in (5). Since the matrix $\mathrm{X}_0(t)$ is a solution of system (4), then for any c and $y(t)$ expression (7) is also a solution of system (4). We choose c and $y(t)$ such that the initial conditions is in the next form

$$x(t) = \mathrm{X}_0(t)c + \int_{-\tau}^{0} \mathrm{X}_0(t-\tau-s)y'(s)ds \equiv \varphi(t).$$

Let put $t = -\tau$. From definition (1.3) and from (5) there follows that

$$\mathrm{X}_0(-\tau) = I, \quad \mathrm{X}_0(-2\tau - s) = \begin{cases} \Theta, -\tau < s \leq 0 \\ I, s = -\tau \end{cases} .$$

So we have $\varphi(-\tau) = c$, and formula (7) takes the form

$$x(t) = \mathrm{X}_0(t)\varphi(-\tau) + \int_{-\tau}^{0} \mathrm{X}_0(t - \tau - s)y'(s)ds.$$

Since $-\tau \leq t \leq 0$, let us break the interval in two parts. We get

$$\varphi(t) = \varphi(-\tau) + \int_{-\tau}^{t} \mathrm{X}_0(t - \tau - s)y'(s)ds + \int_{t}^{0} \mathrm{X}_0(t - \tau - s)y'(s)ds.$$

In the first integral $-\tau \leq s \leq t$, so $-\tau \leq t - \tau - s \leq t$ and late matrix exponential is

$$\mathrm{X}_0(t - \tau - s) \equiv I, \quad -\tau \leq s \leq t.$$

In the second integral $t \leq s \leq 0$, so $t - \tau \leq t - \tau - s \leq -\tau$ and late matrix exponential is

$$\mathrm{X}_0(t - \tau - s) = \begin{cases} \Theta, 0 \leq s < t, \\ I, s = t. \end{cases}$$

Hence in the interval $-\tau \leq t \leq 0$ we get

$$\varphi(-\tau) + \int_{-\tau}^{t} y'(s)ds = \varphi(t). \tag{8}$$

We get

$$\varphi(-\tau) + y(t) - y(-\tau) = \varphi(t). \tag{9}$$

When solving the system of equations (8), (9) we obtain that $y(t) = \varphi(t)$. Substituting this in to (7), we obtain the statement of the theorem.

3 Representation of solution of the Cauchy problem for heterogeneous system

Let us have a linear heterogeneous system with delay

$$\dot{x}(t) = A_0x(t) + A_1x(t - \tau) + f(t), \tag{10}$$

where $f(t)$ is some vector-function, $f(t) = (f_1(t), .., f_n(t))^T$.

Theorem 3.1 *Let matrices A_0, A_1 of system (10) be commutative. Then the solution $\overline{x(t)}$ of the heterogeneous system (10) which satisfies zero initial conditions has the form*

$$\overline{x(t)} = \int_0^t e^{A_0(t-\tau-s)} e_\tau^{D(t-2\tau-s)} f(s)ds, \quad t \geq 0. \tag{11}$$

Proof. Since $\mathrm{X}_0(t)$ is the solution of the homogeneous system (4) then, using the method of variation of an arbitrary constant, the solution $\overline{x(t)}$ of the heterogeneous system is

$$\overline{x(t)} = \int_0^t e^{A_0(t-\tau-s)} e_\tau^{D(t-2\tau-s)} c(s)ds,$$

where $c(s)$, $0 \leq s \leq t$ is an unknown vector-function. When differentiating the expression

$$\dot{\overline{x(t)}} = e^{A_0(t-\tau-s)} e_\tau^{D(t-2\tau-s)} c(s) \mid_{s=t} + \int_0^t \left[A_0 e^{A_0(t-\tau-s)} e_\tau^{D(t-2\tau-s)} + e^{A_0(t-\tau-s)} D e_\tau^{D(t-3\tau-s)}\right] c(s)ds$$

and substituting to (10), we get

$$e^{A_0(-\tau)} e_\tau^{D(-2\tau)} c(t) + \int_0^t \left[A_0 e^{A_0(t-\tau-s)} e_\tau^{D(t-2\tau-s)} + e^{A_0(t-\tau-s)} D e_\tau^{D(t-3\tau-s)}\right] c(s)ds \quad =$$

$$= A_0 \left[\int_0^t e^{A_0(t-\tau-s)} e_\tau^{D(t-\tau-s)} c(s)ds\right] + A_1 \left[\int_0^{t-\tau} e^{A_0(t-2\tau-s)} e_\tau^{D(t-3\tau-s)} c(s)ds\right] + f(t).$$

Hence $e^{A_0(-\tau)} e_\tau^{D(-2\tau)} = I$ and $e^{A_0(t-2\tau-s)} e_\tau^{D(t-3\tau-s)} c(s) = \mathrm{X}_0(t-2\tau-s)$ and we get

$$c(t) + A_1 \int_{t-\tau}^t \mathrm{X}_0(t-2\tau-s)c(s)ds = f(t), \quad \mathrm{X}_0(t-2\tau-s) = \begin{cases} \Theta, t-\tau < s \leq t, \\ I, s = t-\tau, \end{cases}$$

which results in $c(t) = f(t)$. Thus we get (11).

Theorem 3.2 . *The solution of heterogeneous system (10) which satisfies the initial conditions $x(t) \equiv \varphi(t), -\tau \leq t \leq 0$ has the form*

$$x(t) = \mathrm{X}_0(t)\varphi(-\tau) + \int_{-\tau}^0 \mathrm{X}_0(t-\tau-s)\varphi'(s)ds + \int_0^t \mathrm{X}_0(t-\tau-s)f(s)ds.$$

The proof follows from theorems (2.2) and (2.3).

4 The control construction for system with delay with commutative matrices

Let us have the control system of differential equations

$$\dot{x}\,(t) = A_0 x\,(t) + A_1 x(t-\tau) + Bu(t), \tag{12}$$

where $x(t) \in R^n, t \geq 0, \tau\ > 0,\ x(t) \equiv \varphi(t)$ if $-\tau \leq t \leq 0,\ A_0, A_1$ are square matrices, B is the constant matrix of dimension $(n \times n)$, $u(t) = (u_1(t), .., u_n(t))^T$ is the control vector-function.

Theorem 4.1 *For a linear stationary system with delay (12) to be controllable it is necessary that the following condition holds:* $t \geq (k-1)\tau, k = 1, 2, 3, ..$ *and* $\det S_k \neq 0,$ *where*

$$S_k = \{B; e^{-A_0\tau} A_1 B; e^{-2A_0\tau} A_1^2 B; ...; e^{-(k-1)A_0\tau} A_1^{k-1} B\}.$$

Proof. Let the system (12) be controllable. Than for any function $\varphi(t)$, for any time moment t_1 and point $x_1 = x(t_1)$ there exist a control $u^*(t)$ such that for the system (12) there exists solution $x^*(t)$ which satisfies initial conditions $x(t) \equiv \varphi(t), -\tau \leq t \leq 0.$ The representation of the Cauchy problem for heterogeneous equation as the sum is

$$x(t) = \mathrm{X}_0(t)\varphi(-\tau) + \int\limits_{-\tau}^{0} \mathrm{X}_0(t-\tau-s)\varphi'(s)ds + \int\limits_{0}^{t} \mathrm{X}_0(t-\tau-s)Bu(s)ds,$$

where $\mathrm{X}_0(t)$ is defined in (5). If we use the control $u^*(t)$ then in time moment $t = t_1$ we get

$$x(t_1) = x_1 = \mathrm{X}_0(t_1)\varphi(-\tau) + \int\limits_{-\tau}^{0} \mathrm{X}_0(t_1-\tau-s)\varphi'(s)ds + \int\limits_{0}^{t_1} \mathrm{X}_0(t_1-\tau-s)Bu^*(s)ds. \tag{13}$$

Denoted

$$x_1 - \mathrm{X}_0(t_1)\varphi(-\tau) - \int\limits_{-\tau}^{0} \mathrm{X}_0(t_1-\tau-s)\varphi'(s)ds = \mu \tag{14}$$

And using the representation of $\mathrm{X}_0(t)$ from (5) we get

$$\int\limits_{0}^{t} e^{A_0(t_1-\tau-s)} e_{\tau}^{D(t_1-2\tau-s)} Bu^*(s)ds = \int\limits_{0}^{t_1-\tau} e^{A_0\xi} e_{\tau}^{D(\xi-\tau)} Bu^*(t_1-\tau-\xi)d\xi =$$

$$= \int\limits_{-\tau}^{0} e^{A_0\xi} Bu^*(t_1-\tau-\xi)d\xi + \int\limits_{0}^{\tau} e^{A_0\xi}\left[I + D\frac{\xi-\tau}{1!}\right] Bu^*(t_1-\tau-\xi)d\xi +$$

$$\int\limits_{\tau}^{2\tau} e^{A_0\xi}\left[I + D\frac{\xi-\tau}{1!} + D^2\frac{(\xi-2\tau)^2}{2!}\right] Bu^*(t_1-\tau-\xi)d\xi \quad + \ldots$$

$$\int_{(k-2)\tau}^{t_1-\tau} e^{A_0\xi}\left[I+D\frac{\xi-\tau}{1!}+D^2\frac{(\xi-2\tau)^2}{2!}+...+D^{k-2}\frac{[\xi-(k-1)\tau]^{k-1}}{(k-1)!}\right]Bu^*(t_1-\tau-\xi)d\xi+...$$

Denoted $\psi_1(t_1)=\int\limits_{-\tau}^{t_1-\tau} u^*(t_1-\tau-\xi)d\xi;\quad \psi_2(t_1)=\int\limits_{0}^{t_1-\tau}\frac{\xi-\tau}{1!}u^*(t_1-\tau-\xi)d\xi;\quad \ldots;$ $\psi_k(t_1)=\int\limits_{(k-2)\tau}^{t_1-\tau}\frac{[\xi-(k-1)\tau]^{k-1}}{(k-1)!}u^*(t_1-\tau-\xi)d\xi.$ And using (14) correlation (13) we get the form

$$e^{A_0\xi}B\psi_1(t_1)+e^{A_0\xi}DB\psi_2(t_1)+\ldots+e^{A_0\xi}D^{k-1}B\psi_k(t_1)=\mu \tag{15}$$

Since $D=e^{-A_0\tau}A_1$, we rewrite expression (15) as

$$e^{A_0\xi}(B\psi_1(t_1)+e^{-A_0\tau}A_1B\psi_2(t_1)+\ldots+e^{-A_0(k-1)\tau}A_1^{k-1}B\psi_k(t_1))=\mu \tag{16}$$

Since for any matrix A_0, $e^{A_0\xi}\neq\theta$, where θ is zero matrix, and since the system is controllable, we have that (16) has a solution for any vector μ. If $k<n$, then the system is over defined and not always has a solution. Therefore, for controllability it is necessary that $t_1\geq(k-1)\tau\geq(n-1)\tau$. From the Hamilton-Kelly's formula there follows that any power $A_0^n,\ n\geq 2$ of matrix A_0 can be expressed by a linear combination of matrices $I,A_0,A_0^2,...,A_0^{n-1}$. Therefore if $k\geq n$ the system (16) can be substituted

$$e^{A_0\xi}(B\overline{\psi_1(t_1)}+e^{-A_0\tau}A_1B\overline{\psi_2(t_1)}+\ldots+e^{-A_0(k-1)\tau}A_1^{k-1}B\overline{\psi_k(t_1)})=\mu \tag{17}$$

where $\overline{\psi_i(t_1)}$, $i=1,..,n$ are some functions of variable t_1. And if (17) has solution for any μ, then $\det S_k\neq 0$, where $S_k=\{B;e^{-A_0\tau}A_1B;e^{-2A_0\tau}A_1^2B;...;e^{-(k-1)A_0\tau}A_1^{k-1}B\}$.

Theorem 4.2 *System (12) is controllable for* $t\geq(k-1)\tau, k=1,2,3..$ *if holds:* $\det Q_k\neq 0$, $Q_k=\{B;[A_0e^{A_0\tau}e_\tau^{D\tau}+A_1]B;...;[A_0^{k-1}e^{A_0(k-1)\tau}e_\tau^{D(k-1)\tau}+A_1^{k-1}]B\}$, *where* $D=e^{-A_0\tau}A_1$.

Proof. Let the initial condition be zero: $\varphi(t)\equiv 0, -\tau\leq t\leq 0$, $t_1\geq(k-1)\tau$ and $\det S_k\neq 0$. We can show that system (12) is controllable. First, we show that the attainable Q_φ has dimension n. Let the opposite holds: the dimension of Q_φ be less than n. Then there is the constant vector $y\in R^n$, so that for any $u(t)\in\Omega_1(u)$ will be

$$y^Tx(t_1)=0 \tag{18}$$

Since the initial function $\varphi(t)$ is zero, the condition (18) has the form

$$y^T\int_0^{t_1} e^{A_0(t_1-\tau-s)}e_\tau^{D(t_1-2\tau-s)}Bu(s)ds=0 \tag{19}$$

Since (19) holds for arbitrary function $u(t)\in\Omega_1(u)$, the latter equality holds if and only if (the main lemma of variational calculus) $y^Te^{A_0(t_1-\tau-s)}e_\tau^{D(t_1-2\tau-s)}Bu(s)\equiv 0,\quad 0\leq s\leq t_1$.
Since $t_1\geq(n-1)\tau$, then after $(k-1)$-times differentiated this identity, we obtain

$$y^T[A_0e^{A_0(t_1-\tau-s)}e_\tau^{D(t_1-2\tau-s)}+e^{A_0(t_1-\tau-s)}De_\tau^{D(t_1-3\tau-s)}]Bu(s)\equiv 0,\ldots,$$

$$y^T[A_0^{k-1}e^{A_0(t_1-\tau-s)}e_\tau^{D(t_1-2\tau-s)}+e^{A_0(t_1-\tau-s)}D^{k-1}e_\tau^{D(t_1-(k+1)\tau-s)}]Bu(s)\equiv 0$$

Since $D=e^{-A_0\tau}A_1$ and $A_0A_1=A_1A_0$, the last $(k-1)$ identity can be rewritten as

$$y^T[A_0e_\tau^{D(t_1-2\tau-s)}+e^{-A_0\tau}A_1e_\tau^{D(t_1-3\tau-s)}]Bu(s)\equiv 0,\ldots,$$

$$y^T[A_0^{k-1}e_\tau^{D(t_1-2\tau-s)}+e^{-A_0(n-1)\tau}A_1^{k-1}e_\tau^{D(t_1-k\tau-s)}]Bu(s)\equiv 0.$$

Denote $s=t_1-2\tau$, $s=t_1-3\tau$, $\ldots$, $s=t_1-(k+1)\tau$. We get a system of k equations

$$y^TB=0;y^Te^{A_0\tau}[A_0e_\tau^{D\tau}+e^{-A_0\tau}A_1]B=0;\ldots;y^Te^{A_0(k-1)\tau}[A_0^{k-1}e_\tau^{D(k-1)\tau}+e^{-A_0(k-1)\tau}A_1^{k-1}]B=0,$$

or $\quad y^TB=0;y^T[A_0e^{A_0\tau}e_\tau^{D\tau}+A_1]B=0;\ldots;y^T[A_0^{k-1}e^{A_0(k-1)}e_\tau^{D(k-1)\tau}+A_1^{k-1}]B=0.$
A homogeneous system has nontrivial solution if and only if its determinant equals zero. Thus we get $\det\{B;[A_0e^{A_0\tau}e_\tau^{D\tau}+A_1]B;\ldots;[A_0^{k-1}e^{A_0(k-1)}e_\tau^{D(k-1)\tau}+A_1^{k-1}]B\}=0$, which contradicts the condition. Thus the assumption that the dimension of Q_φ is less than n must be false.

Theorem 4.3 *Let $t_1\geq(k-1)\tau, k=1,2,3..$ and the necessary and sufficient conditions for controllability be implemented: $detS_k=det\{B;e^{-A_0\tau}A_1B;e^{-2A_0\tau}A_1^2B;...;e^{-(k-1)A_0\tau}A_1^{k-1}B\}\neq 0$, $detQ_k=\det\{B;[A_0e^{A_0\tau}e_\tau^{D\tau}+A_1]B;\ldots;[A_0^{k-1}e^{A_0(k-1)\tau}e_\tau^{D(k-1)\tau}+A_1^{k-1}]B\}\neq 0$, where $D=e^{-A_0\tau}A_1$. Then the control function can be taken as*

$$u(s)=[\mathrm{X}_0(t_1-\tau-s)B]^T\left[\int_0^{t_1}\mathrm{X}_0(t_1-\tau-s)BB^T[\mathrm{X}_0(t_1-\tau-s)]^Tds\right]^{-1}\mu, \tag{20}$$

where $\mu=x_1-\mathrm{X}_0(t_1)\varphi(-\tau)-\int\limits_{-\tau}^{0}\mathrm{X}_0(t_1-\tau-s)\varphi'(s)ds.$

Proof. Using the Cauchy integral representation, we have that the solution of the system (12) with initial conditions $x_0(t)\equiv\varphi(t)$, $-\tau\leq t\leq 0$ has the form

$$x(t)=\mathrm{X}_0(t)\varphi(-\tau)+\int_{-\tau}^{0}\mathrm{X}_0(t-\tau-s)\varphi'(s)ds+\int_0^t\mathrm{X}_0(t-\tau-s)Bu(s)ds \tag{21}$$

Using the notations (14) we obtain: the system (21) had a solution $x(t)$ that satisfied the initial conditions $x(t)\equiv\varphi(t)$, $-\tau\leq t\leq 0$, $x(t_1)=x_1$, is necessary and sufficient that the integrated equation

$$\int_0^{t_1}e^{A_0(t_1-\tau-s)}e_\tau^{D(t_1-2\tau-s)}Bu(s)ds=\mu \tag{22}$$

has solution $u(s)$, $0\leq s\leq t_1$. We will search a solution as a linear combination

$$u(s)=[e^{A_0(t_1-\tau-s)}e_\tau^{D(t_1-2\tau-s)}B]^TC \tag{23}$$

where $C = (c_1, c_2, ..., c_n)^T$ is unknown vector. After substituting (23) into the system (22), we get

$$\left[\int_0^{t_1} e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-2\tau-s)} BB^T [e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-\tau-s)}]^T ds\right] C = \mu \tag{24}$$

We show that system (24) has only one solution. Since $t_1 \geq (k-1)\tau$ using the Kelly's formula vector $e^{A_0(t-\tau-s)} e_\tau^{D(t-2\tau-s)} B$ for any fixed $0 \leq s \leq t_1$ can be represented as a linear combination $B; e^{-A_0\tau} A_1 B; e^{-2A_0\tau} A_1^2 B; ...; e^{-(k-1)A_0\tau} A_1^{k-1} B$. Since the vectors are linearly independent, when $0 \leq s \leq t_1$ there will be $e^{A_0(t-\tau-s)} e_\tau^{D(t-2\tau-s)} B \neq 0$. Therefore for any vector $l = (l_1, l_2, ..., l_n)^T$ in $0 \leq s \leq t_1$ there will be $([e^{A_0(t-\tau-s)} e_\tau^{D(t-2\tau-s)} B]^T l)^2 \neq 0, 0 \leq s \leq t_1$. And for any $l > 0$

$$\int_0^{t_1} ([e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-2\tau-s)} B]^T l)^2 ds = \left[\int_0^{t_1} e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-2\tau-s)} BB^T [e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-2\tau-s)}]^T ds\right] l^2,$$

or the matrix $\left[\int_0^{t_1} e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-2\tau-s)} BB^T [e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-2\tau-s)}]^T ds\right]$ is positive defined. Hence its determinant is nonzero. Solving system (24), we obtain

$$C = \left[\int_0^{t_1} e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-2\tau-s)} BB^T [e^{A_0(t_1-\tau-s)} e_\tau^{D(t_1-2\tau-s)}]^T ds\right]^{-1} \mu.$$

5 Example

Example 1. Let us have the differential equation of 2^{nd} degree with a constant delay:

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-1) + Bu(t), \text{ where } A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, A_1 = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

As we see $\tau = 1, n = 2$ and $A_0 A_1 = A_1 A_0$. We want to know whether this system is controllable in the moment of time $t_1 = 3$. Let us check the necessary condition. First, we find the matrix $S_3 = \{B, e^{-A_0\tau} A_1 B, e^{-2A_0\tau} A_1^2 B\} = \{B, (e^{-1} I) A_1 B, (e^{-2} I) A_1^2 B\} =$

$$= \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} e^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} e^{-2} & 0 \\ 0 & 0 \end{pmatrix} \right\}$$

We have $rank(S_3) < 2$ so the system is not controllable for $t_1 = 3$.

Example 2. Let us have the differential equation of 2^{nd} degree with a constant delay:

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-1) + Bu(t), \text{ where } A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, A_1 = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

As we see $\tau = 1, n = 2$ and $A_0 A_1 = A_1 A_0$. We want to know whether this system is controllable in the moment of time $t_1 = 2$. Let us check the necessary and sufficient conditions:

$$rank(S_2) = rank(\{B, e^{-A_0\tau} A_1 B\}) = 2, \quad rank(Q_2) = rank(\{B, [A_0 e^{A_0\tau} e_\tau^{D\tau} + A_1] B\}) = 2.$$

It is easy to see that the necessary and sufficient conditions for controllability is implemented ($rank(B) = 2$), so the system is controllable in time moment $t_1 = 2$.
Let us use the solution of the Cauchy problem in general case (21). If $u(t) = 0$, we have that for $1 \le t \le 2, x(t) = 0$, so $x(2) = 0$. Let us construct such control function, that system in time moment $t_1 = 2$ be in point $x_1 = (1,1)^T$, using initial condition $x_0(t) \equiv \varphi(t) = (0,0)^T, -\tau \le t \le 0$. Using the result of the theorem (4.3) we write:

$$u(t) = [e^{A_0(t_1-\tau-t)}e_\tau^{D(t_1-2\tau-t)}B]^T[\int_0^{t_1} e^{A_0(t_1-\tau-s)}e_\tau^{D(t_1-2\tau-s)}BB^T[e^{A_0(t_1-\tau-t)}e_\tau^{D(t_1-2\tau-t}]^T ds]^{-1}\mu,$$

$$\mu = x_1 - e^{A_0(t_1)}e_\tau^{D(t_1)}\varphi(-\tau) - \int_{-\tau}^{0} e^{A_0(t_1-\tau-s)}e_\tau^{D(t_1-2\tau-s)}\varphi'(s)ds$$

So, we have $u(t) = [e^{(1-t)}e_1^{D(-t)}]^T[\int_0^2 e^{(1-s)}e_1^{D(-s)}[e^{(1-s)}e_1^{D(-s)}]^T ds]^{-1}\mu, \mu = (1,1)^T - (e^2 I)e_1^{2D}(0,0)^T - \int_{-1}^{0} e^{(1-s)}(e_1^{D(-s)I)}(0,0)^T ds = (1,1)^T$.

Finally, we get $u(t) = e^{-t}(0.7e^3 - 0.7e^2t + 0.35et^2, \quad -0.1e^3 - 1.1e^2t + 2.35et^2)^T$

And for this control function we have $x(3) = (1.017, 0.986)^T$.

Acknowledgement

This research was supported by the Grant 201/10/1032, 201/08/9469 of Czech Grant Agency and by the Council of Czech Government MSM 0021630529 and by Grant FEKT-S-11-2-921 of Faculty of Electrical Engineering and Communication, BUT.

References

[1] A. Boichuk, J. Diblík, D.Ya. Khusainov, M. Růžičková: *Boundary Value Problems for Delay Differential Systems*, Advances in Difference Equations, vol. 2010, Article ID 593834, 20 pages, 2010. doi:10.1155/2010/593834

[2] A. Boichuk, J. Diblík, D. Khusainov, M. Růžičková: *Fredholm's boundary-value problems for differential systems with a single delay,* Nonlinear Analysis, **72** (2010), 2251–2258. (ISSN 0362-546X)

[3] J. Diblík, D. Khusainov, J. Lukáčová, M. Růžičková: *Control of oscillating systems with a single delay*, Advances in Difference Equations, Volume 2010 (2010), Article ID 108218, 15 pages, doi:10.1155/2010/108218.

[4] J. Diblík, D. Khusainov, M. Růžičková: *Controllability of linear discrete systems with constant coefficients and pure delay,* SIAM Journal on Control and Optimization, **47**, No 3 (2008), 1140–1149. DOI: 10.1137/070689085, url = http://link.aip.org/link/?SJC/47/1140/1. (ISSN Electronic: 1095-7138, Print: 0363-0129)

[5] KHUSAINOV D.Ya., IVANOV A.F., SHUKLIN G.V..: *About a single submission of solution of linear systems with delay.* Differential equations. 2005. No.41,7. 1001-1004. (In Russian)

[6] KHUSAINOV D.Ya., SHUKLIN G.V.: *About relative controllability of systems with pure delay.* Applied Mechanics. 2005. No.41,2. 118-130. (In Russian)

[7] KOLMANOVSKII V.B., NOSOV V.R.: *Stability and periodical regimes of controlled systems with aftereffect*. Nauka. 1981. 448Pp. (In Russian)

[8] KOLMANOVSKII V.B.: *About stability of nonlinear system with delay*. Mathematical Notes Ural University. 1970. 743-751. (In Russian)

Current address

BAŠTINEC Jaromír, doc., RNDr., CSc.
Department of Mathematics,
Faculty of Electrical Engineering and Communication,
Brno University of Technology,
Technická 8, 616 00 Brno, Czech Republic
telefon: +420 541 143 222
email: bastinec@feec.vutbr.cz

PIDDUBNA Ganna, Mgr.
Department of Mathematics,
Faculty of Electrical Engineering and Communication,
Brno University of Technology,
Technická 8, 616 00 Brno, Czech Republic
telefon: +420 541 143 222
email: bastinec@feec.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

ON STABILITY ANALYSIS OF COCYCLES OVER IMPULSE MARKOV DYNAMICAL SYSTEMS

CARKOVS Jevgeņijs, (LV), LIEPA Evija, (LV)

Abstract. This paper deals with Cauchy matrix family of a linear differential equation with right part dependent on a step Markov process and an impulse type dynamical system switched by the above process. All the above mentioned stochastic dynamical objects are also dependent on small positive parameter ε, and the infinitesimal operator of Markov process is proportional to ε^{-1}. This means that impulse dynamical system is rapidly switched and one may simplify this using merger procedur to Markov process and averaging procedures to impulse dynamical system and matrix evolution family. Applying these procedures one achieves more simple linear differential equation for matrix evolution family, which becomes now dependent on more simple dynamical systems such as an ordinary differential equation with a right part switched by a lumped Markov process. It is proved that under some hypotheses one may successfully apply these resulting evolution families not only to approximation of the initial family on an arbitrary finite time interval but also to describe a time asymptotic of it.

Key words and phrases. Cocycles, Random Evolutions, Impulse Stochastic Equations, Stochastic Stability, Averaging Procedures.

Mathematics Subject Classification. Primary 60H10, 60H30; Secondary 37H10.

1 Introduction

As it has been metioned in abstract our mathematical model consists of: a right continuous *Step Markov Process* (**SMP**) $\{y^\varepsilon(t),\, t \geq 0\}$ with switching times $\mathbf{S} := \{\tau_k^\varepsilon,\, k \in \mathbb{N}\}$ given on discrete metric space $\mathbb{Y}$ by a weak infinitesimal operator [2] $Q^\varepsilon v(y) := \frac{1}{\varepsilon}Q_1 v(y) + Q_2 v(y)$, where operator Q_1 has 0 is a simple spectrum point of multiplicity d, ε is small positive parameter, $Q_j v(y) = a(y) \sum\limits_{z\in\mathbb{Y}} [v(z) - v(y)] p_j(y,z),\ j = 1,2$ and $v(y)$ is an arbitrary bounded measurable mapping $\mathbb{Y} \to \mathbb{R}$;

Impulse Markov Dynamical System (**IMDS**)[8] switched by the above **SMP**, and given as right continuous m-dimensional vector-function $\{x^\varepsilon(t),\ t \geq 0\}$ satisfying

- a differential equation for $t \notin \mathbb{S}$

$$\frac{dx^\varepsilon}{dt} = f(x^\varepsilon(t),\ y^\varepsilon(t),\ \varepsilon), \tag{1}$$

- a jump condition for $t \in \mathbb{S}$

$$x^\varepsilon(t) = x^\varepsilon(t-0) + \varepsilon g(x^\varepsilon(t-0), y^\varepsilon(t-0), \varepsilon); \tag{2}$$

*Markov Evolution Family (***MEF***) or Markov Multiplicative Cocycle* [5, 6] given as a two parametric Cauchy matrix family $\{X^\varepsilon(t,s),\ t \geq s \geq 0\}$ satisfying a linear differential equation in $\mathbb{R}^n$:

$$\frac{d}{dt} X^\varepsilon(t,s) = A(x^\varepsilon(t),\ y^\varepsilon(t),\ \varepsilon)\, X^\varepsilon(t,s). \tag{3}$$

The problem of asymptotic analysis of dynamical systems with random switching has been discussed in many mathematical and engineering papers. Apparently, A. V. Skorokhod was the first mathematician to have proved that the probabilistic limit theorems may be successfully used for differential equations with right parts dependent on step Markov process (English edition [7]). The approach proposed by A.V. Skorokhod and developed by many authors (see, for example, [1] and references there) makes it possible under assumption $d = 1$ to apply for asymptotic analysis of **IMDS** (1)–(2)–(3) not only the averaging procedure by time and invariant measure of **SMP** but also diffusion approximation technique. Some of the above mentioned results have been published also in our previous paper. In this paper we will deal with $d > 1$ and will discuss an ability of proposed by V.S.Krolyuk [4] so called *merger procedure* involving averaging procedure by all invariant measures of operator Q_1^*.

2 Assumptions and notations.

To achieve the limiting **MEF** for (1)-(2)-(3) this paper assumes that:

(i) $\forall\, y \in \mathbb{Y}\ : 0 < \hat{a}_1 \leq a(y) \leq \hat{a}_2 < \infty$;

(ii) $\forall\, y, z \in \mathbb{Y}\ |p_2(y,z)| \leq c < \infty$, $p_1(y,z) \geq 0$, $\sum\limits_{z \in \mathbb{Y}} p_1(y,z) = 1$;

(iii) 0 is a simple spectrum point of operator Q_1 of multiplicity d and

$$\exists \rho > 0 :\ \sigma(Q_1) \setminus \{0\} \subset \{z \in \mathbf{C} : \Re z < -\rho\};$$

(iv) $f(x,y,\varepsilon) = f_1(x,y) + \varepsilon f_2(x,y)$, $g(x,y,\varepsilon) = g_1(x,y) + \varepsilon g_2(x,y)$, and $f_j(x,y),\ g_j(x,y)\ j = 1,2$ are boundedly (on x and y) continuously differentiable on x functions;

(v) $A(x,y,\varepsilon) = A_1(x,y) + \varepsilon\, A_2(x,y)$ and matrices $A_j(x,y),\ j = 1,2$ are bounded and continuous on x.

In our presentation we will use the following notations and definitions:

- $F_j(x,y) = f_j(x,y) + a(y)g_j(x,y),\ j = 1,2;$
- μ_j – the probabilistic measures with nonintersecting supports $\mathbb{Y}_j$ defined as the solutions of the equation $Q_1^*\mu_j = 0$, where $(Q_1^*\mu_j)(y) := \sum\limits_{z\in\mathbb{Y}_j} a(z)p_1(z,y)\mu_j(z) - a(y)\mu_j(y),\ 1 \le j \le d;$
- $\mathbf{P}_0$ – projective operator in a kernel of Q_1: $1 \le j \le d,\ y \in \mathbb{Y}_j : (\mathbf{P}_0 v)(y) := \sum\limits_{y\in\mathbb{Y}_j} \mu_j(y)\, v(y);$

  Π - an extension of potential: $(\Pi v)(y) := \int\limits_0^\infty \sum\limits_{z\in\mathbb{Y}} P(t,y,z)[v(z) - (\mathbf{P}_0 v)(z)]dt$, where $P(t,y,z)$ is transition probability of a Markov process corresponding to the infinitesimal operator Q_1;
- $\lambda_p(\varepsilon) := \lim\limits_{t\to\infty} \sup\limits_{x,y} \frac{1}{pt} \ln \mathbb{E}^s_{x,y}\{||X^\varepsilon(t,s)||^p\}$– Lyapunov p-index.

Our paper has for an object to analyse asymptotic behaviuor of **MEF** with $t \to \infty$. **MEF** is said to be asymptotic decreasing with probability one if

$$\lim_{T\to\infty} \mathbb{P}\left\{\sup_{t\ge T} ||X^\varepsilon(t,s)|| \ge \delta/x(s) = x, y(s) = y\right\} = 0$$

for any $\delta > 0$, $x \in \mathbb{R}^n$, $y \in \mathbb{Y}$. Here and elsewhere further probability or expectation with indices denote conditional ones, that is, in the above formula one should read $\mathbb{P}\{\bullet/x(s) = x, y(s) = y\}$ instead of $\mathbb{P}^s_{x,y}\{\bullet\}$.

3 Exponential decreasing of MEF.

In this section we will prove that for analysis of **MEF** behaviour as $t \to \infty$ one may use more simple **MEF**

$$\frac{d\hat{x}(t)}{dt} = \hat{F}_1(\hat{x}(t), \hat{y}(t)) \tag{4}$$

where $\{\hat{y}(t)\}$ is homogeneous Markov process (enlardged MP) with state space $\hat{Y} := \{Y_1, Y_2, ..., Y_d\}$, infinitesimal matrix $\Gamma := \{\gamma_k^j\}$, which elements are given by equalities

$$\gamma_k^j = \begin{cases} \sum\limits_{y\in Y_k} \sum\limits_{z\in Y_j} a(y)p_2(y,z)\mu_k(y), & \text{if} \quad j \ne k; \\ -\sum\limits_{\substack{l=1 \\ l\ne j}}^{d} \gamma_l^j, & \text{if} \quad j = k \end{cases}$$

for $k = 1, 2, ..., d$, and defined by this infinitesimal matrix transition probability function $P_0(t,y,z)$. Due to assumption on spectrum structure of the operator Q_1 one can define [2] the projective operator $\mathcal{P}$ by the equalities

$$\forall\ y \in\ Y_k,\ v \in \mathbb{B}(Y) :\ (\mathcal{P}v)(y) \equiv \sum_{z\in Y_k} v(z)\mu_k(z)$$

for each $k = \overline{1, d}$ and the linear continuous operator $\hat{\Pi} : \mathbb{B}(\hat{Y}) \to \mathbb{B}(\hat{Y})$ by equality

$$(\hat{\Pi}v)(y) := \int_0^\infty \sum_{z \in \hat{Y}} P_0(t, y, z)(v - \mathcal{P}v)(z)dt \tag{5}$$

We will refere to operator (5) as *a potential of enlarged Markov process.*

Theorem 3.1 *(***Merger principle** [8]*). Under the above assumptions the family of processes* $\{x^\varepsilon(s), 0 \le s \le T\}$ *for any* $T > 0$ *weak converges as* $\varepsilon \to 0$ *to the solution of* (4) *with corresponding initial condition.*

Theorem 3.2 *Under the above assumptions if defined by* (4) *evolution family exponentially decrease in the mean with power p then there exists such* $\varepsilon_p > 0$ *that initial* **MEF** *exponentially decrease in the mean with power p for any* $\varepsilon \in (0, \varepsilon_p)$.

Proof. Due to exponential decrease of the p-moments of the solutions of (4) and a boundedness of the x-derivative of $\tilde{F}_1(x, y)$ one can define function

$$y \in Y_k : \; v^{(p)}(x, y) \equiv \hat{v}^{(p)}(x, k) := \int_0^T \mathbb{E}_{x,k} |\hat{x}(t)|^p \, dt, \;\; k = \overline{1, h},$$

with so large constant T that the above function satisfies the inequalities $m_1 |x|^p \le v^{(p)}(x, y) \le m_2 |x|^p$ with some positive constants m_1, m_2, and the inequality

$$(\hat{F}_1(x, k), \, \triangledown) \, \hat{v}^{(p)}(x, k) + \Gamma \hat{v}^{(p)}(x, k) \le -m_3 \, \hat{v}^{(p)}(x, k)$$

is hold with some positive constant m_3 for any $k = \overline{1, h}$ and $x \in \mathbb{R}^n$. To prove the theorem we will use the Lyapunov function

$$v_\varepsilon^{(p)}(x, \, y) := v^{(p)}(x, y) + \varepsilon \tilde{\Pi} \{ F_1(x, \, y), \, \triangledown) \, v^{(p)}(x, y) + Q_1 \, v^{(p)}(x, y) \},$$

which satisfies the inequalities $\hat{m}_1 |x|^p \le v_\varepsilon^{(p)}(x, \, y) \le \hat{m}_2 |x|^p$ with some positive constants $\hat{m}_1$, $\hat{m}_2$ for any $\varepsilon \in (0, 1)$. By definition of the operator $\tilde{\Pi}$ one can write the equality

$$\begin{aligned}
&(F_1(x, y), \, \triangledown) \, v^{(p)}(x, y) + Q_1 \, v^{(p)}(x, y)\} + Q_0 v_1^{(p)}(x, y) \\
&= (\tilde{F}_1(x, y), \, \triangledown) \, v^{(p)}(x, y) + \mathcal{P} Q_1 \, v^{(p)}(x, y)\} + \varepsilon r(x, y, \varepsilon) \\
&= (\hat{F}_1(x, k), \, \triangledown) \, \hat{v}^{(p)}(x, k) + \Gamma \, \hat{v}^{(p)}(x, k)\} + \varepsilon r(x, y, \varepsilon) \\
&\qquad \le -m_3 \hat{v}^{(p)}(x, k) + \varepsilon \alpha(\varepsilon) |x|^p
\end{aligned}$$

and therefore

$$(F_1(x, y), \, \triangledown) \, v^{(p)}(x, y) + Q_1 \, v^{(p)}(x, y)\} + Q_0 v_1^{(p)}(x, y) \le -m_3 \hat{v}^{(p)}(x, k) + \varepsilon \alpha(\varepsilon) |x|^p$$

for any $y \in Y_k$ and $k = \overline{1, d}$, where $\alpha(\varepsilon)$ is infinitesimal as $\varepsilon \to 0$. Due to the above inequalities there exists such positive constants ε_p that

$$\mathbb{L}(\varepsilon)\, v_\varepsilon^{(p)}(x,\, y) \leq -\frac{m_3}{2}\, v_\varepsilon^{(p)}(x,\, y)$$

for any $\varepsilon \in (0,\, \varepsilon_p)$. Using Dynkin formula [2] for the stochastic process

$$\xi(s) := v_\varepsilon^{(p)}(x^\varepsilon(s),\, y(s/\varepsilon))\, e^{\frac{m_3}{4} s}$$

one can get the inequalities

$$e^{\frac{m_3}{4} s}\, \mathbb{E}_{x,y}\, |x^\varepsilon(s)|^p \leq v_\varepsilon^{(p)}(x,\, y) \leq \hat{m}_3\, |x|^p,$$

for any $s \geq 0$ and proof is completed.

Note. By using the supermartingale property of the above defined stochastic process $\xi(s)$ one can make sure that *under the conditions of the Theorem 7 the trivial solution of the* (3) *is asymptotically stochastically stable for all sufficiently small positive* ε.

Acknowledgement

This paper is based upon work supported by the Latvian Scientific Council under grant No.09.1220.

References

[1] B.V. Anisimov, Random Processes with Discrete Component.Limit Theorems (Russian)., Kiev Univ., Kiev, 1988.

[2] E.B. Dynkin, Markov Processes, Springer-Verlag, Berlin, 1965.

[3] L.Katafygiotis and Ye.Tsarkov (J.Carkovs), *Mean square stability of linear dynamical systems with small Markov perturbations.I. Bounded coefficients*, Random Oper. and Stoch. Equ., 4 (1996), pp. 149–170

[4] V.S. Korolyuk and A.F. Turbin, *Limit theorems for Markov random evolutionsin the schemeof asymptotic phase lamping* in Lect. Not. Math., 1021 (1983), pp. 83–88.

[5] V.S. Korolyuk and A. V. Swishchyuk, Semi-Markov Random Evolution, Kluwer Academic Publisher, Holland, 1994.

[6] V.I. Oseledec, *A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems*, in Transactions of Moscow Mathematical Society, 19 (1968), pp. 197–231.

[7] A.V. Skorokhod, Asymptotic Methods of the Theory of Stochastic Differentia Equations, AMS, Providence, RI, 1989.

[8] Ye.Tsarkov (J.Carkovs, *Asymptotic methods for stability analysis Markov impulse dynamical systems*, in Advances of Stability Theory of the End of XXth Century. Stability and Control: Theory, Methods and Applications, Gordon and Breach Science Publishers, London, 2000. p.p. 251–264.

Current address

Jevgeņijs Carkovs, professor
Riga Technical University,
Kaļķu 1, Riga, Latvia, Riga LV-1658
phone: +371 67089517
e-mail:carkovs@latnet.lv

Evija Liepa, asoc. professor
Baltic International Academy
Lomonosova street 4, Riga, Latvia, LV-1003
phone: +371 67100610
e-mail: evija.liepa@bsa.edu.lv

Journal of Applied Mathematics

volume IV (2011), number I

A FUNCTIONAL-INTEGRAL MODEL FROM ECONOMIC DYNAMICS

CIPLEA Sorina Anamaria, (RO), LUNGU Nicolaie, (RO)

Abstract. In this paper we investigate some new applications of the Gronwall lemma to Ulam stability of Volterra and functional Volterra integral equations. In this case we consider two models for the price fluctuations in a single commodity market. First by a integral equation and second by a functional integral equation. Two types of Ulam stability are considered: Ulam-Hyers and generalized Ulam-Hyers-Rassias stability.

Key words and phrases. Volterra integral equations, functional Volterra integral equations, Ulam-Hyers stability, generalized Ulam-Hyers-Rassias stability, Gronwall lemmas, economic dynamics, integral equation

Mathematics Subject Classification. 45G10, 45M10, 45N05, 47H10, 47H20

1 Introduction

In this paper we investigate some new results of Ulam stability for Volterra integral equation and functional Volterra integral equation. In this case we present two types of Ulam stability for Volterra and functional Volterra integral equations: Ulam-Hyers and generalized Ulam-Hyers-Rassias stability. This results are based on the papers [1], [2], [3], [5], [7], [8], [9], [10].

2 Ulam-Hyers-Rassias stability of Volterra integral equation

In what follows we consider the integral equation

$$u(x,y) = h(x,y) + \int_0^x \int_0^y f(x,y,s,t,u(s,t),g(u(s,t)))dsdt, \quad x,y,s,t \in [0,a) \tag{1}$$

$(\mathbb{B}, |\cdot|)$ a (real or complex) Banach space, $f \in C([0,a)^4 \times \mathbb{B}^2, \mathbb{B})$, $h \in C([0,a)^2, \mathbb{B})$, $g \in C([0,a)^2 \times C([0,a)^2))$ and $a \in (0,\infty]$.

2.1 Ulam-Hyers stability

First we consider Ulam-Hyers stability for the equation (1). Let $a \in (0,\infty]$ and $\varepsilon > 0$.

We consider the inequation:

$$\left| u(x,y) - h(x,y) - \int_0^x \int_0^y f(x,y,s,t,u(s,t),g(v(s,t)))dsdt \right| \le \varepsilon, \; x,y,s,t \in [0,a). \tag{2}$$

Definition 2.1 ([7]) The equation (1) is Ulam-Hyers stable if there exists a real number $c_{f,g} > 0$ such that for each solution $u(x,y)$ of (2) there exists a solution $u^*(x,y)$ of (1) with

$$|u(x,y) - u^*(x,y)| \le c_f \cdot \varepsilon \tag{3}$$

Theorem 2.2 *If we have:*

(i) $f \in C([0,a)^4 \times \mathbb{B}^2, \mathbb{B})$, $h \in C([0,a)^2, \mathbb{B})$, $g \in C([0,a)^2 \times C([0,a)^2))$;

(ii) there exist $L_1, L_2 > 0$ *such that*

$$|f(x,y,s,t,u,v) - f(x,y,s,t,\overline{u},\overline{v})| \le L_1|u-\overline{u}| + L_2|v-\overline{v}|, \; \forall\ s,t,x,y \in [0,a], \; u,v,\overline{u},\overline{v} \in \mathbb{B};$$

(iii) there exists $L_3 > 0$ *such that*

$$|g(u) - g(v)| \le L_3|u-v|, \text{ for all } u,v \in \mathbb{B}.$$

Then:

(a) The equation (1) has in $C([0,a],\mathbb{B})$ *a unique solution* u^**;*

(b) For each $\varepsilon > 0$*, if* $u \in C([0,a],\mathbb{B})$ *is a solution of the inequation (2), then*

$$|u(x,y) - u^*(x,y)| \le c_{f,g} \cdot \varepsilon$$

where

$$c_{f,g} = \exp(L_1 + L_2L_3)a^2 \tag{4}$$

hence, the equation (1) is Ulam-Hyers stable.

Proof.(a) This is a known result (see for example [11]).

Let us prove (b). We have

$$|u(x,y) - u^*(x,y)| \le \left| u(x,y) - h(x,y) - \int_0^x \int_0^y f(x,y,s,t,u(s,t),g(u(s,t)))dsdt \right|$$

$$+ \int_0^x \int_0^y |f(x,y,s,t,u(s,t),g(u(s,t))) - f(x,y,s,t,u^*(s,t),g(u^*(s,t)))|dsdt$$

$$\leq \varepsilon + (L_1 + L_2 L_3) \int_0^x \int_0^y |u(s,t) - u^*(s,t)| dsdt.$$

From the Gronwall lemma ([8]), we have

$$|u(x,y) - u^*(x,y)| \leq \varepsilon \exp(L_1 + L_2 L_3) a^2 = c_{f,g} \cdot \varepsilon \tag{5}$$

where

$$c_{f,g} = \exp(L_1 + L_2 L_3) a^2.$$

So, the equation (1) is Ulam-Hyers stable.

2.2 Generalized Ulam-Hyers-Rassias stability of equation (1)

In what follows we have a stability result of Ulam-Hyers-Rassias type for the equation (1) ([7], [8], [9]). Let $a \in (0, \infty]$ and $\varphi \in C([0,a] \times [0,a], \mathbb{R}_+)$. We consider the inequation:

$$\left| u(x,y) - h(x,y) - \int_0^x \int_0^y f(x,y,s,t,u(s,t), g(u(s,t))) dsdt \right| \leq \varphi(x,y), \ \forall \ x,y \in [0,a). \tag{6}$$

Definition 2.3 ([7]) The equation (1) is generalized Ulam-Hyers-Rassias stable if there exists a real number $c_{f,\varphi} > 0$ such that for each solution $u(x,y)$ of (6) there exists a solution $u^*(x,y)$ of (1) with

$$|u(x,y) - u^*(x,y)| \leq c_{f,g} \varphi(x,y), \ \forall \ x,y \in [0,a). \tag{7}$$

Theorem 2.4 *We suppose that*

(i) $f \in C([0,a)^4 \times \mathbb{B}^2, \mathbb{B})$, $h \in C([0,a)^2, \mathbb{B})$, $g \in C([0,a)^2 \times C([0,a)^2))$;

(ii) there exist $L_1, L_2 > 0$ such that

$$|f(x,y,s,t,u,v) - f(x,y,s,t,\overline{u},\overline{v})| \leq L_1 |u - \overline{u}| + L_2 |v - \overline{v}|, \ \forall \ s,t,x,y \in [0,a], \ u,v,\overline{u},\overline{v} \in \mathbb{B};$$

(iii) there exists $L_3 > 0$ such that

$$|g(u) - g(v)| \leq L_3 |u - v|, \ \textit{for all } u,v \in \mathbb{B};$$

(iv) the function φ is an increasing function.

Then:

(a) The equation (1) has in $C([0,a] \times [0,a], \mathbb{B})$ a unique solution $u^*(x,y)$;

(b) If $u \in C([0,a] \times [0,a], \mathbb{B})$ is a solution of the inequation (6) then

$$|u(x,y) - u^*(x,y)| \leq \varphi(x,y) \cdot c_{f,g} \tag{8}$$

where

$$c_{f,g} = \exp((L_1 + L_2 L_3) a^2).$$

Proof. Is analogous as in Theorem 2.2.

3 Stability of functional Volterra integral equation

Let $(\mathbb{B}, |\cdot|)$ a (real or complex) Banach space and $\varepsilon > 0$, a real number.

We consider the following functional Volterra integral equation:

$$u(x) = g(x, h(u)(x)) + \int_0^x K(x, s, u(s), f(u(s)))ds \tag{9}$$

where $K \in C([0,a]^2 \times \mathbb{B}^2, \mathbb{B})$, $g \in C([0,a] \times \mathbb{B}, \mathbb{B})$.

3.1 Ulam-Hyers stability of equation (9)

First we study the Ulam-Hyers stability of the equation (9).

Theorem 3.1 *If we have:*

(i) $K \in C([0,a]^2 \times \mathbb{B}^2, \mathbb{B})$, $g \in C([0,a] \times \mathbb{B}, \mathbb{B})$, $\varepsilon > 0$ *a real number;*

(ii) there exist $l_{K_1}, l_{K_2} > 0$ *such that*

$$|K(x, s, u, v) - K(x, s, \overline{u}, \overline{v})| \leq l_{K_1}|u - \overline{u}| + l_{K_2}|v - \overline{v}|,$$

for all $x, s \in [0, a]$, $u, v, \overline{u}, \overline{v} \in \mathbb{B}$*;*

(iii) there exists $l_g > 0$ *such that*

$$|g(x, e_1) - g(x, e_2)| \leq l_g|e_1 - e_2|;$$

(iv) there exists $l_h > 0$ *such that*

$$|h(u) - h(v)| \leq l_h|u - v|;$$

(v) there exists $l_f > 0$ *such that*

$$|f(u) - f(v)| \leq l_f|u - v|;$$

(vi) $l_g \cdot l_h < 1$.

Then:

(a) The equation (9) has in $C([0,a], \mathbb{B})$ *a unique solution* $u^*(x)$*;*

(b) If $u \in C([0,a], \mathbb{B})$ *is such that*

$$\left| u(x) - g(x, h(u)(x)) - \int_0^x K(x, s, u(s), f(u(s)))ds \right| \leq \varepsilon \tag{10}$$

for all $x, s \in [0, a]$*, then*

$$|u(x) - u^*(x)| \leq c_{K,g,h,f} \cdot \varepsilon, \ \forall \ x \in [0, a]$$

where

$$c_{K,g,h,f} = \frac{1}{1 - l_g l_h} \exp\left(\frac{l_{K_1} + l_{K_2} l_f}{1 - l_g l_h}\right) a \tag{11}$$

i.e., the equation (9) is Ulam-Hyers stable.

Proof.(a) It is a known result (see [5]).
(b) We have

$$|u(x) - u^*(x)| \leq \left| u(x) - g(x, h(u)(x)) - \int_0^x K(x, s, u(s), f(u(s)))ds \right|$$

$$+|g(x, h(u)(x)) - g(x, h(u^*)(x))| + \int_0^x |K(x, s, u(s), f(u(s))) - K(x, s, u^*(s), f(u^*(s)))|ds$$

$$\leq \varepsilon + l_g|h(u) - h(u^*)| + \int_0^x (l_{K_1}|u(s) - u^*(s)| + l_{K_2}|f(u(s)) - f(u^*(s))|)ds$$

$$\leq \varepsilon + l_g l_h|u(x) - u^*(x)| + \int_0^x (l_{K_1}|u(s) - u^*(s)| + l_{K_2} l_f|u(s) - u^*(s)|)ds.$$

Hence, we have

$$|u(x) - u^*(x)| \leq \varepsilon + l_g l_h|u(x) - u^*(x)| + \int_0^x (l_{K_1} + l_{K_2} l_f)|u(s) - u^*(s)|ds.$$

Then,

$$|u(x) - u^*(x)| \leq \frac{\varepsilon}{1 - l_g l_h} + \frac{l_{K_1} + l_{K_2} l_f}{1 - l_g l_h} \int_0^x |u(s) - u^*(s)|ds.$$

From Gronwall-lemma it follows that ([8])

$$|u(x) - u^*(x)| \leq \frac{1}{1 - l_g l_h} \exp\left(\frac{l_{K_1} + l_{K_2} l_f}{1 - l_g l_h} a\right) \cdot \varepsilon$$

and

$$|u(x) - u^*(x)| \leq c_{K,g,h,f} \cdot \varepsilon \tag{12}$$

where $c_{K,g,h,f}$ is given by (11) and the equation (9) is Ulam-Hyers stable.

3.2 Generalized Ulam-Hyers-Rassias stability of equation (9)

In what follows we study the generalized Ulam-Hyers-Rassias stability for the equation (9).

Theorem 3.2 *If we have:*
(i) $K \in C([0, a]^2 \times \mathbb{B}^2, \mathbb{B})$, $g \in C([0, a] \times \mathbb{B}, \mathbb{B})$;
(ii) there exist $l_{K_1}, l_{K_2} > 0$ *such that*

$$|K(x, s, u, v) - K(x, s, \overline{u}, \overline{v})| \leq l_{K_1}|u - \overline{u}| + l_{K_2}|v - \overline{v}|,$$

for all $x, s \in [0, a]$, $u, v, \overline{u}, \overline{v} \in \mathbb{B}$;
(iii) there exists $l_g > 0$ *such that*

$$|g(x, e_1) - g(x, e_2)| \leq l_g|e_1 - e_2|;$$

(iv) there exists $l_h > 0$ *such that*

$$|h(u) - h(v)| \leq l_h |u - v|;$$

(v) there exists $l_f > 0$ *such that*

$$|f(u) - f(v)| \leq l_f |u - v|;$$

(vi) $l_g l_h < 1$*;*
(vii) there exists the increasing function $\varphi \in C([0,a], \mathbb{R}_+)$*.*
Then:
(a) The equation (9) has in $C([0,a], \mathbb{B})$ *a unique solution* $u^*(x)$*;*
(b) If $u \in C([0,a], \mathbb{B})$ *is such that*

$$|u(x) - g(x, h(u)(x))| - \int_0^x K(x, s, u(s), f(u(s)))ds| \leq \varphi(x), \tag{13}$$

for all $x, s \in [0, a]$*, then*

$$|u(x) - u^*(x)| \leq c_{K,g,h,f} \cdot \varphi(x), \ \forall \ x \in [0, a] \tag{14}$$

where

$$c_{K,g,h,f} = \frac{1}{1 - l_g l_h} \exp\left(\frac{l_{K_1} + l_{K_2} l_f}{1 - l_g l_h}\right) a, \tag{15}$$

hence, the equation (9) is generalized Ulam-Hyers-Rassias stable.

Proof. Let us prove (b). We have

$$|u(x) - u^*(x)| \leq \left| u(x) - g(x, h(u)(x)) - \int_0^x K(x, s, u(s), f(u(s)))ds \right|$$

$$+|g(x, h(u)(x)) - g(x, h(u^*)(x))| + \int_0^x |K(x, s, u(s), f(u(s))) - K(x, s, u^*(s), f(u^*(s)))|ds$$

$$\leq \varphi(x) + l_g l_h |u(x) - u^*(x)| + \int_0^x (l_{K_1}|u(s) - u^*(s)| + l_{K_2} l_f |u(s) - u^*(s)|)ds.$$

Hence, we have

$$|u(x) - u^*(x)| \leq \frac{\varphi(x)}{1 - l_g l_h} + \frac{l_{K_1} + l_{K_2} l_f}{1 - l_g l_h} \int_0^x |u(s) - u^*(s)|ds.$$

From Gronwall-lemma it follows that ([8])

$$|u(x) - u^*(x)| \leq \frac{1}{1 - l_g l_h} \exp\left(\frac{l_{K_1} + l_{K_2} l_f}{1 - l_g l_h} a\right) \cdot \varphi(x),$$

and

$$|u(x) - u^*(x)| \leq c_{K,g,h,f} \cdot \varphi(x)$$

where $c_{K,g,h,f}$ is given by (15) and the equation (9) is generalized Ulam-Hyers-Rassias stable.

References

[1] D. H. HYERS, G. ISAC AND TH. M. RASSIAS, *Stability of Functional Equations in Several Variables*, Birkhäuser, Basel, 1998.

[2] JUNG, S.-M.: *Hyers-Ulam stability of first order linear differential equations with constant coefficients*, J. Math. Anal. Appl., **320**(2006), 549-561.

[3] JUNG, S.-M.: *A fixed point approach to the stability of a Volterra integral equation*, Fixed Point Theory and Applications, vol. 2007, 9 pages.

[4] LUNGU, N.: *On some Volterra integral inequalities*, Fixed Point Theory, **8**(2007), no. 1, 39-45.

[5] LUNGU, N. AND RUS, I. A.: *On a functional Volterra-Fredholm integral equation, via Picard operators*, Journal of Math. Inequalities, **3**(2009), no. 4, 519-527.

[6] POPA, D.: *Hyers-Ulam-Rassias stability of a linear recurrence*, J. Math. Anal. Appl., **309**(2005), 591-597.

[7] RUS, I. A. AND LUNGU, N.: *Ulam stability of a nonlinear hyperbolic partial differential equation*, Carpathian J. Math., **24**(2008), no. 3, 403-408.

[8] RUS, I. A.: *Gronwall lemma approach to the Hyers-Ulam-Rassias stability of an integral equation*, in: Nonlinear Analysis and Variational Problems, Springer, 2009, 147-152.

[9] RUS, I. A.: *Remarks on Ulam stability of the operatorial equations*, Fixed Point Theory, **10**(2009), no. 2, 305-320.

[10] RUS, I. A.: *Ulam stability of ordinary differential equations*, Studia Univ. Babeş-Bolyai, Mathematica, **54**(2009), no. 4, 125-133.

[11] RUS, I. A.: *Differential equations, Integral equations and Dynamical Systems*, Transilvania Press, Cluj-Napoca, 1996.

Current address

Sorina Anamaria Ciplea
Technical University of Cluj-Napoca, Department of Mathematics
28 Memorandumului Street, 400114, Cluj-Napoca, Romania

Nicolaie Lungu
Technical University of Cluj-Napoca, Department of Mathematics
28 Memorandumului Street, 400114, Cluj-Napoca, Romania
e-mail: nlungu@math.utcluj.ro

Journal of Applied Mathematics

volume IV (2011), number I

A NOTE ON A REPRESENTATION OF SOLUTIONS OF IMPULSIVE LINEAR DISCRETE SYSTEMS WITH CONSTANT COEFFICIENTS AND A SINGLE DELAY

DIBLÍK Josef, (CZ), MORÁVKOVÁ Blanka, (CZ)

Abstract. The purpose of this paper is to develop a method for the construction of solutions of linear discrete systems with constant coefficients, with pure delay and with impulses. Solutions are expressed by means of a special function called a discrete matrix delayed exponential.

Key words and phrases. Impulses, linear discrete system with constant coefficients.

Mathematics Subject Classification. Primary 39A10; Secondary 39A11

1 Introduction

We use the following notation throughout this paper: For integers s, q, $s \le q$, we define a set $\mathbb{Z}_s^q := \{s, s+1, \dots, q-1, q\}$. Similarly, we define sets $\mathbb{Z}_{-\infty}^q := \{\dots, q-1, q\}$ and $\mathbb{Z}_s^\infty := \{s, s+1, \dots\}$. The function $\lfloor\cdot\rfloor$ is the floor function.

Consider a initial Cauchy problem

$$\Delta x(k) = Bx(k-m), \quad k \in \mathbb{Z}_0^\infty, \tag{1}$$

$$x(k) = \varphi(k), \quad k \in \mathbb{Z}_{-m}^0 \tag{2}$$

where $m \ge 1$ is a fixed integer, $B = (b_{ij})$ is a constant $n \times n$ matrix, $x\colon \mathbb{Z}_{-m}^\infty \to \mathbb{R}^n$, $\varphi\colon \mathbb{Z}_{-m}^0 \to \mathbb{R}^n$ and $\Delta x(k) = x(k+1) - x(k)$.

We add impulses $J_i \in \mathbb{R}^n$ to x at points having a form $i(m+1)+p$ where the index $i \geq 0$ is defined as $i = \lfloor \frac{k-1}{m+1} \rfloor$ for every $k \in \mathbb{Z}_0^\infty$, i.e., we set

$$x(i(m+1)+p) = x(i(m+1)+p-0) + J_i \tag{3}$$

where $p = 1, 2, 3, ..., m+1$ and investigate the solution of problem (1) – (3).

A particural case of this problem (if $p = 1$) was solved in [1, 2]. We give a representation of the solution of this problem when, in every interval $\mathbb{Z}_{(\ell-1)(m+1)+1}^{\ell(m+1)}$, $\ell = 1, 2, \ldots,$ exactly one impuls J_i, $i = \lfloor \frac{k-1}{m+1} \rfloor$, $k = (l-1)(m+1)+p$, with fixed $p \in \mathbb{Z}_1^{(m+1)}$, acts on the solution of (1), (2). Our result generalizes the result in [1, 2].

In this paper, we use a special matrix function called a discrete function delayed exponential. Such a discrete matrix function was first defined in [3, 4].

Definition 1.1 *For an $n \times n$ constant matrix B, $k \in \mathbb{Z}$ and fixed $m \in \mathbb{N}$, we define a discrete matrix delayed exponential e_m^{Bk} as follows:*

$$\mathrm{e}_m^{Bk} := \begin{cases} \Theta & \text{if} \quad k \in \mathbb{Z}_{-\infty}^{-m-1}, \\ I & \text{if} \quad k \in \mathbb{Z}_{-m}^{0}, \\ I + B \cdot \binom{k}{1} & \text{if} \quad k \in \mathbb{Z}_1^{m+1}, \\ I + B \cdot \binom{k}{1} + B^2 \cdot \binom{k-m}{2} & \text{if} \quad k \in \mathbb{Z}_{(m+1)+1}^{2(m+1)}, \\ I + B \cdot \binom{k}{1} + B^2 \cdot \binom{k-m}{2} + B^3 \cdot \binom{k-2m}{3} & \text{if} \quad k \in \mathbb{Z}_{2(m+1)+1}^{3(m+1)}, \\ \cdots \\ I + B \cdot \binom{k}{1} + B^2 \cdot \binom{k-m}{2} + \cdots + B^\ell \cdot \binom{k-(\ell-1)m}{\ell} \\ \qquad \text{if} \quad k \in \mathbb{Z}_{(\ell-1)(m+1)+1}^{\ell(m+1)}, \ \ell = 0, 1, 2, \ldots \end{cases} \tag{4}$$

where Θ is $n \times n$ null matrix and I is $n \times n$ unit matrix.

Next, Theorem 1.2 is proved in [3].

Theorem 1.2 *Let B be a constant $n \times n$ matrix. Then, for $k \in \mathbb{Z}_{-m}^\infty$,*

$$\Delta\, \mathrm{e}_m^{Bk} = B\mathrm{e}_m^{B(k-m)}. \tag{5}$$

The following example illustrates the influence of impulses on the solution and serves as a motivation for the formulation of a general case.

Example 1.3 *We consider a particular case of* (1) *if* $n = 1$, $B = b$, $m = 3$, $p = 1$ *together with an initial problem* (2) *for* $\varphi(k) = 1$, $k \in \mathbb{Z}_{-3}^{0}$ *and with impulses* $J_i \in \mathbb{R}$ *at points* $i(m+1)+1 = 4i+1$ *where* $i \geq 0$, $i = \lfloor \frac{k-1}{m+1} \rfloor = \lfloor \frac{k-1}{4} \rfloor$*:*

$$\Delta x(k) = bx(k-3), \tag{6}$$

$$x(-3) = x(-2) = x(-1) = x(0) = 1, \tag{7}$$

$$x(4i+1) = x(4i+1-0) + J_i, \tag{8}$$

where $b \in \mathbb{R}$, $b \neq 0$. *Rewriting the equation as*

$$x(k+1) = x(k) + bx(k-3)$$

and solving it by the method of steps, we conclude that the solution of the problem, can be written in the form:

$$x(k) = b^0 \binom{k+3}{0} \qquad \text{if} \quad k \in \mathbb{Z}_{-3}^{0},$$

$$x(k) = b^0 \binom{k+3}{0} + b^1 \binom{k}{1} + J_0\, b^0 \binom{k-1}{0} \qquad \text{if} \quad k \in \mathbb{Z}_{1}^{4},$$

$$x(k) = b^0 \binom{k+3}{0} + b^1 \binom{k}{1} + b^2 \binom{k-3}{2} + J_0 \left[b^0 \binom{k-1}{0} + b^1 \binom{k-4}{1} \right] + J_1\, b^0 \binom{k-5}{0}$$
$$\text{if} \quad k \in \mathbb{Z}_{5}^{8},$$

$$\begin{aligned} x(k) = {} & b^0 \binom{k+3}{0} + b^1 \binom{k}{1} + b^2 \binom{k-3}{2} + b^3 \binom{k-6}{3} + J_0 \Bigg[b^0 \binom{k-1}{0} + b^1 \binom{k-4}{1} \\ & + b^2 \binom{k-7}{2} \Bigg] + J_1 \left[b^0 \binom{k-5}{0} + b^1 \binom{k-8}{1} \right] + J_2\, b^0 \binom{k-9}{0} \qquad \text{if} \quad k \in \mathbb{Z}_{9}^{12}, \end{aligned}$$

$$\vdots$$

$$\begin{aligned} x(k) = {} & b^0 \binom{k+3}{0} + b^1 \binom{k}{1} + b^2 \binom{k-3}{2} + \dots + b^{\ell} \binom{k-3(\ell-1)}{\ell} \\ & + J_0 \left[b^0 \binom{k-1}{0} + b^1 \binom{k-4}{1} + b^2 \binom{k-7}{2} + \dots + b^{\ell-1} \binom{k-4-3(\ell-2)}{\ell-1} \right] \\ & + J_1 \left[b^0 \binom{k-5}{0} + b^1 \binom{k-8}{1} + b^2 \binom{k-11}{2} + \dots + b^{\ell-2} \binom{k-8-3(\ell-3)}{\ell-2} \right] \\ & + J_2 \left[b^0 \binom{k-9}{0} + b^1 \binom{k-12}{1} + b^2 \binom{k-15}{2} + \dots + b^{\ell-3} \binom{k-12-3(\ell-4)}{\ell-3} \right] \\ & + J_3 \left[b^0 \binom{k-13}{0} + b^1 \binom{k-16}{1} + b^2 \binom{k-19}{2} + \dots + b^{\ell-4} \binom{k-16-3(\ell-5)}{\ell-4} \right] \end{aligned}$$

$$
\begin{aligned}
&+\cdots\\
&+J_i\left[b^0\binom{k-4(i+1)+3}{0}+b^1\binom{k-4(i+1)}{1}+b^2\binom{k-4(i+1)-3}{2}\right.\\
&\qquad\left.+b^3\binom{k-4(i+1)-6}{3}+\cdots+b^{\ell-(i+1)}\binom{k-4(i+1)-3(\ell-(i+2))}{\ell-(i+1)}\right]\\
&\qquad\qquad \textit{if}\quad k\in\mathbb{Z}_{4(\ell-1)+1}^{4(\ell-1)+4},\ \ell=0,1,2,\ldots,\ i=\left[\tfrac{k-1}{4}\right],\ i\geq 0.
\end{aligned}
$$

The solution of the problem (6) – (8) *can be shortened to*

$$
x(k)=\sum_{j=0}^{\ell}b^j\binom{k-3(j-1)}{j}+\sum_{q=0}^{i}J_q\sum_{j=0}^{\ell-(q+1)}b^j\binom{k-4(q+1)-3(j-1)}{j}, \tag{9}
$$

for $k\in\mathbb{Z}_{4(\ell-1)+1}^{4(\ell-1)+4},\ \ell=0,1,2,\ldots,\ i=\left[\frac{k-1}{4}\right],\ i\geq 0.$

2 Representation of a Solution of a Homogeneous Initial Problem

The following result gives a representation of a solution of problem (1) – (3).

Theorem 2.1 *Let B be a constant $n\times n$ matrix, m be a fixed integer, $J_i\in\mathbb{R}^n$. Then the solution of the initial Cauchy problem with impulses*

$$
\Delta x(k)=Bx(k-m),\quad k\in\mathbb{Z}_0^{\infty}, \tag{10}
$$

$$
x(k)=\varphi(k),\quad k\in\mathbb{Z}_{-m}^{0}, \tag{11}
$$

$$
x(i(m+1)+p)=x(i(m+1)+p-0)+J_i,\quad p=1,2,...,m+1,\quad i\geq 0,\quad i=\lfloor\tfrac{k-1}{m+1}\rfloor \tag{12}
$$

can be expressed in the form:

$$
x(k)=\mathrm{e}_m^{Bk}\varphi(-m)+\sum_{j=-m+1}^{0}\mathrm{e}_m^{B(k-m-j)}\Delta\varphi(j-1)+\sum_{q=0}^{i}J_q\mathrm{e}_m^{B(k-(p-1)-(q+1)(m+1))} \tag{13}
$$

where $k\in\mathbb{Z}_{-m}^{\infty}$.

Proof. We substitute (13) into the left-hand side $\mathcal{L}$ of the equation (10):

$$\begin{aligned}\mathcal{L} &= \Delta x(k)\\ &= \Delta\left[\mathrm{e}_m^{Bk}\varphi(-m) + \sum_{j=-m+1}^{0}\mathrm{e}_m^{B(k-m-j)}\Delta\varphi(j-1) + \sum_{q=0}^{i} J_q\mathrm{e}_m^{B(k-(p-1)-(q+1)(m+1))}\right]\\ &= [\text{according to the Theorem 1.2}]\\ &= \Delta\mathrm{e}_m^{Bk}\varphi(-m) + \sum_{j=-m+1}^{0}\Delta\mathrm{e}_m^{B(k-m-j)}\Delta\varphi(j-1) + \sum_{q=0}^{i} J_q\Delta\mathrm{e}_m^{B(k-(p-1)-(q+1)(m+1))}\\ &= B\mathrm{e}_m^{B(k-m)}\varphi(-m) + \sum_{j=-m+1}^{0}B\mathrm{e}_m^{B(k-2m-j)}\Delta\varphi(j-1) + \sum_{q=0}^{i} J_qB\mathrm{e}_m^{B(k-m-(p-1)-(q+1)(m+1))}\\ &= B\left[\mathrm{e}_m^{B(k-m)}\varphi(-m) + \sum_{j=-m+1}^{0}\mathrm{e}_m^{B(k-2m-j)}\Delta\varphi(j-1) + \sum_{q=0}^{i} J_q\mathrm{e}_m^{B(k-m-(p-1)-(q+1)(m+1))}\right].\end{aligned}$$

Now we substitute (13) into the right-hand side $\mathcal{R}$ of the equation (10):

$$\begin{aligned}\mathcal{R} &= Bx(k-m)\\ &= B\left[\mathrm{e}_m^{B(k-m)}\varphi(-m) + \sum_{j=-m+1}^{0}\mathrm{e}_m^{B(k-2m-j)}\Delta\varphi(j-1) + \sum_{q=0}^{i} J_q\mathrm{e}_m^{B(k-m-(p-1)-(q+1)(m+1))}\right].\end{aligned}$$

Since $\mathcal{L} = \mathcal{R}$, (13) is a solution of (10), (11).

Now we have to prove that (12) holds, too. We substitute (13) into the left-hand side $\mathcal{L}^*$ and right-hand side $\mathcal{R}^*$ of (12):

$$\begin{aligned}\mathcal{L}^* &= x(i(m+1)+p)\\ &= \mathrm{e}_m^{B(i(m+1)+p)}\varphi(-m) + \sum_{j=-m+1}^{0}\mathrm{e}_m^{B(i(m+1)+p-m-j)}\Delta\varphi(j-1) + \sum_{q=0}^{i} J_q\mathrm{e}_m^{B(i(m+1)+p-(p-1)-(q+1)(m+1))},\\ \mathcal{R}^* &= x(i(m+1)+p-0) + J_i\\ &= \mathrm{e}_m^{B(i(m+1)+p)}\varphi(-m) + \sum_{j=-m+1}^{0}\mathrm{e}_m^{B(i(m+1)+p-m-j)}\Delta\varphi(j-1) + \sum_{q=0}^{i-1} J_q\mathrm{e}_m^{B(i(m+1)+p-(p-1)-(q+1)(m+1))}\\ &\qquad + J_i.\end{aligned}$$

Since

$$\begin{aligned}\sum_{q=0}^{i} J_q \mathrm{e}_m^{B(i(m+1)+p-(p-1)-(q+1)(m+1))} &= \sum_{q=0}^{i-1} J_q \mathrm{e}_m^{B(i(m+1)+p-(p-1)-(q+1)(m+1))} \\ &\quad + J_i \mathrm{e}_m^{B(i(m+1)+p-(p-1)-(i+1)(m+1))} \\ &= \sum_{q=0}^{i-1} J_q \mathrm{e}_m^{B(i(m+1)+p-(p-1)-(q+1)(m+1))} + J_i \mathrm{e}_m^{B(-m)} \\ &= \left[\text{according to the Definition 1.1 , } \mathrm{e}_m^{B(-m)} = I\right] \\ &= \sum_{q=0}^{i-1} J_q \mathrm{e}_m^{B(i(m+1)+p-(p-1)-(q+1)(m+1))} + J_i\end{aligned}$$

it is obvious that $\mathcal{L}^* = \mathcal{R}^*$ and (12) holds.

Acknowledgement

Research was supported by the Grant P201/10/1032 of Czech Grant Agency (Prague) and by the Council of Czech Government MSM 00216 30503 and MSM 00216 30519.

References

[1] J. DIBLÍK, B. MORÁVKOVÁ, *Representation of solutions of linear discrete systems with constant coefficients, a single delay and with impulses*. Journal of Applied Mathematics, 2010, roc. III.(2010), c. 2, s. 45-52. ISSN: 1337- 6365.

[2] J. DIBLÍK, B. MORÁVKOVÁ, *Representation of solutions of linear discrete systems with constant coefficients, a single delay and with impulses*. In 9th International Conference APLIMAT 2010, Proceedings. Bratislava: STU, 2010. s. 141-148. ISBN: 978-80-89313-47-1.

[3] J. DIBLÍK, D. YA. KHUSAINOV, *Representation of solutions of linear discrete systems with constant coefficients and pure delay*. Adv. Difference Equ. 2006, Art. ID 80825, 13 pp.

[4] J. DIBLÍK, D. YA. KHUSAINOV, *Representation of solutions of discrete delayed system $x(k+1) = Ax(k) + Bx(k-m) + f(k)$ with commutative matrices*. J. Math. Anal. Appl. 318 (2006), no. 1, 63–76.

[5] J. DIBLÍK, D. YA. KHUSAINOV, Z. ŠMARDA, *Construction of the general solution of planar linear discrete systems with constant coefficients and weak delay*. Adv. Difference Equ. 2009, Art. ID 784935, 18 pp.

Current address

Diblík Josef, Prof. RNDr., DrSc.
Brno University of Technology, Czech Republic

tel.: +420-514147601, +420-541143155
email: diblik.j@fce.vutbr.cz, diblik@feec.vutbr.cz

Morávková Blanka, Mgr.
Brno University of Technology, Czech Republic
tel.: +420-541143154, +420-541147620
email: xmorav17@stud.feec.vutbr.cz, moravkova.b@fce.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

SPECTRAL ANALYSIS IN A MICROPOLAR FLUID FLOW PROBLEM

DRAGOMIRESCU Florica Ioana, (RO),
SIDDHESHWAR Ganapathi Pradeep (IND)

Abstract. The effects exhibited by micropolar fluids arising from the local structure and microrotation of the fluid elements have been widely investigated in the literature. Thermal instability conditions in a micropolar fluid layer heated from below are investigated in this paper based on the critical eigenvalues of the resulting two-point eigen-boundary-value problem governing the linear stability of the flow in the case of free, isothermal boundary conditions. By standard variational arguments a weak formulation of the problem is obtained. The neutral hypersurface separating the domain of stability from the linear instability domain is deduced and neutral curves emphasizing the influence of all physical parameters are graphically presented. The numerical results are obtained employing a weighted residual method based on generalized Jacobi polynomials.

Key words and phrases. micropolar fluids, polynomial based spectral methods, Rayleigh number.

Mathematics Subject Classification. 65L10, 76M22.

1 Introduction

The use of micropolar fluids in a number of processes that occur in industries concerning lubricants, colloidal suspensions, polymer fluids, liquid crystals, cooling of metallic plate in a bath [7] led to a large volume of research on this subject for applied mathematicians, physicists and engineers. Their extensive use is in fact the reason why micropolar fluid aspects have been widely investigated, emphasizing in each case the influence they have on the considered system stability. We have also investigated a Ginzburg -Landau model for Rayleigh - Bénard convection in a micropolar fluid with free, isothermal boundaries using nonlinear stability analysis techniques [14]. A very elegant expression of the Nusselt number as a function of the finite amplitude of convection was obtained.

In this paper, the conditions for the onset of convection in a micropolar fluid layer heated from below are investigated in the case when two sides restrict the normal flow and are shear stress free, microrotation free surfaces [1]. The linear stability of the fluid layer is governed by a fourth order differential system of equations with constant coefficients, supplied with free, isothermal boundary conditions, which is solved using spectral method. The strategy applied to lower the order into the ordinary differential equations defining the system transforms the boundary conditions from hinged boundary conditions to Dirichlet type boundary conditions.

Spectral methods arise from the fundamental problem of approximation of a function on an interval and are very much successful in obtaining a numerical solution of ordinary differential equations. The basis of spectral methods to solve classes of equations is to expand the solution function as a finite series of very smooth basis functions. In practice, summations in the expansions of the unknown functions are truncated to some finite values of the number of terms in the approximation, N, for which the higher order terms become essentially negligible. During the last few decades optimization became a very important objective which has expanded in many directions. For this purpose new algorithmic and theoretical techniques have been developed. Polynomials based spectral methods are widely used since they have the characteristic of very rapid convergence and also reduce the computational evaluations especially in the orthogonal polynomials case.

The paper is organized as follows. In the next section the eigenvalue problem obtained by the linearization procedure against normal mode perturbations for the initial governing equations problem is presented. The third section describes the method applied to solve the problem, emphasizing several classes of polynomials that can be used as trial and/or test basis functions in spectral methods and pointing out the advantages and disadvantages in each case. The numerical results section completes the analytical investigation. The main results of the paper are summarized in the conclusions section.

2 Basic equations and formulation of the problem

The onset of convection in the horizontal layer of micropolar fluid heated from below is governed by the conservation of momentum, internal angular momentum and internal energy balance equations [1]

$$\begin{array}{l} -\nabla p + (1+K)\nabla^2 \mathbf{u} + K\nabla \times \phi + R\theta \mathbf{k} = 0, \\ K(\nabla \times \mathbf{u} - 2\phi) - C_0 \nabla \times \nabla \times \phi + C_1 \nabla\nabla \cdot \phi = 0, \\ \nabla^2\theta + Rw = 0, \end{array} \tag{1}$$

with $\mathbf{u}$, p, θ, ϕ the perturbations of the governing fields of velocity, pressure, temperature and spin, C_0 accounts for the spin diffusion, K the coupling between vorticity and spin effects, $Ra = R^2$ is the Rayleigh number. The curl operator is then applied on $(1)_2$, and the $z-$component is taken. In [1] it is proven that the principle of exchange of stability holds, so the marginal stability is characterized by non-oscillatory motion, i.e., one can assume the following form of solution for the perturbations:

$$(w, \theta, \xi) = (W, \Theta, Z)(z)e^{i(a_x x + a_y y)}, \tag{2}$$

where $\xi = (\nabla \times \phi)_z = \dfrac{\partial \phi_y}{\partial x} - \dfrac{\partial \phi_z}{\partial y}$ and $a^2 = a_x^2 + a_y^2$. Following [1], we assume the principle of exchange of stability holds. The neutral stability of the thermal conduction against normal mode perturbations is governed by an eigenvalue problem consisting of a system of linear, ordinary differential equations with constant coefficients and a set of boundary conditions taken on the two boundaries:

$$\begin{cases} (1+K)(D^2-a^2)^2 W + K(D^2-a^2)Z - a^2 R\Theta = 0, \\ K[(D^2-a^2)W + 2Z] - C_0(D^2-a^2)Z = 0, \\ (D^2-a^2)\Theta + RW = 0, \end{cases} \tag{3}$$

where $D = \dfrac{d}{dz}$ represents the derivative with respect to the spatial independent variable z, a is the wavenumber, $Ra = R^2$ is the Rayleigh number, W, Θ, Z are the perturbation amplitudes for the vertical component of the velocity, temperature and the vertical component of the spin, respectively.

The boundary conditions for the free, isothermal, spin-vanishing boundaries case have the form

$$W = D^2 W = Z = \Theta = 0 \text{ at } z = 0, 1. \tag{4}$$

The problem (3) - (4) generates a linear differential operator $L : \mathcal{D}(L) \subset (L^2(0,1))^3 \to (L^2(0,1))^3$, with the domain of definition defined as follows,

$$\mathcal{D}_f(L) = \{(W, Z, \Theta) \in (C^\infty[0,1])^3 | W = D^2 W = Z = \Theta = 0 \text{ at } z = 0, 1\}.$$

The main idea is to determine in the parameter space (R, K, C_0, a) the neutral hypersurface which separates the domain of stability from the linear instability domain. The exact eigensolutions describing the thermal convection were investigated in [9] using the direct method for the hydrodynamic case. However, the presence of a large number of physical parameters allowed a complete analysis only in some particular cases. In [5] micropolar magnetic fluid stability domain was deduced by us in the rigid boundaries case, focussing our attention on the analytical Budianski-DiPrima method. For the discussions presented above, it is noticeable that the problem of stability analysis of micropolar fluid layer heated from below has been addressed mostly in the rigid boundaries case or in the free boundaries case with severe truncation of series in the Galerkin method used. With this in mind, we are interested in investigating the problem in the case of free, isothermal, spin-vanishing boundaries using a more sophisticated weighted - residual method.

3 Mathematical analysis and dispersion relation

Galerkin type spectral method has been extensively used in solving high order (especially fourth order) differential eigenvalue problems in hydrodynamic stability theory. However, in practical computations, imposing boundary conditions which involves derivatives of order higher than one can lead to tedious evaluations with significant rounding errors with respect to the selected trial and test functions. The construction of suitable spaces of functions satisfying as many as possible boundary conditions becomes a very important process. The problem of approximating solutions of ODE by polynomials, i.e., the polynomials-based Galerkin approximation, involves the projection onto the span of some appropriate set of basis functions, typically arising as the eigenfunctions of a singular Sturm-Liouville problem. The members of the basis automatically satisfy the auxiliary conditions imposed on the problem, such as initial, boundary or more general conditions. In the case of linear stability problems of hydrodynamic stability theory characterized by linear and homogeneous boundary conditions, these conditions are satisfied by a linear combination of such functions too. The expansion functions must have a basic property: they must be easy to evaluate. That is why, orthogonal polynomials can be used in the discretization process. Another property concerns the completeness of these families of functions such that each function of the given space can be represented as a limit of a linear combination of such functions. Since the standard interval of definition for these orthogonal polynomials is $[-1, 1]$, a linear transformation of the form $x := 2z - 1$ is used to map the interval $[0, 1]$ onto the interval $[-1, 1]$.

The finite dimensional space X_N on which we choose to construct the eigensolution (W, Z, Θ) is a subspace of the Hilbert space $(L^2(-1, 1))^3$ and the same it is considered here for the trial functions as well as the test functions space. An approximation $\mathbf{U}_N$ of the solution has the form $\mathbf{U}_N = (W_N\ Z_N\ \Theta_N)^T$, $\mathbf{U}_N \in X_N \setminus \{\mathbf{0}\}$ with the index N related to the discretization process. The approximations W_N, Z_N, Θ_N of the unknown functions are then defined as truncated series of trial functions

$$W_N = \sum_{k=0}^{N} \widehat{W}_z \phi_k^W(x), \quad Z_N = \sum_{k=0}^{N} \widehat{Z}_z \phi_k^Z(x), \quad \Theta_N = \sum_{k=0}^{N} \widehat{\Theta}_k \phi_k^\Theta(x). \tag{5}$$

with the Fourier coefficients $\widehat{W}_k$, $\widehat{Z}_k$, $\widehat{\Theta}_k$, $k = 0, 1, 2, ..., N$ unknown and the sets $\{\phi_k^{\mathbf{U}_l}(x)\}_{k=0,1,...,N}$, $l = 1, 2, 3$, $\mathbf{U} = (W, Z, \Theta)$ representing the sets of trial functions. The spectral approximation $\mathbf{U}_N$ is obtained by imposing the vanishing of the projection of the residual on the finite dimensional space X_N, i.e.,

$$\left(\mathbf{L_1}\left(\mathbf{U}_N\right), \mathbf{V}_N\right)_w + R \cdot \left(\mathbf{L}_2\left(\mathbf{U}_N\right), \mathbf{V}_N\right)_w = 0,\ \forall\ \mathbf{V}_N \in X_N \tag{6}$$

where $\mathbf{L_1} = \begin{pmatrix} (1+K)(4D^2 - a^2\mathbf{I})^2 & K(4D^2 - a^2\mathbf{I}) & O \\ K(4D^2 - a^2\mathbf{I}) & 2K\mathbf{I} - C_0(4D^2 - a^2\mathbf{I}) & O \\ O & O & (4D^2 - a^2\mathbf{I}) \end{pmatrix}$ and $\mathbf{L_2} = \begin{pmatrix} O & O & -a^2\mathbf{I} \\ O & O & O \\ \mathbf{I} & O & O \end{pmatrix}$.

The weight function w is taken with respect to the corresponding orthogonal polynomials, $\mathbf{I}$ represents the unit matrix and O the null matrix.

In order for the boundary conditions for the disturbance's amplitude of the vertical component of the velocity field W to be satisfied, a suitable trial set of functions $\phi_k(x) = j_k^{-2,-2}(x)$, $k = 1, 2, ..., N$, defining the space X_N is considered [11]

$$\phi_k(x) = \frac{4(k-2)(k-3)}{(2k-3)(2k-5)}\Big(L_{k-4}(x) - \frac{2(2k-3)}{2k-1}L_{k-2}(x) + \frac{2k-5}{2k-1}L_k(x)\Big), \tag{7}$$

where $L_k(x)$ are the Legendre polynomial of the k^{th} order. The $j_k^{-2,-2}(x)$ notation is given for the generalized Jacobi polynomials with index $\alpha = \beta = -2$ defined in details in [11].

Several such classes of generalized Jacobi polynomials $j_k^{\alpha,\beta}(x)$ are constructed in [11]. This selection, however, introduces in our case, for a large value of the spectral parameter N, tedious computations without always offering a convergence of the algorithm. For a small parameter N the numerical values are acceptable, the not so high accuracy is given by the selection of the $j_k^{-2,-2}(x)$ functions as expansion functions for all the unknown fields even if the boundary conditions allow a relaxation on the selection.

These results motivate the introduction of an alternative approach, i.e. the D^2 strategy [10]. By taking into account the order of differentiation in (3) - (4) we introduce the new variable $\Psi := (D^2 - a^2)W$. Thus, the two-point boundary value problem (3) - (4) can be rewritten as the second order system

$$\begin{cases} (D^2 - a^2)W - \Psi = 0, \\ (1+K)(D^2 - a^2)\Psi + K(D^2 - a^2)Z - a^2 R\Theta = 0, \\ K[(D^2 - a^2)W + 2Z] - C_0(D^2 - a^2)Z = 0, \\ (D^2 - a^2)\Theta + RW = 0, \end{cases} \tag{8}$$

with the homogeneous Dirichlet boundary conditions

$$\Psi = W = Z = \Theta = 0 \text{ at } x = -1 \text{ and } 1 \ . \tag{9}$$

This means a generalized eigenvalue problem.

A linear matricial n^{th} order differential operator $\mathbf{L}$ given by a $n \times n$ matrix (a_{ij}), where $a_{ij} = \sum_{k=1}^{n} a_{ij}^k D^k$, $D^k = \dfrac{\partial^k}{\partial x^k}$, a_{ij}^k are constants, is selfadjoint if we have $a_{ji} = \sum_{k=1}^{n} (-1)^k a_{ij}^k D^k$. In our case, for the operator $\mathbf{L}$ defined by the system (8) written as $\mathbf{L}\underline{\mathbf{U}} = \mathbf{0}$, with $\underline{\mathbf{U}} = (\Psi, W, Z, \Theta)$, the condition is fullfield if we multiply the first equation of (8) by $(1+K)$ and the last one by $-a^2$. The associated matrix of the operator is now a symmetric one given by

$$\begin{pmatrix} -(1+K)\mathbf{I} & (1+K)(4D^2 - a^2\mathbf{I}) & \mathbf{O} & \mathbf{O} \\ (1+K)(4D^2 - a^2\mathbf{I}) & \mathbf{O} & (4D^2 - a^2\mathbf{I}) & -a^2R\mathbf{I} \\ \mathbf{O} & K(4D^2 - a^2\mathbf{I}) & 2K\mathbf{I} - C_0(4D^2 - a^2\mathbf{I}) & \mathbf{O} \\ \mathbf{O} & -a^2R\mathbf{I} & \mathbf{O} & -a^2(4D^2 - a^2\mathbf{I}) \end{pmatrix}.$$

Also, a variational formulation of the problem led us to the following adjoint problem

$$\begin{cases} -(1+K)\Psi^* + (1+K)(D^2 - a^2)W^* = 0, \\ (1+K)(D^2 - a^2)\Psi^* + K(D^2 - a^2)Z^* - a^2R\Theta^* = 0, \\ K[(D^2 - a^2)W^*] + 2Z^* - C_0(D^2 - a^2)Z^* = 0, \\ -a^2RW^* - a^2(D^2 - a^2)\Theta^* = 0, \end{cases} \tag{10}$$

$$\Psi^* = W^* = Z^* = \Theta^* = 0 \text{ at } x = -1 \text{ and } 1, \tag{11}$$

where $\underline{\mathbf{U}}^* = (\Psi^*, W^*, Z^*, \Theta^*)$ is such that $(\mathbf{L}\underline{\mathbf{U}}, \underline{\mathbf{U}}^*) = (\mathbf{L}\underline{\mathbf{U}}^*, \underline{\mathbf{U}})$.

The finite dimensional space X_N was made up in this case by functions which satisfy the Dirichlet boundary conditions at $x = \pm 1$. Two families of functions $\{S_{jk}\}_{k=0,1,\ldots,N}$, $j = 1, 2$ are considered. The first one, a Shen's basis, denoted as (SB1), is defined by (see [10])

$$S_{1k}(x) = T_k(x) - T_{k+2}(x), \ k = 0, 1, \ldots, N. \tag{12}$$

The second Shen's basis, $\{S_{2k}\}_{k=0,1,\ldots,N}$, denoted this time as (SB2), has the form

$$S_{2k}(x) = T_k(x) - \frac{2(k+2)}{k+3}T_{k+2}(x) + \frac{k+1}{k+3}T_{k+4}(x),\ k = 0, 1, \ldots, N. \tag{13}$$

The functions from both the Shen's basis satisfy boundary conditions of the form $S_{jk}(\pm 1) = 0,\ j = 1, 2;\ k = 0, 1, \ldots, N$. The problem (4) can be written $\mathbf{L}\cdot\mathbf{U}_M^{coef} = \mathbf{0}$, with

$$\mathbf{U}_M^{coef} = (\widehat{\Psi}_0 \ldots \widehat{\Psi}_N \widehat{W}_0 \ldots \widehat{W}_N \widehat{Z}_0 \ldots \widehat{\widehat{Z}}_N \widehat{\Theta}_0 \ldots \widehat{\Theta}_N)$$

the vector of the unknown coefficients. This means we have a homogeneous system of $4(N+1)$ linear algebraic equations in $4(N+1)$ unknowns.

Now we consider for the numerical approximation the functions

$$\phi_k(x) = j_k^{-1,-1}(x) = \frac{2(k-1)}{2k-1}(L_{k-2}(x) - L_k(x)).$$

These are polynomials that satisfy Dirichlet type boundary conditions at ± 1. Polynomials of the form

$$\phi_k(x) = (1-x^2)\cdot P_k^{\alpha,\beta}(x), \quad k = 1, \ldots, N \tag{14}$$

which can be viewed as generalized Jacobi polynomials [6] and that which fulfill the Dirichlet boundary conditions, were also considered for numerical investigation.

Numerical results for all these basis are presented in the next section of the paper.

4 Numerical results and discussions

For abbreviations in the presentation of numerical results we denote the trial sets of function as follows: for the $j_k^{-2,-2}$ polynomials the abreviation Jacobi0 is used, for the $j_k^{-1,-1}$ polynomials - Jacobi1 is used, for the polynomials defined in (14) we use the notation Jacobi2. As we stated in the previous section, SB1 and SB2 notations will be used for the two Shen basis.

a^2	K	C_0	$R_{Jacobi0}$
4.92	0.00001	0.001	29.64156
4.92	0.00001	0.01	30.508985
4.92	0.00001	0.1	31.111212
4.92	0.00001	1	29.447312
4.92	0.00001	10	29.919165

Table 1. Selective numerical evaluations of the Rayleigh number for the Jacobi0 trial set of functions for various values of the micropolar physical parameters - the problem (3) - (4).

The numerical results for the original eigenvalue problem (3) - (4) obtained by using the trial functions Jacobi0 are summarized in Table 1. These results show that the accuracy of derivatives obtained by direct, term-by-term differentiation of such truncated expansions naturally deteriorates. It is well known that for sufficiently high-order truncations this deterioration is negligible, compared to the restrictions in accuracy introduced by typical difference approximations. In our case, however, for large values of the spectral parameter N, say $N > 6$, the convergence of the algorithm is no longer assured. For small values of the index N, the error has acceptable limits. The relative error for the values of the Rayleigh number $Ra = R^2$ is for some cases almost 30% (for instance, for the case $K = 10^{-5}$, $C_0 = 10^{-3}$ the eigenparameter R has a numerical value around 25.70 and here this value is almost 30). That is why, the D^2 strategy, transforming the eigenvalue problem into a generalized one proved to be very successful. For the case $K = 0$, $C_0 = 0$ we get the results of the classical Rayleigh-Bénard convection case [3]. The numerical evaluations also show that the Ra_{min} values for each expansion set of functions occurs around $a = 2.22$, similarly to the classical case of Rayleigh - Bénard convection. The present numerical computations were performed with the dimensionless perturbation wavenumber a ranging from 2 to 3 in order to observe the behavior of the critical Rayleigh number values, known to be obtainable in this interval for nonvanishing micropolar parameters.

a^2	K	C_0	R_{SB1}	R_{SB2}
4.92	10^{-5}	10^{-3}	25.73867468	26.46203686
4.92	10^{-5}	10^{-2}	25.63612410	26.46203696
4.92	10^{-5}	10^{-1}	25.63612413	26.46203695
4.92	10^{-5}	1	25.63612409	26.46203697
4.92	10^{-5}	10	25.63612412	26.46203700
4.92	0.1	10	26.88643399	27.75264555
4.92	1	10	36.19424673	37.36159888
4.92	10	10	82.69073630	85.39988570
4.92	11	10	8612970773	88.95572796
4.92	12	10	89.40375390	92.34140685
4.92	14	10	95.52826109	98.67567207
4.00	10	20	84.31620116	87.24243418
4.50	10	20	83.87496903	86.68316945
4.92	10	20	83.79202619	86.51668151
5.00	10	20	83.80067755	86.51075204
7.00	10	20	85.72735763	88.16704606
9.00	10	20	89.55986275	91.83548932

Table 2. Selective numerical evaluations of the critical eigenparameter R for various values of the other physical parameters involved using SB1, SB2 - the problem (8) - (9).

Clearly, the best results are offered by the first Shen basis. For this basis, the investigation of the classical case gives us the results: at $N = 3$ the relative error is around $4.66E - 4$, at $N = 4$ this error is less than $6.08E - 6$ offering yet a small computational time. A larger value of the spectral parameter increases the computational time without offering a major decrease of the relative error.

In Fig.1 a), b) and c), the influence of each of the micropolar parameters on the critical values of the eigenparameter defining the stability domain is shown. We have varied the wavenumber, the coupling parameter K and the spin diffusion coefficient, respectively, keeping all other parameters values fixed. More than that, the investigations were performed for all the presented expansion functions.

a^2	K	C_0	R
4.92	10^{-5}	10^{-3}	25.9821
4.92	10^{-5}	10^{-2}	25.9826
4.92	10^{-5}	10^{-1}	25.9821
4.92	10^{-5}	1	25.9819
4.92	10^{-5}	10	25.9822
4.92	10^{-1}	10	26.1119
4.92	1	10	36.6824
4.92	10	10	84.0349
4.92	11	10	87.4980
4.92	12	10	90.4326
4.92	14	10	96.3868
4.00	10	20	85.5700
4.50	10	20	84.9809
4.92	10	20	85.2077
5.00	10	20	84.8394
7.00	10	20	86.1382
9.00	10	20	89.8374

Table 3. Numerical evaluations of the critical eigenparameter R for various values of the other physical parameters involved for Jacobi1.

a^2	K	C_0	R_0	$R_{-1/2}$	$R_{1/2}$
4.92	10^{-5}	10^{-3}	25.9816	24.9441	27.7427
4.92	10^{-5}	10^{-2}	25.9808	24.9441	27.7427
4.92	10^{-5}	10^{-1}	25.9829	24.9441	27.7427
4.92	10^{-5}	1	25.9814	25.6609	27.7427
4.92	10^{-5}	10	25.9822	25.6609	25.6609
4.92	10^{-1}	10	27.2487	26.1607	26.9124
4.92	1	10	36.6809	35.2162	36.1942
4.92	10	10	84.1710	82.7721	82.69073
4.92	11	10	87.1457	86.2145	86.1297
4.92	12	10	90.0645	89.4919	89.4037
4.92	14	10	97.2951	95.5282	95.5282
4.00	10	20	85.0402	81.8665	91.7846
4.50	10	20	84.6074	81.5232	91.0400
4.92	10	20	85.0728	81.5303	90.7423
5.00	10	20	85.1870	81.5096	90.7134
7.00	10	20	86.9149	83.6811	91.9445
9.00	10	20	89.9149	87.6494	95.3546

Table 4. Numerical evaluations of the critical eigenparameter $R_i$for various values of the other physical parameters involved for Jacobi2. The index i point out $\alpha = \beta = i$.

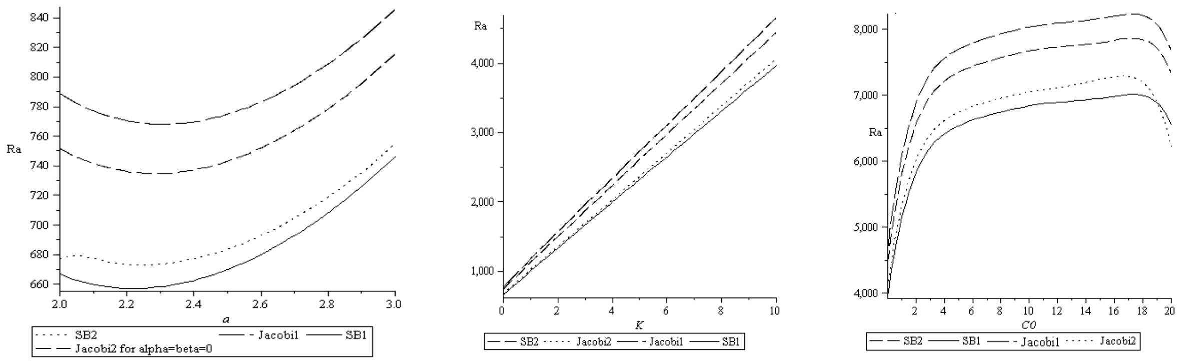


Fig.1. a) Dependence on a of the Rayleigh number for the case $K = 0.00001$, $C_0 = 0.01$.

b) Dependence on K of the Rayleigh number for the case $a^2 = 4.92$, $C_0 = 0.01$.

c) Dependence on C_0 of the Rayleigh number for the case $a^2 = 4.92$, $K = 10$.

The dependence of the Rayleigh number on the micropolar parameter K is linear and a sensitivity of its values on K is emphasized by the numerical evaluations, so we can point out that an increase in the values of K is responsible for a delayed onset of convection. For larger values of C_0, say $C_0 > 100$, the Rayleigh number values is very small less than 10^{-5}%. In this case also best numerical results are provided by the SB1 trial set of functions without any major discrepancies in the results.

5 Conclusions

In this paper an analytical and numerical investigation of the thermal instability problem in a micropolar fluid flow is provided. Several polynomials type trial basis of functions are proposed for the numerical investigation using the Galerkin method. In each case the secular equation led to neutral manifolds in the parameters spaces offering the possibility to investigate the physical influence of each micropolar parameter on the stability domain. It was shown that for increasing values of these parameters the stability domain also increases. The accuracy of the methods are also evaluated with respect to previous existing results only in the classical Rayleigh-Bénard convection since eigenvalue evaluations were performed usually in the rigid boundaries case. The accuracy of the method is increased by far by applying a D^2 strategy which transforms the original eigenvalue problem into a generalized one with lower order derivatives and Dirichlet type boundary conditions.

The conclusion is that the micropolar fluid layer heated from below is more stable than Newtonian liquid. One other conclusion is that severely truncated polynomial based Galerkin methods, should not be used in estimating eigenvalues in a two-point boundary value problem. The reason is severely truncated methods may only serve the purpose of qualitative prediction, though not always. For the case of trigonometric based Galerkin methods, however, the truncation to one term leads to very good numerical results. This will be proved in a paper that is under preparation.

References

[1] AHMADI, G.: *Stability of a micropolar fluid layer heated from below*, Int. J. Engng Sci., Vol. 14, pp. 81-89, 1976.

[2] ANNAMMA A.:*Rayleigh Bénard convection in a micropolar ferromagnetic fluid*, Int. J. Engng. Sci., Vol. Vol 40, No. 4, pp. 449-460, 2002.

[3] CHANDRASEKHAR, S., Hydrodynamic and hydromagnetic stability, Oxford Univ. Press, 1961.

[4] DATTA, A.B., SASTRY, V.U.K, *Thermal instability of a horizontal layer of micropolar fluid heated from below*, Int. J. Engng. Sci., Vol. 14, No. 7 1976, 631-637.

[5] DRAGOMIRESCU, F. I., *Rayleigh number in a stability problem for a micropolar fluid*, Turkish Journal of Mathematics, Vol. 31, No. 2, 2007, 123-137.

[6] DRAGOMIRESCU, F. I., *A $P_n^{\alpha,\beta}$-based method for linear nonconstant coefficients high order eigenvalue problems*, Lecture Notes in Computational Science and Engineering, Vol. 76, 2010, pp. 383-391.

[7] EZZAT, M.A., OTHMAN, I.A., *Thermal instability in a rotating micropolar fluid layer subject to an electric field*, Int. J. Engng. Sci., **38**, 16, 2000, 1851-1867.

[8] ERINGEN, A.C., *Simple microfluids*, Int. J. Engng. Sci., Vol. 2, (1964), 205-217.

[9] GEORGESCU, A., GAVRILESCU, M., PALESE, L., *Neutral thermal hydrodynamic and hydromagnetic stability hypersurface for a micropolar fluid layer*, Indian J. Pure Appl.Math., Vol. 29, No. 6, 1998, 575-582.

[10] GHEORGHIU C. I., DRAGOMIRESCU, F. I., *Spectral methods in linear stability. Applications to thermal convection with variable gravity field*, Appl. Num. Math., Vol. 59, No. 6, 2009, 1290-1302.
[11] GUO, B.Y., SHEN, J., WANG, L.L., *Generalized Jacobi polynomials/functions and their applications*, Appl. Num. Math., Vol. 59, No. 5, 2009, 1011-1028.
[12] SASTRY, V. U. K, RAMAMOHAN RAO, V., *Numerical solution of thermal instability of a rotating micropolar fluid layer*, Int. J Engng. Sci., Vol. 21 , No. 5, 1983, 449-461.
[13] SIDDHESHWAR, P. G., PRANESH, S., *Magnetoconvection in a micropolar fluid* Int. J. Engng. Sci., Vol. 36 (1998), pp. 1173-1181.
[14] SIDDHESHWAR, P. G., DRAGOMIRESCU, F.I., *Ginzburg-Landau model for Rayleigh-Bnard convection in a micropolar fluid with free isothermal boundaries*, Nonlinear convection in micropolar liquids (submitted paper to Int. J. Thermal Sciences, 2010).

Current address

Dragomirescu Florica Ioana, Lecturer
Deparment of Mathematics, "Politehnica" University of Timisoara,
P-ta Victoriei, No.2, Romania,
e-mail: ioana.dragomirescu@mat.upt.ro

Siddheshwar Ganapathi Pradeep, Professor
Department of Mathematics, Bangalore University,
Central College Campus Bangalore - 560 001, INDIA,
e-mail: pgsmath@gmail.com

Journal of Applied Mathematics

volume IV (2011), number I

ON SHARPNESS OF TWO-SIDED DISCRETE MAXIMUM PRINCIPLES FOR REACTION-DIFFUSION PROBLEMS

FARAGÓ István, (HU), KOROTOV Sergey, (E), SZABÓ Tamás, (HU)

Abstract. We present and discuss several numerical tests for checking the sharpness of the discrete maximum principles in the form of two-sided estimates for the reaction-diffusion problems of the elliptic type solved by the finite element and finite difference methods.

Key words and phrases. reaction-diffusion problem, two-sided maximum principles.

Mathematics Subject Classification. 35B50, 65N06, 65N30, 65N50

1 Model problem and maximum principles

We shall deal with the following boundary-value problem: Find a function $u \in C^2(\overline{\Omega})$ such that

$$-\Delta u + cu = f \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbf{R}^d$ is a bounded domain with Lipschitz continuous boundary $\partial\Omega$ and the reactive coefficient $c(x) \geq 0$ for all $x \in \overline{\Omega}$. We also assume that c and the right-hand side (RHS) function f are both from $C(\overline{\Omega})$.

The following two-sided estimation (the so-called modified (or *two-sided*) maximum principle (MP) [5]) holds.

Theorem 1.1 *Let functions c and f in (1) be from $C(\overline{\Omega})$, and let, in addition,*

$$c(x) \geq c_0 > 0 \quad \textit{for all} \quad x \in \overline{\Omega}. \tag{2}$$

Then

$$\min\Big\{0, \min_{s\in\overline{\Omega}} \frac{f(s)}{c(s)}\Big\} \leq u(x) \leq \max\Big\{0, \max_{s\in\overline{\Omega}} \frac{f(s)}{c(s)}\Big\} \quad \textit{for any} \quad x \in \overline{\Omega}. \tag{3}$$

After discretization of (1) by FEM or FDM [3] we arrive at the problem of solving $n \times n$ system of linear algebraic equations

$$\mathbf{A}\mathbf{u} = \mathbf{F}, \tag{4}$$

where the vector of unknowns $\mathbf{u} = [u_1, \dots, u_n]^T$ approximates the unknown solution u at certain selected points $B_1, \dots, B_n$ of the solution domain Ω, and the vector $\mathbf{F} = [F_1, \dots, F_n]^T$ approximates (in the sense related to the nature of a concrete numerical method used) the values $f(B_i),\ i = 1, \dots, n$.

In what follows, the entries of matrix $\mathbf{A}$ will be denoted by a_{ij}, and all matrix and vector inequalities appearing in the text are always understood component-wise.

Definition 1.2 *The square $n \times n$ matrix $\mathbf{A}$ (with entries m_{ij}) is called strictly diagonally dominant (SDD) if the values*

$$\alpha_i(\mathbf{A}) := |a_{ii}| - r_i > 0 \quad \textit{for} \quad \textit{all} \quad i = 1, \dots, n, \tag{5}$$

where r_i is the sum of absolute values of all off-diagonal entries in the i-th row of $\mathbf{A}$, i.e. $r_i := \sum\limits_{j=1,\, j\neq i}^{n} |a_{ij}|$.

Theorem 1.3 *Let matrix $\mathbf{A}$ in system* (4) *be SDD and monotone* [12]*. Then the following two-sided estimates for the entries of the solution $\mathbf{u}$ are valid*

$$\min\Big\{0, \min_{j=1,\dots,n} \frac{F_j}{\alpha_j(\mathbf{A})}\Big\} \le u_i \le \max\Big\{0, \max_{j=1,\dots,n} \frac{F_j}{\alpha_j(\mathbf{A})}\Big\}, \quad i = 1, \dots, n. \tag{6}$$

For the proof see [5].

As (6) actually resembles the estimates (3), it is natural to give the following definition.

Definition 1.4 *We say that the solution $\mathbf{u}$ of system (4) with SDD matrix $\mathbf{A}$ satisfies the two-sided (or modified) discrete maximum principle (DMP), corresponding to MP (3), if estimates (6) are valid and if, in addition,*

$$\max_{j=1,\dots,n} \frac{F_j}{\alpha_j(\mathbf{A})} \le \max\Big\{0, \max_{s\in\overline{\Omega}} \frac{f(s)}{c(s)}\Big\}, \ \min_{j=1,\dots,n} \frac{F_j}{\alpha_j(\mathbf{A})} \ge \min\Big\{0, \min_{s\in\overline{\Omega}} \frac{f(s)}{c(s)}\Big\}. \tag{7}$$

Remark 1.5 *The conditions (7) are really important in order to produce reliable (i.e. controllable) numerical simulations as, for example, linear finite element (FE) and finite difference (FD) approximations do stay within the same (a priori known) limits (and preserve right signs if needed) then as the exact solutions they do approximate.*

Remark 1.6 *While the SDD-property of* $\mathbf{A}$ *is almost automatically guaranteed after discretization by the nature of the reaction-diffusion problems (namely, due to the presence of nonvanishing reactive terms), its monotonicity, required in Theorem 1.3, should be provided a priori (or proved separately in each concrete case). One common approach for this in FEM is to impose certain a priori geometric requirements on the FE meshes employed so that all the off-diagonal entries* $a_{ij} \leq 0$ *(see e.g. [1, 2, 4, 7, 8, 10, 11] for more details on this subject). As far it concerns FDM, this property for the off-diagonal entries of* $\mathbf{A}$ *is often guaranteed a priori by many standard FD schemes producing the so-called* M*-matrices [6].*

Remark 1.7 *One of advantages for dealing namely with the property* $a_{ij} \leq 0$ $(i \neq j)$ *is an easy calculation (or estimation) of values* $\alpha_i(\mathbf{A})$ *and their relation to* c *in general (see e.g. [5]). However, the property* $a_{ij} \leq 0$ $(i \neq j)$ *is only sufficient for monotonicity (and therefore for DMPs) (cf. [9]).*

Remark 1.8 *The validity of DMP for various (classical) FE and FD schemes is proved in [5]. In the next section we present some numerical tests supporting the theoretical analysis done in [5].*

2 On sharpness of two-sided discrete maximum principles

Here, we present and shortly discuss several numerical tests checking the sharpness of the discrete maximum principle from the previous section.

Test 1: First, we consider the following one-dimensional problem

$$-\frac{d^2u}{dx^2} + u = f_1 := 4xe^x, \ x \in (0,1), \quad u(0) = u(1) = 0.$$

Its exact solution is $u(x) = x(1-x)e^x$. It is easy to show that

$$\min_{x\in[0,1]} u(x) = 0, \quad \max_{x\in[0,1]} u(x) \approx 0.4380.$$

In this test $c(x) \equiv 1$, and therefore, in a view of Theorem 1.1, a priori estimation of the solution results in (formally correct) bounds

$$0 = \min_{x\in[0,1]} f_1(x) \leq u(x) \leq \max_{x\in[0,1]} f_1(x) = 4e \approx 10.8730,$$

which, however, considerably overestimate the true solution (from above).

In Table 1 we present the results of estimations and computations by FDM and FEM, where n is the number of nodes in the uniform partitions of the solution domain $(0,1)$. We used linear finite elements and the standard 3-nodes FD stencil, so that all the conditions of Theorem 1.3 are valid. We also computed all entries in the corresponding system of linear equations exactly (also for all the other tests), so the validity of DMP (7) can be easily proved (see [5] for details), which is also demonstrated in our calculations (Table 1).

	Estimated discrete				*Real discrete*			
	FDM		**FEM**		**FDM**		**FEM**	
n	**min**	**max**	**min**	**max**	**min**	**max**	**min**	**max**
11	0	8.8546	0	7.1633	0	0.4354	0	0.4294
101	0	10.6573	0	10.4452	0	0.4379	0	0.4379
1001	0	10.8514	0	10.8297	0	0.4380	0	0.4380

Table 1: Results of theoretical estimations and calculations for Test 1.

Test 2: In this case we consider the problem with a varying coefficient c

$$-\frac{d^2u}{dx^2} + (1+x)u = f_2 := (4x + x^2 - x^3)e^x,\ x \in (0,1), \quad u(0) = u(1) = 0,$$

whose exact solution is $u(x) = x(1-x)e^x$.

The minimum and maximum of u are the same as in the previous test, i.e.,

$$0 \leq u(x) \leq 0.4380.$$

The estimation of the minimum and the maximum by Theorem 1.1 gives now the following (correct) bounds

$$0 = \min_{x\in[0,1]} \frac{f_2(x)}{1+x} \leq u(x) \leq \max_{x\in[0,1]} \frac{f_2(x)}{1+x} = 2e \approx 5.4370,$$

which are less overestimating than those in the first test, which can be explained by the fact that now $c = 1 + x$ is not constant.

We notice that in case of varying coefficient c there have been no theoretical results on the validity of DMP done in [5]. However, DMP (7) seems to be valid in this case anyway, which is also supported by our numerical experiments summarized in Table 1. Here and in the next Tests 3 and 4, we used the same numerical schemes as in the first test.

	Estimated discrete				*Real discrete*			
	FDM		**FEM**		**FDM**		**FEM**	
n	**min**	**max**	**min**	**max**	**min**	**max**	**min**	**max**
11	0	4.7652	0	4.1310	0	0.4354	0	0.4299
101	0	5.3687	0	5.3011	0	0.4379	0	0.4379
1001	0	5.4298	0	5.4230	0	0.4380	0	0.4380

Table 2: Results of theoretical estimations and calculations for Test 2.

Test 3: Now we shall consider the problem with the RHS not preserving its sign

$$-\frac{d^2u}{dx^2} + u = f_3 := 2(x \sin x - \cos x),\ x \in (0,\pi), \quad u(0) = u(\pi) = 0.$$

The exact solution is $u(x) = x\,\sin x$, and therefore we have

$$0 \le u(x) \le 1.8197.$$

The estimated minimum and maximum (due to Theorem 1.1) are

$$-2.0000 = \min_{x\in[0,\pi]} f_3(x) \le u(x) \le \max_{x\in[0,\pi]} f_3(x) \approx 4.7632,$$

which presents less overestimating bounds than in the previous cases.

The results of theoretical estimation and real calculations presented in Table 3 again support the theoretical analysis of DMPs as shown in [5].

	Estimated discrete				*Real discrete*			
	FDM		**FEM**		**FDM**		**FEM**	
n	**min**	**max**	**min**	**max**	**min**	**max**	**min**	**max**
11	-1.7080	4.7338	-1.2852	4.6185	0	1.8022	0	1.7786
101	-1.9970	4.7632	-1.9872	4.7620	0	1.8196	0	1.8193
1001	-2.0000	4.7632	-1.9999	4.7632	0	1.8197	0	1.8197

Table 3: Results of theoretical estimations and calculations for Test 3.

Test 4: Neither the solution nor the RHS of the problem is preserving its sign

$$-\frac{d^2u}{dx^2} + u = f_4 = f_3, \quad x \in (0, 2\pi), \quad u(0) = u(2\pi) = 0. \tag{8}$$

The exact solution of the above problem is $u(x) = x\sin x$. The minimum and the maximum of the solution are

$$-4.8145 \le u(x) \le 1.8197.$$

The minimum and maximum estimation results in more accurate bounds than those in Tests 1, 2, and 3:

$$-10.2003 \approx \min_{x\in[0,2\pi]} f_4(x) \le u(x) \le \max_{x\in[0,2\pi]} f_4(x) \approx 4.7632.$$

The estimation and computational details are summarized in Table 4.

	Estimated discrete				*Real discrete*			
	FDM		**FEM**		**FDM**		**FEM**	
n	**min**	**max**	**min**	**max**	**min**	**max**	**min**	**max**
11	-10.1791	4.5726	-9.4550	4.0823	-4.8600	1.8461	-4.6610	1.7628
101	-10.2002	4.7606	-10.1925	4.7558	-4.8148	1.8197	-4.8130	1.8189
1001	-10.2002	4.7632	-10.2001	4.7632	-4.8145	1.8197	-4.8145	1.8197

Table 4: Results of theoretical estimations and calculations for Test 4.

Test 5: We consider a two-dimensional problem over a square domain, with sign preserving both, the solution and the right-hand side

$$-\Delta u + u = f_5(x,y) := (7xy - 3x^2y - 3xy^2 - x^2y^2)e^{x+y} \quad \text{in} \quad \Omega := (0,1)\times(0,1),$$

$$u = 0 \quad \text{on} \quad \partial\Omega,$$

where the exact solution is $u(x,y) = x(1-x)y(1-y)e^{x+y}$. We can easily compute that

$$\min_{\bar\Omega} u(x,y) = 0, \quad \max_{\bar\Omega} u(x,y) \approx 0.1918.$$

Theorem 1.1 produces the following bounds:

$$0 = \min_{\bar\Omega} f_5 \le u(x,y) \le \max_{\bar\Omega} f_5 \approx 4.9577.$$

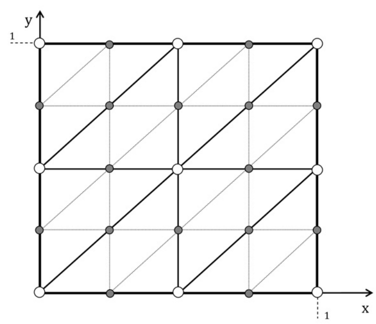


Figure 1: FDM and FEM meshes (and their refinements) used in Test 5.

The estimation and computational results are given in Table 5 (n denotes the total number of nodes). We used linear finite elements and the standard 5-nodes FD stencil (see Figure 1).

	Estimated discrete				*Real discrete*			
	FDM		**FEM**		**FDM**		**FEM**	
n	**min**	**max**	**min**	**max**	**min**	**max**	**min**	**max**
6x6	0	3.6335	0	5.2831	0	0.1877	0	0.1668
11x11	0	4.9451	0	7.1464	0	0.1904	0	0.1850
51x51	0	4.9562	0	7.4245	0	0.1918	0	0.1915

Table 5: Results of theoretical estimations and calculations for Test 5.

In this test, the theoretical estimation of FE approximations by Theorem 2, does not obey the upper bound provided by Theorem 1. This is explained by the fact that in this case the matrices being still monotone have some positive off-diagonal entries which leads to values $\alpha_j(\mathbf{A})$ be smaller than the corresponding values of $c(B_j)$. However, such an effect does not appear in the case of finite differences.

Test 6: This example is also a two-dimensional problem but now posed over a trianglar domain, and for which the solution and the RHS are both preserving their sign

$$-\Delta u + u = f_6(x,y) := y^3 - 2y^2 + (-4x^2 + 4x + 2)y + 4 \quad \text{in} \quad \Omega,$$

$$u = 0 \quad \text{on} \quad \partial\Omega.$$

The exact solution $u(x,y) = -y(2x-y)(2x+y-2)$.

The real maximum and the minimum:

$$0 \leq u(x,y) \leq \frac{4}{27} \approx 0.1481.$$

The estimated minimum and maximum for this problem are as follows

$$0 = \min_{\bar{\Omega}} f_6 \leq u(x,y) \leq \max_{\bar{\Omega}} f_6 = 6.0000.$$

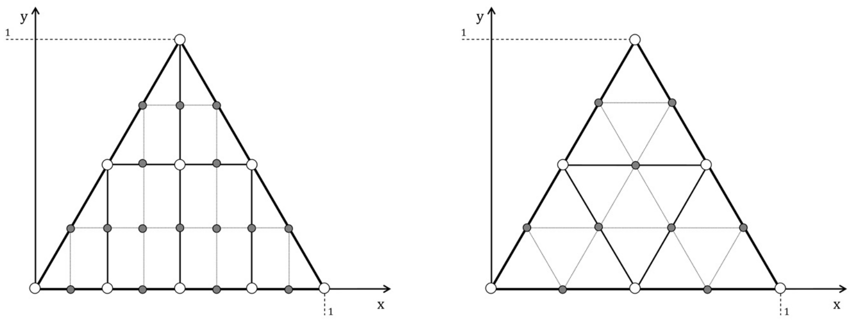


Figure 2: FD and FE meshes (and their refinements) used in Test 6.

We used linear finite elements and the standard 5-nodes FD stencil (see Figure 2 for details on meshes employed and their refinements). The results of estimations and computations are given in Tables 6 and 7, where n again denotes the total number of nodes in the meshes/stencils.

	Estimated discrete		*Real discrete*	
n	**min**	**max**	**min**	**max**
25	0	5.1250	0	0.1406
81	0	5.5469	0	0.1465
289	0	5.7637	0	0.1477
1089	0	5.8787	0	0.1480
4356	0	5.9384	0	0.1481

Table 6: Results of FD estimations and calculations for Test 6.

	Estimated discrete		*Real discrete*	
n	**min**	**max**	**min**	**max**
15	0	0.0852	0	0.0848
45	0	5.1146	0	0.1306
153	0	5.5443	0	0.1440
561	0	5.7630	0	0.1471
2145	0	5.8785	0	0.1479

Table 7: Results of FE estimations and calculations for Test 6.

As in this case we use acute FE meshes, and therefore all appearing matrices are M-matrices, the phenomenon observed in Test 5 does not appear here.

3 Conclusions

On the base of computations we can conclude that in the quite many situations both, estimation of the exact solution and estimation of FE/FD approximations [5] can overestimate quite a lot the real exact solution and real approximations, so more sophisticated a priori solution estimation tools are needed.

Acknowledgement

The first author was supported by Hungarian National Research Fund OTKA No. K67819.

The second author was supported by Grant MTM2008-03541 of the MICINN, Spain, the ERC Advanced Grant FP7-246775 NUMERIWAVES and Grant PI2010-04 of the Basque Government.

The first and the third authors were supported by Jedlik project "ReCoMend" 2008–2011. The European Union and the European Social Fund have provided financial support to the project under the grant agreement no. TÁMOP 4.2.1./B-09/1/KMR-2010-0003

References

[1] BRANDTS, J., KOROTOV, S., KŘÍŽEK, M., The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, *Linear Algebra Appl.* 429 (2008), 2344–2357.

[2] BRANDTS, J., KOROTOV, S., KŘÍŽEK, M., ŠOLC, J., On nonobtuse simplicial partitions, *SIAM Rev.* 51 (2009), 317–335.

[3] CIARLET, P.G., Discrete maximum principle for finite-difference operators, *Aequationes Math.* 4 (1970), 338–352.

[4] CIARLET, P.G., RAVIART, P.-A., Maximum principle and uniform convergence for the finite element method, *Comput. Methods Appl. Mech. Engrg.* 2 (1973), 17–31.

[5] FARAGÓ, I., KOROTOV, S., SZABÓ, T., On modifications of continuous and discrete maximum principles for reaction-diffusion problems, *Adv. Appl. Math. Mech.*, 3 (2011), 109–120.

[6] FORSYTHE, G. E., WASOW, W. R., *Finite-Difference Methods for Partial Differential Equations*. Reprint of the 1960 original. Dover Phoenix Editions. Dover Publications, Inc., Mineola, NY, 2004.

[7] HANNUKAINEN, A., KOROTOV, S., VEJCHODSKÝ, T., Discrete maximum principle for FE solutions of the diffusion-reaction problem on prismatic meshes, *J. Comput. Appl. Math.* 226 (2009), 275–287.

[8] KARÁTSON, J., KOROTOV, S., Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, *Numer. Math.* 99 (2005), 669–698.

[9] KOROTOV, S., KŘÍŽEK, M., NEITTAANMÄKI, P., Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle, *Math. Comp.* 70 (2001), 107–119.

[10] KOROTOV, S., VEJCHODSKÝ, T., *A comparison of simplicial and block finite elements.* In "Proc. of The Eighth Europ. Conf. on Numerical Mathematics and Advanced Applications (ENUMATH–2009), Uppsala, Sweden", (ed. by G. Kreiss et al.), Springer-Verlag, Heidelberg, 2010, 531–540.

[11] KŘÍŽEK, M., QUN LIN, On diagonal dominance of stiffness matrices in 3D, *East-West J. Numer. Math.*, 3 (1995), 59–69.

[12] VARGA, R., *Matrix Iterative Analysis*, Prentice Hall, 1962.

Current address

István Faragó, Tamás Szabó
Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University
H–1117, Budapest, Pázmány P. s. 1/c., Hungary
e-mail: faragois@cs.elte.hu, szabot@cs.elte.hu

Sergey Korotov
BCAM – Basque Center for Applied Mathematics
Bizkaia Technology Park, Building 500, E – 48160 Derio
Basque Country, Spain
e-mail: korotov@bcamath.org

Journal of Applied Mathematics

volume IV (2011), number I

BIFURCATION OF PERIODIC ORBITS IN DISCONTINUOUS SYSTEMS

FEČKAN Michal, (SK), POSPÍŠIL Michal, (SK)

Abstract. We study persistence of periodic orbit in discontinuous autonomous equations in $\mathbb{R}^n$ under nonautonomous perturbations. Also the bifurcation of a periodic orbit from a family of trajectories under autonomous perturbations is investigated. In the both cases, sufficient conditions of Melnikov type are stated.

Key words and phrases. Discontinuous system, periodic orbit, bifurcation.

Mathematics Subject Classification. Primary 34C23; Secondary 34C25, 37G15.

1 Introduction

This is a survey note on our study of periodic bifurcations in piecewise smooth systems [2, 3]. Now we formulate the problem. Let $\Omega \subset \mathbb{R}^n$ be an open set in $\mathbb{R}^n$ and h be a C^r-function on $\overline{\Omega}$, with $r \geq 3$. We set $\Omega_\pm := \{x \in \Omega \mid \pm h(x) > 0\}$, $\Omega_0 := \{x \in \Omega \mid h(x) = 0\}$. Let $f_\pm \in C^r(\overline{\Omega})$, i.e. the derivative of $f_\pm$ is continuous up to the r-th order. Let $\varepsilon \in \mathbb{R}$, $\alpha \in \mathbb{R}$, $\mu \in \mathbb{R}^p, p \geq 1$ be parameters and $\langle \cdot, \cdot \rangle$ be an inner product in $\mathbb{R}^n$. In the first part of the paper, we consider the problem of persistence of periodic orbits in nonautonomous equations given by

$$\dot{x} = f_\pm(x) + \varepsilon g(x, t + \alpha, \varepsilon, \mu), \quad x \in \overline{\Omega}_\pm \tag{1}$$

with $g \in C^r(\overline{\Omega} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p)$ being T-periodic in $t \in \mathbb{R}$ and autonomous one

$$\dot{x} = f_\pm(x) + \varepsilon g(x, \varepsilon, \mu), \quad x \in \overline{\Omega}_\pm, \tag{2}$$

where $g \in C^r(\overline{\Omega} \times \mathbb{R} \times \mathbb{R}^p)$.

We say that a function $x(t)$ is a solution of the equation (1), resp. (2), if it is continuous, piecewise C^1, satisfies equation (1), resp. (2) on $\Omega_\pm$ and, moreover, the following holds: if for some t_0 we have $x(t_0) \in \Omega_0$, then there exists $r > 0$ such that for any $t \in (t_0 - r, t_0)$ we have $x(t) \in \Omega_\pm$, and for any $t \in (t_0, t_0 + r)$ we have $x(t) \in \Omega_\mp$.

2 Nonautonomous perturbation

Let $x_+(\tau,\xi)(t,\varepsilon,\mu,\alpha)$ denote a solution of

$$\begin{aligned}\dot{x} &= f_+(x) + \varepsilon g(x, t+\alpha, \varepsilon, \mu),\\ x_+(\tau,\xi)(\tau,\varepsilon,\mu,\alpha) &= \xi\end{aligned} \tag{3}$$

and $x_-(\tau,\xi)(t,\varepsilon,\mu,\alpha)$ denote a solution of

$$\begin{aligned}\dot{x} &= f_-(x) + \varepsilon g(x, t+\alpha, \varepsilon, \mu),\\ x_-(\tau,\xi)(\tau,\varepsilon,\mu,\alpha) &= \xi.\end{aligned} \tag{4}$$

We assume

H1) For $\varepsilon = 0$ equation (1) has a T-periodic solution $\gamma(t)$ with an initial point $x_0 \in \Omega_+$ and consists of three branches

$$\gamma(t) = \begin{cases} \gamma_1(t) & \text{if} \quad t \in [0, t_1],\\ \gamma_2(t) & \text{if} \quad t \in [t_1, t_2],\\ \gamma_3(t) & \text{if} \quad t \in [t_2, T],\end{cases} \tag{5}$$

where $\gamma_1(t) \in \Omega_+$ for $t \in [0, t_1)$, $\gamma_2(t) \in \Omega_-$ for $t \in (t_1, t_2)$ and $\gamma_3(t) \in \Omega_+$ for $t \in (t_2, T]$ and

$$\begin{aligned} x_1 &:= \gamma_1(t_1) = \gamma_2(t_1) && \in \Omega_0,\\ x_2 &:= \gamma_2(t_2) = \gamma_3(t_2) && \in \Omega_0,\\ x_0 &:= \gamma_3(T) = \gamma_1(0) && \in \Omega_+.\end{aligned} \tag{6}$$

H2) $\mathrm{D}h(x_1)f_\pm(x_1) < 0$ and $\mathrm{D}h(x_2)f_\pm(x_2) > 0$.

In [2] the following lemma is stated.

Lemma 2.1 *Assume* H1) *and* H2). *Then there exist $\varepsilon_0, r_0 > 0$ and a Poincaré mapping*

$$P(\cdot,\varepsilon,\mu,\alpha) : U \to \Sigma$$

for all fixed $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, $\mu \in \mathbb{R}^p$, $\alpha \in \mathbb{R}$, where $\Sigma = \{x \in \mathbb{R}^n \mid \langle x - x_0, f_+(x_0)\rangle = 0\}$, $U = \Sigma \cap B(x_0, r_0)$ and $B(x,r)$ is the ball of radius r and center in x. Moreover, P is C^r-smooth in all arguments.

Using this Poicaré mapping and the fact, that we are looking for T-periodic solutions, the initial point $\xi \in \Sigma$ of such a solution has to satisfy

$$\xi - \widetilde{P}(\xi,\varepsilon,\mu,\alpha) = 0 \tag{7}$$

where

$$\begin{aligned}\widetilde{P}(\xi,\varepsilon,\mu,\alpha) = x_+(t_2(0,\xi,\varepsilon,\mu,\alpha), x_-(t_1(0,\xi,\varepsilon,\mu,\alpha), x_+(0,\xi)(t_1(0,\xi,\varepsilon,\mu,\alpha),\varepsilon,\mu,\alpha))\\ (t_2(0,\xi,\varepsilon,\mu,\alpha),\varepsilon,\mu,\alpha))(T,\varepsilon,\mu,\alpha)\end{aligned} \tag{8}$$

and

$$h(x_+(\tau,\xi)(t_1(\tau,\xi,\varepsilon,\mu,\alpha),\varepsilon,\mu,\alpha)) = 0,$$
$$h(x_-(t_1(\tau,\xi,\varepsilon,\mu,\alpha),x_+(\tau,\xi)(t_1(\tau,\xi,\varepsilon,\mu,\alpha),\varepsilon,\mu,\alpha))(t_2(\tau,\xi,\varepsilon,\mu,\alpha),\varepsilon,\mu,\alpha)) = 0$$

for τ close to 0, $\xi \in B(x_0, r_0)$, $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, $\mu \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$.

Mapping $\widetilde{P}$ has the following properties (cf. [2] for the proof).

Lemma 2.2 *Let $\widetilde{P}(\xi,\varepsilon,\mu,\alpha)$ be defined by (8) for $\xi \in B(x_0,r_0)$, $\varepsilon \in (-\varepsilon_0,\varepsilon_0)$, $\mu \in \mathbb{R}^p$, $\alpha \in \mathbb{R}$. Then $\widetilde{P}_\xi(x_0,0,\mu,\alpha)$ has an eigenvalue 1 with the corresponding eigenvector $f(x_0)$, i.e.*

$$\widetilde{P}_\xi(x_0,0,\mu,\alpha)f(x_0) = f(x_0).$$

Moreover,

$$\widetilde{P}_\xi(x_0,0,\mu,\alpha) = X_3(T)S_2X_2(t_2)S_1X_1(t_1), \tag{9}$$
$$\widetilde{P}_\varepsilon(x_0,0,\mu,\alpha) = \int_0^T A(s)g(\gamma(s),s+\alpha,0,\mu)ds, \tag{10}$$

where $X_1(t)$, $X_2(t)$ and $X_3(t)$ are matrix solutions of equations

$$\begin{array}{lll} \dot{X}_1(t) = \mathrm{D}f_+(\gamma(t))X_1(t), & \dot{X}_2(t) = \mathrm{D}f_-(\gamma(t))X_2(t), & \dot{X}_3(t) = \mathrm{D}f_+(\gamma(t))X_3(t), \\ X_1(0) = \mathbb{I}, & X_2(t_1) = \mathbb{I}, & X_3(t_2) = \mathbb{I}, \end{array}$$

$$S_1 = \mathbb{I} + \frac{(f_-(x_1) - f_+(x_1))\mathrm{D}h(x_1)}{\mathrm{D}h(x_1)f_+(x_1)}, \qquad S_2 = \mathbb{I} + \frac{(f_+(x_2) - f_-(x_2))\mathrm{D}h(x_2)}{\mathrm{D}h(x_2)f_-(x_2)}$$

and

$$A(t) = \begin{cases} X_3(T)S_2X_2(t_2)S_1X_1(t_1)X_1^{-1}(t) & \text{if } t \in [0,t_1), \\ X_3(T)S_2X_2(t_2)X_2^{-1}(t) & \text{if } t \in [t_1,t_2), \\ X_3(T)X_3^{-1}(t) & \text{if } t \in [t_2,T]. \end{cases} \tag{11}$$

From the last lemma one can see that $(\mathbb{I} - \widetilde{P}_\xi(x_0,0,\mu,\alpha))$ has the nonempty null space. For simplicity we add one more assumption

H3) dim $\mathcal{N}(\mathbb{I} - \widetilde{P}_\xi(x_0,0,\mu,\alpha)) = 1$,

where $\mathcal{N}(\mathbb{I} - \widetilde{P}_\xi(x_0,0,\mu,\alpha))$ denotes the null space of the corresponding operator. We use Lyapunov-Schmidt reduction method to solve equation (7). Denote $Y := \mathcal{R}(\mathbb{I} - \widetilde{P}_\xi(x_0,0,\mu,\alpha))$ the range of $\mathbb{I} - \widetilde{P}_\xi(x_0,0,\mu,\alpha)$ and define orthogonal projections $\mathcal{P}: \mathbb{R}^n \to Y^\perp$ and $\mathcal{Q}: \mathbb{R}^n \to Y$ by

$$\mathcal{P}y = \frac{\langle y,\psi\rangle}{|\psi|^2}\psi, \qquad \mathcal{Q}y = (\mathbb{I} - \mathcal{P})y$$

for arbitrary fixed $\psi \in Y^\perp$. Consequetly, (7) splits into a couple of equations

$$\mathcal{Q}(\xi - \widetilde{P}(\xi,\varepsilon,\mu,\alpha)) = 0, \qquad \mathcal{P}(\xi - \widetilde{P}(\xi,\varepsilon,\mu,\alpha)) = 0.$$

Implicit Function Theorem [1] applied on the first one implies the existence of a unique C^r-function $\xi(\varepsilon,\mu,\alpha)$ for ε close to 0, which solves the equation. Moreover $\xi(0,\mu,\alpha)=x_0$. Using the form of $\mathcal{P}$, we obtain the bifurcation equation

$$\langle \xi(\varepsilon,\mu,\alpha)-\widetilde{P}(\xi(\varepsilon,\mu,\alpha),\varepsilon,\mu,\alpha),\psi\rangle = 0.$$

Differentiation with respect to ε at 0 gives the discontinuous Melnikov function

$$M^{\mu}(\alpha) = -\int_0^T \langle g(\gamma(s),s+\alpha,0,\mu),A(s)^*\psi\rangle ds, \tag{12}$$

where

$$A(t)^* = \begin{cases} X_1^{-1*}(t)X_1(t_1)^*S_1^*X_2(t_2)^*S_2^*X_3(T)^* & \text{if} \quad t\in[0,t_1), \\ X_2^{-1*}(t)X_2(t_2)^*S_2^*X_3(T)^* & \text{if} \quad t\in[t_1,t_2), \\ X_3^{-1*}(t)X_3(T)^* & \text{if} \quad t\in[t_2,T]. \end{cases}$$

solves the adjoint variational equation [5]

$$\begin{array}{ll} X' = -f_+^*(\gamma(t))X & \text{if} \quad 0<t<t_1, \\ X' = -f_-^*(\gamma(t))X & \text{if} \quad t_1<t<t_2, \\ X' = -f_+^*(\gamma(t))X & \text{if} \quad t_2<t<T. \end{array}$$

As a result we get the next theorem.

Theorem 2.3 *Let conditions* H1), H2), H3) *hold,* $M^{\mu}(\alpha)$ *be defined by* (12) *and* $\psi\in Y^{\perp}$ *is fixed. If there is a* $(\mu_0,\alpha_0)\in\mathbb{R}^p\times\mathbb{R}$*, such that* $M^{\mu_0}(\alpha_0)=0$ *and* $\mathrm{D}_{\alpha}M^{\mu_0}(\alpha_0)\neq 0$ *then there exists a neighbourhood* U *of the point* $(0,\mu_0)$ *in* $\mathbb{R}\times\mathbb{R}^p$ *and a* C^{r-1}*-function* $\alpha(\varepsilon,\mu)$*, with* $\alpha(0,\mu_0)=\alpha_0$*, such that perturbed equation* (7) *possesses a unique* T*-periodic piecewise* C^1*-smooth solution for each* $(\varepsilon,\mu)\in U$*.*

3 Autonomous perturbation

In this section following [3], we derive a sufficient condition for persistence of a single trajectory from a family of periodic orbits. Similarly to previous section, let $x_{\pm}(\tau,\xi)(t,\varepsilon,\mu)$ denotes the solution of the initial value problem

$$\begin{aligned} \dot{x} &= f_{\pm}(x)+\varepsilon g(x,\varepsilon,\mu), \\ x(\tau) &= \xi \end{aligned}$$

with the corresponding sign. We assume

H1') For $\varepsilon=0$ equation (2) has a smooth family of T^{β}-periodic orbits $\{\gamma(\beta,t)\}$ parametrized by $\beta\in V\subset\mathbb{R}^k$, $0<k<n$, V is an open set in $\mathbb{R}^k$, T is a C^r-function of β. Each orbit is uniquely determined by its initial point $x_0(\beta)\in\Omega_+$, $x_0\in C^r$, and consists of three branches

$$\gamma(\beta,t) = \begin{cases} \gamma_1(\beta,t) & \text{if} \quad t\in[0,t_1^{\beta}], \\ \gamma_2(\beta,t) & \text{if} \quad t\in[t_1^{\beta},t_2^{\beta}], \\ \gamma_3(\beta,t) & \text{if} \quad t\in[t_2^{\beta},T^{\beta}], \end{cases} \tag{13}$$

where $0 < t_1^\beta < t_2^\beta < T^\beta$, $\gamma_1(\beta,t) \in \Omega_+$ for $t \in [0, t_1^\beta)$, $\gamma_2(\beta,t) \in \Omega_-$ for $t \in (t_1^\beta, t_2^\beta)$, $\gamma_3(\beta,t) \in \Omega_+$ for $t \in (t_2^\beta, T^\beta]$ and

$$\begin{aligned} x_1(\beta) &:= \gamma_1(\beta, t_1^\beta) = \gamma_2(\beta, t_1^\beta) \in \Omega_0, \\ x_2(\beta) &:= \gamma_2(\beta, t_2^\beta) = \gamma_3(\beta, t_2^\beta) \in \Omega_0, \\ x_0(\beta) &= \gamma_3(\beta, T^\beta) = \gamma_1(\beta, 0) \in \Omega_+. \end{aligned} \tag{14}$$

Note t_1^β and t_2^β are C^r-functions of β. We suppose in addition that vectors

$$\frac{\partial x_0(\beta)}{\partial \beta_1}, \ldots, \frac{\partial x_0(\beta)}{\partial \beta_k}, f_+(x_0(\beta))$$

are linearly independent whenever $\beta \in V$.

H2') $\mathrm{D}h(x_1(\beta))f_\pm(x_1(\beta)) < 0$ and $\mathrm{D}h(x_2(\beta))f_\pm(x_2(\beta)) > 0$ in V.

Now we study local bifurcations from $\gamma(\beta,t)$, so we fix $\beta_0 \in V$ and set $x_0^0 = x_0(\beta_0)$, $t_1^0 = t_1^{\beta_0}$, etc. Analogically to the previous case, we have the Poincaré mapping.

Lemma 3.1 *Assume* H1') *and* H2')*. Then there exist $\varepsilon_0, r_0 > 0$, a neighborhood $W \subset V$ of β_0 in $\mathbb{R}^k$ and a Poincaré mapping*

$$P(\cdot, \beta, \varepsilon, \mu) : B(x_0^0, r_0) \to \Sigma_\beta$$

for all fixed $\beta \in W$, $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, $\mu \in \mathbb{R}^p$, where

$$\Sigma_\beta = \{y \in \mathbb{R}^n \mid \langle y - x_0(\beta), f_+(x_0(\beta))\rangle = 0\}.$$

Moreover, $P : B(x_0^0, r_0) \times W \times (-\varepsilon_0, \varepsilon_0) \times \mathbb{R}^p \to \mathbb{R}^n$ is C^r-smooth in all arguments and $x_0(W) \subset B(x_0^0, r_0) \subset \Omega_+$.

Mapping P is defined by

$$\begin{aligned} P(\xi, \beta, \varepsilon, \mu) = x_+(t_2(0,\xi,\varepsilon,\mu), x_-(t_1(0,\xi,\varepsilon,\mu), x_+(0,\xi)(t_1(0,\xi,\varepsilon,\mu), \varepsilon, \mu)) \\ (t_2(0,\xi,\varepsilon,\mu), \varepsilon, \mu))(t_3(0,\xi,\beta,\varepsilon,\mu), \varepsilon, \mu). \end{aligned} \tag{15}$$

for $\xi \in B(x_0^0, r_0)$, $\beta \in W$, $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$ and $\mu \in \mathbb{R}^p$. It has the following properties.

Lemma 3.2 *For any $\xi \in B(x_0^0, r_0)$, $\beta \in W$, $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$ and $\mu \in \mathbb{R}^p$ $P_\xi(\xi, \beta, \varepsilon, \mu)$ defined by* (15) *has an eigenvalue* 0 *with the corresponding eigenvector $f_+(\xi) + \varepsilon g(\xi, \varepsilon, \mu)$, i.e.*

$$P_\xi(\xi, \beta, \varepsilon, \mu)[f_+(\xi) + \varepsilon g(\xi, \varepsilon, \mu)] = 0.$$

Moreover,

$$P_\xi(x_0(\beta), \beta, 0, \mu) = (\mathbb{I} - S_\beta)A(\beta, 0), \tag{16}$$

$$P_\beta(x_0(\beta), \beta, 0, \mu) = S_\beta \mathrm{D}x_0(\beta), \tag{17}$$

$$P_\varepsilon(x_0(\beta), \beta, 0, \mu) = (\mathbb{I} - S_\beta)\left(\int_0^{T^\beta} A(\beta, s)g(\gamma(\beta, s), 0, \mu)ds\right), \tag{18}$$

where S_β *is the orthogonal projection onto the 1-dimensional space* $[f_+(x_0(\beta))]$ *defined by*

$$S_\beta u = \frac{\langle u, f_+(x_0(\beta))\rangle f_+(x_0(\beta))}{\|f_+(x_0(\beta))\|^2}$$

and A *is given by*

$$A(\beta,t) = \begin{cases} X_3(\beta,T^\beta)S_2(\beta)X_2(\beta,t_2^\beta)S_1(\beta)X_1(\beta,t_1^\beta)X_1^{-1}(\beta,t) & \text{if } t\in[0,t_1^\beta),\\ X_3(\beta,T^\beta)S_2(\beta)X_2(\beta,t_2^\beta)X_2^{-1}(\beta,t) & \text{if } t\in[t_1^\beta,t_2^\beta),\\ X_3(\beta,T^\beta)X_3^{-1}(\beta,t) & \text{if } t\in[t_2^\beta,T^\beta], \end{cases} \tag{19}$$

where

$$S_1(\beta) = \mathbb{I} + \frac{(f_-(x_1(\beta)) - f_+(x_1(\beta)))\mathrm{D}h(x_1(\beta))}{\mathrm{D}h(x_1(\beta))f_+(x_1(\beta))},$$
$$S_2(\beta) = \mathbb{I} + \frac{(f_+(x_2(\beta)) - f_-(x_2(\beta)))\mathrm{D}h(x_2(\beta))}{\mathrm{D}h(x_2(\beta))f_-(x_2(\beta))},$$

and X_1*,* X_2 *and* X_3 *solve the following equations, respectively,*

$$\dot X_1(\beta,t) = \mathrm{D}f_+(\gamma(\beta,t))X_1(\beta,t), \qquad \dot X_2(\beta,t) = \mathrm{D}f_-(\gamma(\beta,t))X_2(\beta,t),$$
$$X_1(\beta,0)=\mathbb{I}, \qquad X_2(\beta,t_1^\beta)=\mathbb{I},$$

$$\dot X_3(\beta,t) = \mathrm{D}f_+(\gamma(\beta,t))X_3(\beta,t),$$
$$X_3(\beta,t_2^\beta)=\mathbb{I}.$$

For any $\xi\in\mathbb{R}^n$ we define the orthogonal projection $\widetilde{S}_\beta$ onto Σ_β as follows

$$\widetilde{S}_\beta : \xi \mapsto \xi - S_\beta(\xi - x_0(\beta)).$$

Denoting $F(\xi,\beta,\varepsilon,\mu) = \widetilde{S}_\beta(\xi) - P(\xi,\beta,\varepsilon,\mu)$, ξ is an initial point from Σ_β of periodic orbit of perturbed system (2) if and only if it satisfies

$$F(\xi,\beta,\varepsilon,\mu) = 0, \qquad \xi\in\Sigma_\beta. \tag{20}$$

Using Lemma 3.2, it can be shown [3] that for any $v\in\left\{\frac{\partial x_0(\beta)}{\partial\beta_1},\dots,\frac{\partial x_0(\beta)}{\partial\beta_k}, f_+(x_0(\beta))\right\}$ it holds

$$F_\xi(x_0(\beta),\beta,0,\mu)v = 0, \qquad \forall\beta\in W.$$

For simplicity, we assume one more condition.

H3') The set

$$\left\{\frac{\partial x_0(\beta)}{\partial\beta_1},\dots,\frac{\partial x_0(\beta)}{\partial\beta_k}, f_+(x_0(\beta))\right\}$$

spans the null space of the operator $F_\xi(x_0(\beta),\beta,0,\mu)$.

Note $\Sigma_\beta = [f_+(x_0(\beta))]^\perp + x_0(\beta)$. Let us denote

$$Z_\beta = \mathcal{N}F_\xi(x_0(\beta),\beta,0,\mu) \cap [f_+(x_0(\beta))]^\perp, \quad Y_\beta = \mathcal{R}F_\xi(x_0(\beta),\beta,0,\mu)$$

the restricted null space and the range of the corresponding operator, respectively. Condition H3') implies $Z_\beta = (\mathbb{I} - S_\beta)\mathrm{D}x_0(\beta)$. Using the Gram-Schmidt orthogonalization we find an orthonormal basis $\{y_1,\dots,y_{n-k-1}\}$ for vector space $Z_\beta^\perp$ such that $Z_\beta \perp Z_\beta^\perp$ and $Z_\beta + Z_\beta^\perp = [f_+(x_0(\beta))]^\perp$. Define projections

$$\mathcal{Q}_\beta : \Sigma_\beta \to Y_\beta, \qquad \mathcal{P}_\beta : \Sigma_\beta \to Y_\beta^\perp,$$

where $Y_\beta^\perp$ is an orthogonal complement to Y_β in $[f_+(x_0(\beta))]^\perp$, and the decomposition for any z close to $x_0(W)$

$$z = x_0(\beta) + \xi \quad \text{for} \quad \beta \in W, \xi \in \left[\frac{\partial x_0(\beta)}{\partial \beta_1},\dots,\frac{\partial x_0(\beta)}{\partial \beta_k}\right]^\perp.$$

Assuming only $\xi \in \Sigma_\beta$ slightly changes the decomposition to a form

$$z = x_0(\beta) + \xi, \qquad \beta \in W,\ \xi \in Z_\beta^\perp.$$

We apply the Lyapunov-Schmidt reduction to equation (20) and write it as a couple of equations

$$\mathcal{Q}_\beta F(x_0(\beta)+\xi,\beta,\varepsilon,\mu) = 0, \qquad \mathcal{P}_\beta F(x_0(\beta)+\xi,\beta,\varepsilon,\mu) = 0.$$

First one is solved via Implicit Function Theorem, giving the existence of positive constants ε_1, r_1 and a unique C^{r-1}-function $\xi(\beta,\varepsilon,\mu)$ such that $\mathcal{Q}_\beta F(x_0(\beta)+\xi(\beta,\varepsilon,\mu),\beta,\varepsilon,\mu) = 0$ for any $\beta \in B(\beta_0,r_1)$, $\varepsilon \in (-\varepsilon_1,\varepsilon_1)$ and $\mu \in \mathbb{R}^p$. Moreover $\xi(\beta,0,\mu) = 0$, since $\mathcal{Q}_\beta F(x_0(\beta),\beta,0,\mu) = 0$. The second one is the bifurcation equation

$$G(\beta,\varepsilon,\mu) := \mathcal{P}_\beta F(x_0(\beta)+\xi(\beta,\varepsilon,\mu),\beta,\varepsilon,\mu) = 0$$

for β close to β_0 and parameters $\varepsilon \in (-\varepsilon_1,\varepsilon_1)$, $\mu \in \mathbb{R}^p$. We note [4] that there exists an orthogonal basis $\{\psi_1(\beta),\dots,\psi_k(\beta)\}$ of $Y_\beta^\perp$, i.e. $Y_\beta^\perp = [\psi_1(\beta),\dots,\psi_k(\beta)]$ for each $\beta \in W$ and ψ_i are C^{r-1}-smooth. It allows us to rewrite the projection $\mathcal{P}_\beta$ into a form

$$\mathcal{P}_\beta y = \sum_{i=1}^{k} \frac{\langle y,\psi_i(\beta)\rangle \psi_i(\beta)}{\|\psi_i(\beta)\|^2}.$$

Using this notation, we get

$$G_\varepsilon(\beta,0,\mu) = -\sum_{i=1}^{k} \frac{\langle P_\varepsilon(x_0(\beta),\beta,0,\mu),\psi_i(\beta)\rangle\, \psi_i(\beta)}{\|\psi_i(\beta)\|^2}.$$

Linear independency of $\psi_1(\beta),\dots,\psi_k(\beta)$ and Lemma 3.2 imply that

$$M^\mu(\beta) = 0 \quad \text{if and only if} \quad G_\varepsilon(\beta,0,\mu) = 0,$$

where

$$M^\mu(\beta) = (M_1^\mu(\beta), \dots, M_k^\mu(\beta)),$$

$$M_i^\mu(\beta) = \int_0^{T^\beta} \langle g(\gamma(\beta,t),0,\mu), A^*(\beta,t)\psi_i(\beta)\rangle \, dt, i = 1,\dots,k, \tag{21}$$

is a discontinuous Melnikov function and

$$A^*(\beta,t) = \begin{cases} X_1^{-1*}(\beta,t)X_1^*(\beta,t_1^\beta)S_1^*(\beta)X_2^*(\beta,t_2^\beta)S_2^*(\beta)X_3^*(\beta,T^\beta) & \text{if} \quad t \in [0,t_1^\beta), \\ X_2^{-1*}(\beta,t)X_2^*(\beta,t_2^\beta)S_2^*(\beta)X_3^*(\beta,T^\beta) & \text{if} \quad t \in [t_1^\beta,t_2^\beta), \\ X_3^{-1*}(\beta,t)X_3^*(\beta,T^\beta) & \text{if} \quad t \in [t_2^\beta,T^\beta]. \end{cases} \tag{22}$$

For A^* we derived in [3] the next result.

Lemma 3.3 *Let $A(t) \in C([0,T],L(\mathbb{R}^n))$, $B_1,B_2,B_3 \in L(\mathbb{R}^n)$, $0 < t_1 < t_2 < T$ and $h \in \mathcal{C} := C([0,t_1],\mathbb{R}^n) \cap C([t_1,t_2],\mathbb{R}^n) \cap C([t_2,T],\mathbb{R}^n)$. Then the nonhomogeneous problem*

$$\begin{gathered} \dot{x} = A(t)x + h(t), \\ x(t_i+) = B_i x(t_i-),\ i = 1,2, \\ B_3(x(T) - x(0)) = 0 \end{gathered} \tag{23}$$

has a solution $x \in \mathcal{C}^1 := C^1([0,t_1],\mathbb{R}^n) \cap C^1([t_1,t_2],\mathbb{R}^n) \cap C^1([t_2,T],\mathbb{R}^n)$ if and only if

$$\int_0^T \langle h(t), v(t)\rangle dt = 0$$

for any solution $v \in \mathcal{C}^1$ of the adjoint system given by

$$\begin{gathered} \dot{v} = -A^*(t)v, \\ v(t_i-) = B_i^* v(t_i+),\ i = 1,2, \\ v(T) = v(0) \in \mathcal{N}B_3^\perp. \end{gathered} \tag{24}$$

Therefore, the adjoint variational system of (2) is given by

$$\begin{gathered} \begin{array}{lll} \dot{X} = -\mathrm{D}f_+^*(\gamma(\beta,t))X & \text{if} & t \in [0,t_1^\beta], \\ \dot{X} = -\mathrm{D}f_-^*(\gamma(\beta,t))X & \text{if} & t \in [t_1^\beta,t_2^\beta], \\ \dot{X} = -\mathrm{D}f_+^*(\gamma(\beta,t))X & \text{if} & t \in [t_2^\beta,T^\beta], \end{array} \\ X(t_i^\beta+) = S_i^*(\beta)^{-1}X(t_i^\beta-),\ i = 1,2, \quad X(T^\beta) = X(0) \in [f_+(x_0(\beta))]^\perp. \end{gathered} \tag{25}$$

One can show that for each $i = 1,\dots,k$, $A^*(\beta,t)\psi_i(\beta)$ is the solution of (25). Consequently, we can take in (21) any basis of T^β-periodic solutions of the adjoint variational equation (25). We obtain the sufficient condition for the existence of a unique periodic solution of (2) with $\varepsilon \neq 0$:

Theorem 3.4 *Let conditions* H1'), H2'), H3') *be satisfied and $M^\mu(\beta)$ be defined by* (21). *If $\beta_0 \in V$ is a simple root of M^{μ_0}, i.e.*

$$M^{\mu_0}(\beta_0) = 0, \qquad \det \mathrm{D}M^{\mu_0}(\beta_0) \neq 0,$$

then there exists a neighborhood U *of the point* $(0, \mu_0)$ *in* $\mathbb{R}\times\mathbb{R}^p$ *and a* C^{r-2}*-function* $\beta(\varepsilon, \mu)$*, with* $\beta(0, \mu_0) = \beta_0$*, such that perturbed equation* (2) *possesses a unique persisting closed trajectory. Moreover, it contains a point*

$$x^*(\varepsilon, \mu) := x_0(\beta(\varepsilon, \mu)) + \xi(\beta(\varepsilon, \mu), \varepsilon, \mu) \in \Sigma_{\beta(\varepsilon,\mu)} \tag{26}$$

and has a period $t_3(0, x^*(\varepsilon, \mu), \beta(\varepsilon, \mu), \varepsilon, \mu)$*.*

In the special case of the manifold of initial points when $k = n-1$, i.e. $x_0(V)$ is an immersed submanifold of codimension 1, we can use slightly another approach. We can suppose that $x_0(\beta) = (\beta_1, \dots, \beta_{n-1}, 0) = (\beta, 0)$. Let us denote $\bar{\xi} = (\xi, 0)$ for $\xi \in V$. We take $\Sigma = \mathbb{R}^{n-1} \times \{0\}$ and a new Poincaré mapping $\widetilde{P} : V \times (-\varepsilon_0, \varepsilon_0) \times \mathbb{R}^p \to \Sigma$ defined as

$$\begin{aligned}\widetilde{P}(\xi, \varepsilon, \mu) = x_+(t_2(0, \bar{\xi}, \varepsilon, \mu), x_-(t_1(0, \bar{\xi}, \varepsilon, \mu), x_+(0, \bar{\xi})(t_1(0, \bar{\xi}, \varepsilon, \mu), \varepsilon, \mu))\\ (t_2(0, \bar{\xi}, \varepsilon, \mu), \varepsilon, \mu))(t_3(0, \xi, \varepsilon, \mu), \varepsilon, \mu),\end{aligned}$$

where $t_3(\cdot, \cdot, \cdot, \cdot)$ is a solution close to T^ξ of equation

$$\begin{gathered}\langle x_+(t_2(0, \bar{\xi}, \varepsilon, \mu), x_-(t_1(0, \bar{\xi}, \varepsilon, \mu), x_+(0, \bar{\xi})\\ (t_1(0, \bar{\xi}, \varepsilon, \mu), \varepsilon, \mu))(t_2(0, \bar{\xi}, \varepsilon, \mu), \varepsilon, \mu))(t, \varepsilon, \mu), e_n\rangle = 0\end{gathered}$$

with $e_n = (0, \dots, 0, 1) \in \mathbb{R}^n$. Moreover $t_3(0, \xi, 0, \mu) = T^\xi$. Then one can easily derive

$$\begin{aligned}\widetilde{P}_\xi(\xi, 0, \mu) &= (\mathbb{I} - T_\xi)A(\xi, 0),\\ \widetilde{P}_\varepsilon(\xi, 0, \mu) &= (\mathbb{I} - T_\xi)\left(\int_0^{T^\xi} A(\xi, s)g(\gamma(\xi, s), 0, \mu)ds\right)\end{aligned}$$

with $T_\xi u = \frac{\langle u, e_n\rangle f_+(\bar{\xi})}{\langle f_+(\bar{\xi}), e_n\rangle}$ and A given by (19). Note $\widetilde{P}(\xi, 0, \mu) = \bar{\xi} \in \Sigma$ and

$$\widetilde{P}(\xi, \varepsilon, \mu) = \bar{\xi} + \varepsilon(\mathbb{I} - T_\xi)\int_0^{T^\xi} A(\xi, t)g(\gamma(\xi, t), 0, \mu)dt + O(\varepsilon^2).$$

Consequently, we have the following theorem:

Theorem 3.5 *Let conditions* H1'), H2'), H3') *be satisfied and* $k = n-1$*. Let there be* $(\mu_0, \xi_0) \in \mathbb{R}^p \times V$ *such that* $M^{\mu_0}(\xi_0) = 0$ *and* $\det \mathrm{D}M^{\mu_0}(\xi_0) \neq 0$*, where*

$$M^{\mu_0}(\xi) = \left[(\mathbb{I} - T_\xi)\int_0^{T^\xi} A(\xi, t)g(\gamma(\xi, t), 0, \mu_0)dt\right]_{\mathbb{R}^{n-1}}$$

and the lower index $\mathbb{R}^{n-1}$ *denotes the restriction on first* $n-1$ *coordinates. Then there is a unique periodic solution* $x^*(\varepsilon, \mu_0, t)$ *of* (2) *with* $\mu = \mu_0$ *near* $\gamma(\xi_0, t)$*. Moreover, if*

1. $\Re\sigma(\mathrm{D}M^{\mu_0}(\beta_0)) \subset (-\infty, 0)$ *then* $x^*(\varepsilon, \mu_0, t)$ *is stable,*

2. $\Re\sigma(\mathrm{D}M^{\mu_0}(\beta_0)) \cap (0,\infty) \neq \emptyset$ *then* $x^*(\varepsilon,\mu_0,t)$ *is unstable,*

3. $0 \notin \Re\sigma(\mathrm{D}M^{\mu_0}(\beta_0))$ *then* $x^*(\varepsilon,\mu_0,t)$ *is hyperbolic with the same hyperbolicity type as* $\mathrm{D}M^{\mu_0}(\xi_0)$.

Acknowledgement

First author was partially supported by Grants VEGA-MS 1/0098/08, VEGA-SAV 2/7140/27, APVV-0414-07. The second one was supported by Grant VEGA-SAV 2/7140/27.

References

[1] DEIMLING, K.: *Nonlinear Functional Analysis.* Springer-Verlag, Berlin 1985.
[2] FEČKAN, M., POSPÍŠIL, M.: *On the bifurcation of periodic orbits in discontinuous systems.* In Commun. Math. Anal., Vol. 8, No. 1, pp. 87-108, 2010.
[3] FEČKAN, M., POSPÍŠIL, M.: *Bifurcation from family of periodic orbits in discontinuous systems: the transverse autonumous case.* submitted
[4] GOLUBITSKY, M., GUILLEMIN, V.: *Stable Mappings and Their Singularities.* Springer-Verlag, New York 1973.
[5] HARTMAN, P.: *Ordinary Differential Equations.* John Wiley & Sons, Inc., New York 1964.

Current addresses

Michal Fečkan, prof. RNDr. DrSc.
Department of Mathematical Analysis and Numerical Mathematics, Comenius University
Mlynská dolina, 842 48 Bratislava, Slovakia
and
Mathematical Institute, Slovak Academy of Sciences
Štefánikova 49, 814 73 Bratislava, Slovakia
e-mail: Michal.Feckan@fmph.uniba.sk

Michal Pospíšil, Mgr.
Mathematical Institute, Slovak Academy of Sciences
Štefánikova 49, 814 73 Bratislava, Slovakia
e-mail: Michal.Pospisil@mat.savba.sk

Journal of Applied Mathematics

volume IV (2011), number I

TESTING HESTON MODEL CONSISTENCY AND EVALUATION OF PARAMETERS THOUGHT REPRESENTATION IN DISCRETE TIME

FJODOROVS Jegors, (LV), MATVEJEVS Andrejs, (LV), PAVLENKO Oksana, (LV)

Abstract. A methodology for the estimation of parameter of a stochastic model using discontinuous models (ARIMA class) and based on the financial market data is introduced. This approach helps to simplify financial derivative pricing problems under various underlying stochastic processes. We show how to apply our technique to the financial index VIX - a market mechanism that measures the 30-day forward implied volatility of the underlying index, the S&P 500. Also the results with regression model of time series which produced by Heston volatility model are considered.

Key words: diffusion processes, time series, VIX index, CIR model, Heston model, ARIMA, ARCH.

Mathematics Subject Classification: 60J70, 62M10, 47D07

1. Introduction

The purpose of this paper is to give a reader an overview of the ARIMA techniques which could be useful in the case of the continuous Markov process representation in discrete time. As we know, every continuous Markov process can be considered as a limiting case of a discontinuous Markov process and that the solutions of the Kolmogorov diffusion equations can be approximated by solutions of the Kolmogorov differential equations [1].

For further discussion, let take a look to the Heston volatility model. This approach is adopted to give the reader an intuitive understanding of the Heston model, rather than an overly technical one, so that the sections that follow are easily absorbed. If further technical details are desired, the reader is directed to the relevant references.

(Heston 1993) proposed the following the model:

$$dS_t = \mu S_t dt + \sqrt{V_t}\, S_t dW_t^1, \tag{1}$$

$$dV_t = k(\theta - V_t)dt + \sigma\sqrt{V_t}\,dW_t^2, \tag{2}$$
$$dW_t^1 dW_t^2 = \rho dt,$$

where $\{S_t\}_{t\geq 0}$ and $\{V_t\}_{t\geq 0}$ are the price and volatility processes, respectively, and $\{W_t^1\}_{t\geq 0}$, $\{W_t^2\}_{t\geq 0}$ are correlated Brownian motion processes (with correlation parameter ρ), μ is deterministic risk free interest rate. $\{V_t\}_{t\geq 0}$ is a square root mean reverting process, first used by (Cox, Ingersoll & Ross 1985 [3]), with long-run mean θ, σ standard deviation and rate of reversion k. ρ is referred to as the volatility of diffusion. All the parameters $k, \mu, \theta, \rho, \sigma$ are time and state homogenous.
ρ , which can be interpreted as the correlation between the log-returns and volatility of the asset, affects the heaviness of the tails. ρ > 0, then volatility will increase as the asset price/return increases. This will spread the right tail and squeeze the left tail of the distribution creating a fat right-tailed distribution.
Conversely, if ρ<0, then volatility will increase when the asset price/return decreases, thus spreading the left tail and squeezing the right tail of the distribution
creating a fat left-tailed distribution (emphasizing the fact that in most cases equity returns and its related volatility are negatively correlated). Besides, ρ affects of the skew of the distribution.

There are many economic, empirical, and mathematical reasons for choosing a model with such a form (see [2] for a detailed statistical/ empirical analysis). Empirical studies have shown that an asset's log-return distribution is non-Gaussian. It is characterised by heavy tails and high peaks (leptokurtic). There is also empirical evidence and economic arguments that suggest that equity returns and implied volatility are negatively correlated (also termed 'the leverage effect'). This departure from normality plagues the Black-Scholes-Merton model with many problems. In contrast, Heston's model can imply a number of different distributions.
A contingent claim is dependent on one or more *tradable* assets in the Black-Scholes-Merton model. The randomness in the option value is solely due to the randomness of these assets. Since the assets are tradable, the option can be hedged by continuously trading the underlying. This makes the market complete, i.e., every contingent claim can be replicated.

In a stochastic volatility model, a contingent claim is dependent on the randomness of the asset $\{S_t\}_{t\geq 0}$ and the randomness associated with the volatility of the asset's return $\{V_t\}_{t\geq 0}$. Only one of these is tradable. Volatility is not a traded asset. This renders the market incomplete and has many implications to the pricing of options and other financial instruments [4].

Therefore we have to find a discrete representation of the Heston model and find possible ARIMA class model to show the hypothesis existence about residuals heteroscedastity using time series technique. Otherwise, we can't use Heston model for the chosen financial instrument pricing.

2. Heston model in discrete time

2.1. Discretizing of the price process

Firstly, we must find the solution for the equation (1) and make the discrete representation of this stochastic model. The following diffusion process postulated in the (1) model:

$$dS_t = \mu S_t dt + \sqrt{V_t}\, S_t dW_t \ .$$

Let's make substitution

$$y_t = \ln S_t \quad \text{and consider } V_t = const$$

The next step is to apply Ito's theorem. In particular,

$$dy_t = (\ln(S_t))' \mu S_t dt + (\ln(S_t))' \sqrt{V_t}\, S_t dW_t + \frac{1}{2}(\ln(S_t))'' V_t^2 \mu S_t^2 dt =$$

$$= \mu dt + \sqrt{V}\, dW_t - \frac{1}{2} V^2 \mu dt =$$

$$= (\mu - \frac{1}{2}V^2)dt + \sqrt{V}\, dW_t \ . \tag{3}$$

Solving equation (3) we get:

$$y_t - y_0 = \int_0^t (\mu - \frac{1}{2}V^2)dt + \int_0^t \sqrt{V}\, dW_s = (\mu - \frac{1}{2}V^2)dt + \sqrt{V}\, dW_t \ .$$

Or

$$\ln S_t = (\mu - \frac{1}{2}V^2)dt + \sqrt{V}\, dW_t \ .$$

Then

$$S_t = S_0 \exp((\mu - \frac{1}{2}V^2)t + \sqrt{V}\, W_t \ .$$

In a sense, we get option (or other financial instrument) price equation which depends from previous price level.

But, it is easy to deal with substitution $y_t = \ln S_t$ and get a recursive expression for y_t in terms of its previous value. Firstly, we evenly subdivide the interval [*0; T*] into N subinterval and let $t_i = i\frac{T}{N}$ for $t_i = i, \ldots, n$. In addition, we denote each time-step as $\Delta t = t_i - t_{i-1}$ (in practice we can take $\Delta t = 1$) and $\mathrm{W_t} \sim \varepsilon_\mathrm{t}$, where $\varepsilon_\mathrm{t} \sim \mathrm{N}(0,1)$.

In general, we have

$$y_t - y_{t-1} = (\mu - \frac{1}{2}V^2) + \sqrt{V}\, \varepsilon_t. \tag{4}$$

After evaluation of the equation (4) we can see that continuous Markov process could be represented in discrete time. The main part of the Heston model now is ARIMA (AR(1)) process. Nowadays, AR(1) process is quite enough studied and for further model building or testing purposes we can use ARIMA technique to reject or to accept model consistency with observable financial instrument's price data.

2.2 Discretizing of the volatility process

For the second part of Heston model (equation (2)) we can use the following approach. While a variety of alternatives exist for discretizing the stochastic differential equations that we are dealing with, in the case of the CIR model we can actually solve for the diffusion process explicitly [5]. To see how this is done, consider the following diffusion process postulated in the CIR model:

$$dV_t = k(\theta - V_t)dt + \sigma\sqrt{V_t}\,dW_t \ . \tag{5}$$

We now strip out the drift term and denote it as

$$Y_t = k(\theta - V_t)dt \tag{6}$$

The next step is a bit odd, but necessary, and involves pre-multiplying equation (6) by e^{kt} . This operation yields

$$e^{kt}Y_t = f(V_t) = e^{kt}(k(\theta - V_t) \tag{7}$$

This is, in fact, the trick in the derivation. We have a function that depends on a stochastic process, $\{V_t\}_{t\geq 0}$.
We would like to describe its differential dynamics, and to do so we may apply Ito's theorem. Before we do this, let us compute the required partial derivatives:

$$\frac{\partial f(V_t)}{\partial t} = ke^{kt}Y_t;$$

$$\frac{\partial f(V_t)}{\partial V_t} = -ke^{kt};$$

$$\frac{\partial f(V_t)}{\partial V_t^2} = 0.$$

We now have everything that we need to apply Ito's theorem. In particular,

$$f(V_t) - f(V_0) = \int_0^t \frac{\partial f(V_t)}{\partial t}ds + \int_0^t \frac{\partial f(V_t)}{\partial V_t}dV_s + \frac{1}{2}\int_0^t \frac{\partial f(V_t)}{\partial V_t^2}ds =$$

$$= \int_0^t ke^{ks}Y_s ds - \int_0^t ke^{ks}[Y_s ds + \sigma dW_s] = -\int_0^t ke^{ks}\sigma dW_s. \tag{8}$$

Inspection of equation (8) reveals that we have a recursive expression for V_t in terms of its previous value, V_0. A bit of manipulation will make this clearer:

$$e^{kt}Y_t - e^{kt}Y_0 = -\int_0^t ke^{ks}\sigma dW_s;$$

$$ke^{kt}(\theta - V_t - k(\theta - V_0)) = -\int_0^t ke^{ks}\sigma dW_s;$$

$$-ke^{kt}V_t = -k\theta e^{kt} + k\theta - kV_0 - \int_0^t ke^{ks}\sigma dW_s;$$

$$V_t = \theta(1-e^{kt}) + e^{kt}V_0 + \int_0^t ke^{ks}\sigma dW_s. \tag{9}$$

In a sense, we are finished, as we have a recursive expression for V_t in terms of its previous value. Firstly, we evenly subdivide the interval [*0; T*] into *N* subinterval and let $t_i = i\frac{T}{N}$ for $t_i = i,...,n$. In addition, we denote each time-step as $\Delta t = t_i - t_{i-1}$.

In general, we have

$$V_{t_i} = \theta(1-e^{-k\Delta t}) + e^{-k\Delta t}V_{t_{i-1}} + \varepsilon_{t_i}, \tag{10}$$

where

$$\varepsilon_{t_i} = \int_{t_{i-1}}^{t_i} ke^{-k(t_i-s)}\sigma dW_s.$$

In other words, we have the first two moments of the Gaussian transition density of V_t .

Specifically,

$$V_{t_i} / F_{t_{i-1}} \sim \mathrm{N}(\theta(1-e^{kt}) + e^{kt}V_{t_{i-1}}, \varepsilon_{t_i}^2).$$

All that remains, to get this expression into a form that can aid us in our simulations, is to find a more convenient way to express ε_{t_i}. In fact, it might not yet be obvious that ε_{t_i} is actually the variance of our transition density. This is true by virtue of the fact that ε_{t_i} is a stochastic integral and, as such, it has a zero expectation. We also recall that the quadratic variation process of the Brownian motion (i.e., W_t)
is t. In particular, this means that

$$E[\varepsilon_{t_i} / F_{t_{i-1}}] = 0.$$

With the CIR model, however, the transition density follows a non-central χ^2-quared distribution, which is rather difficult to handle. Fortunately, Ball and Torous (1996) show that, over small time intervals, diffusions arising from stochastic differential equations behave like Brownian motion and, thus, to assume a normal transition density is probably a good approximation. Thus, for the purposes of simulation we can use the first two moments of the non-central χ^2-squared distribution and assume that

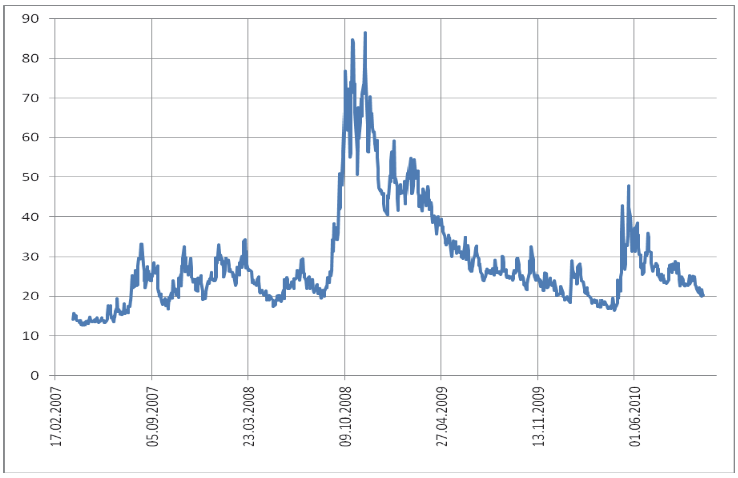
$$V_{t_i} / F_{t_{i-1}} \sim \mathrm{N}[\theta(1-e^{kt}) + e^{kt}V_{t_{i-1}}, \frac{\theta\sigma^2}{2k}(1-e^{kt})^2 + \frac{\sigma^2}{k}(e^{-kt} - e^{-2kt})V_{t_{i-1}}].$$

Taking into account, that studying financial instruments which have χ^2-quared distributions could take a lot of time and lead to unpredictable results, the situation could be improved if we used regressions (autoregressions).

This fact allows us to capture achieved results with regression model of time series which produced by Heston volatility model.

3. Time series estimation based on ARMA models

Let consider one example of combine of previous methods and ARIMA techniques. We'll analyze the VIX - **Market Volatility Index** – daily data from 27.03.2007. to 21.10.2010. The VIX is a market mechanism that measures the 30-day forward implied volatility of the underlying index, the S&P 500. Being able to meaningfully interpret movements in the VIX and its reaction to market events can give investors an edge in managing the risk and profitability of their trading book and in designing portfolio strategies using VIX derivatives to capitalize on the dynamic and time-varying correlation of the VIX with its underlying S&P 500 Index.



Graph 1 . Market Volatility Index – daily data from 27.03.2007. to 21.10.2010.

One can see from the graph 1 that the series seem to be nonstationary, especially in short time consideration. Besides, we observe some change of tendencies at the begin of October, 2008. The correlogram of the whole taken series shows us that autocorrelations decay slowly, this is typical for nonstationary series. And the augmented Dickey-Fuller test states presence of the unit root. So in present form whole data from 27.03.2007 to 21.10.2010 cannot be used for the stationary model construction and forecasting.

We can analyse first differences of our data. These new series DVIX seem to be conditionally heteroscedastic. They have not got any unit root, as shows Dickey-Fuller test. And the correlogram hasn't slowly decayed autocorrelations. We have chosen an appropriate model AR(1) with ARCH(2) residuals. But the stability test (Chow Breakpoint Test) certifies changing of the model about October, 2008.

So we decided to analyse the data after October, 2008 and tried to do it for the initial series. Series have clearly defined trend in this period.

So we exclude the time trend and analyse new series vix2 during the period from 10.10.2008 to 21.10.2010:

$$vix2_t = vix_t - 0.0450460613 \cdot t\,.$$

The system of equations (4),(13) which was received by discretezation of Heston model approximately corresponds to the AR(1) model with GARCH-M(1,0) residuals, where the squared variance with the fixed coefficient 0.5 is included in AR(1) equation for $vix2_t$.

We are successfully estimated this model for time series $vix2_t$ with the programm WINRATS and received the system in the following form:

$$\begin{cases} vix\,2_t = -16.743 + 0.5\,V_t^2 + 0.954 \cdot vix\,2_{t-1} + \varepsilon_t \\ \varepsilon_t = v_t \sqrt{V_t} \\ V_t = 5.9934 + 0.04138979 \cdot V_{t-1} \end{cases}.$$

Moreover, the coefficient at V_t^2 was fixed, i.e. it is equaled by 0.5. Unfortunately, we couldn't assert that this is the best model for our data. At first, the model is appeared sensible to initial data. Secondly, if we decided to give up fixing of the coefficient at V_t^2, the actual model will appear a little different, and other coefficients will change .

$$\begin{cases} vix\,2_t = -10.3265 + 0.24347\,V_t^2 + 0.95319 \cdot vix\,2_{t-1} + \varepsilon_t \\ \varepsilon_t = v_t \sqrt{V_t} \\ V_t = 6.7326 + 0.0849 \cdot V_{t-1} \end{cases}.$$

We can assume that some other ARMA / GARCH model will be better than this one. We have considered many similar models and the above-mentioned model indeed appears the most suitable model for our data.

Thus, the regression analysis gives us more possibilities for the analysis of such data, but the model of Heston allows us to narrow down the scopes of search of suitable model, that conversely is a pretty labor intensive process.

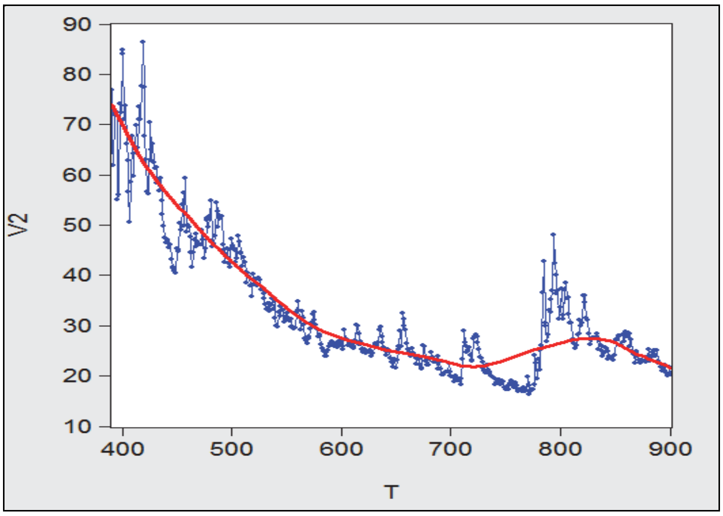
Graphically the estimation problem can be also successfully decides in the package of E-views by the method of local regression by the direct choice of suitable kernel and evaluation of standard deviation of realization of initial row from a hypothetical curve (this curve is represented as a red color curve in the picture).

Conclusions

A discrete representation of Heston stochastic model was applied to solve the problem of stochastic model consistency with financial time series data. Model discretization shows that evaluation of parameters using ARIMA technique is quite easier. However, VIX data simulation as a result forward option price evaluation via Heston model is not the best idea. That is why above mentioned approach can be used for different class of diffusion models as a quick test of the model consistency.

References

[1] Albert T. BHARUCHA-REID: *Elements of the Theory of Markov Processes and Their Applications*, Courier Dover Publications, 1997.



Graph 2. The time series *vix2$_t$* estimation with method of local regression

[2] John C. COX; Johathan E.INGERSOLL, Jr,; Stephen A .ROSS: *A Theory of the Term Structure of Interest Rates*, Econometrica, Vol. 53, No 2.

[3] CONT,R.: *Empirical properties of asset returns: stylized facts and statistical issues*, Quantative Finance, 2001;

[4] D.J.BOLDER: *Affine Term-Structure Models: Theory and implementations*, Bank of Canada Working Paper 2001 -15.

[5] A.SEPP: *Pricing Options on Realized Variance in the Heston Model with Jumps in Returns and Volatility*, Journal on Computational Finanace, 2008, 11(4), 33-70.

[6] N.MOODLEY: *The Heston Model: a Practical Approach with Matlab Code*, University of the Witwatersrand, Johannesburg, 2005.

[7] W.FELLER: *Two Singular Diffusion Problems*, Annals of Mathematics, 1951, p. 173 – 182.

Current address

Jegors Fjodorovs, Phd student
Affiliation: Department of Probability Theory and Mathematical Statistics, Riga Technical University, Address: 1 Kalku Street, LV-1658, Riga, Latvia
e-mail: jegorsf@apollo.lv

Andrejs Matvejevs, Dr.sc.ing., Professor
Affiliation: Department of Probability Theory and Mathematical Statistics, Riga Technical University, Address: 1 Kalku Street, LV-1658, Riga, Latvia
e-mail: Andrejs.Matvejevs@rtu.lv

Oksana Pavlenko, Dr.math., Docent
Affiliation: Department of Probability Theory and Mathematical Statistics, Riga Technical University, Address: 1 Kalku Street, LV-1658, Riga, Latvia
e-mail: Oksana.Pavlenko@rtu.lv

Journal of Applied Mathematics

volume IV (2011), number I

ON SOME APPLICATIONS OF FRACTIONAL CALCULUS

HALFAROVÁ Hana, (CZ), Šmarda Zdeněk, (CZ)

Abstract. In the paper, we will apply the Laplace transform to solve some fractional order differential equations in the sense of the Riemann-Liouville fractional derivative. Solutions will be described using the Mittag-Leffler functions, the Gamma function and the Mellin-Ross function.

Key words and phrases. Fractional differential equations, Riemann-Liouville fractional derivative.

Mathematics Subject Classification. Primary 26A33; Secondary 34A08.

1 Introduction

Fractional calculus owes its origin to a question of whether the meaning of a derivative to an integer order n could be extended to still be valid when n is not an integer. This question was first raised by L'Hospital in 1695. In a letter to Leibniz, he posed a question about $d^n y/dx^n$ if $n = 1/2$. Leibniz responded that it would be "an apparent paradox", from which one day useful consequences will be drawn.
In 1819 Lacroix[2] became the first mathematician to publish a paper that mentioned a fractional derivative. Starting with $y = x^m$, where M is a positive integer, Lacroix found the n^{th} derivative

$$\frac{d^n y}{dx^n} = \frac{m!}{(m-n)!} x^{m-n}, \quad n \geq n,$$

and using the Legendre's symbol Γ for the generalized factorial, he wrote

$$\frac{d^n y}{dx^n} = \frac{\Gamma(m+1)}{\Gamma(m-n+1)} x^{m-n}.$$

Finally by letting $m = 1$ and $n = 1/2$, he obtained

$$\frac{d^{1/2}}{dx^{1/2}} = \frac{2\sqrt{x}}{\sqrt{\pi}}. \tag{1}$$

The first serious attempt to give a logical definition of a fractional derivative is due to Liouville [3]. He published nine papers on the subject between 1832 and 1837, the last in the field in 1855. Liouville started in 1832 with the well known result $D^n e^{ax} = a^n e^{ax}$, where $D = d/dx,\ n \in N$, and extended it at first in the particular case $n = 1/2,\ a = 2$, and then to arbitraty order $\nu \in R^+$ by

$$D^\nu e^{ax} = a^\nu e^{ax}.$$

Over the years, many mathematicians, using their own notation and approach, have found various definitions that fit the idea of a non-integer order integral or derivative. One version that has been popularized in the world of fractional calculus is the Riemann- Liouville definition. It is interesting to note that the Riemann-Liouville definition of a fractional derivative gives the same result as that obtained by Lacroix in equation (1).

Before looking at the definition of the Riemann-Liouville fractional integral or derivative, we will first discuss some useful mathematical definitions that are inherently tied to fractional calculus and will commonly be encountered (see [1],[4],[5]). These include the Error function, the Mittag-Leffler function and the Mellin-Ross fuction.

1.1 Error function

The definition of the Error function is given by

$$Erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt, \quad x \in R.$$

The complementary Error function $Erfc$ is a closely related function that can be written in terms of the Error function as

$$Erfc(x) = 1 - Erf(x).$$

We note that $Erf(0) = 0$ and $Erf(\infty) = 1$.

1.2 Mittag-Leffler function

The Mittag-Leffler type which is defined by the series expansion

$$E_{\alpha,\beta}(x) = \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(\alpha k + \beta)}, \quad \alpha > 0,\ \beta > 0. \tag{2}$$

It follows from (2) that

$$E_{1/2,1}(x) = \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(k/2 + 1)} = e^{x^2} Erfc(-x).$$

$$E_{1,1}(x) = \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(k + 1)} = \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x.$$

$$E_{1,2}(x) = \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(k+2)} = \sum_{k=0}^{\infty} \frac{x^k}{(k+1)!} = \frac{1}{x} \sum_{k=0}^{\infty} \frac{x^{k+1}}{(k+1)!} = \frac{e^x - 1}{x}$$

and in general

$$E_{1,m}(x) = \frac{1}{x^{m-1}} \left(e^x - \sum_{k=0}^{m-2} \frac{x^k}{k!} \right).$$

For $\beta = 1$ we obtain the Mittag-Leffler function in one parameter

$$E_{\alpha,1}(x) = \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(\alpha k + 1)} \equiv E_\alpha(x).$$

Note that $E_{\alpha,\beta}(0) = 1$. Some modifications of (2) we use for a description of solutions of certain classes fractional differential equations.

1.3 The Mellin-Ross function

The Mellin-Ross function, $E_t(\nu, a)$, arises when finding the fractional integral of an exponential function. The function is closely related to both the incomplete Gamma function

$$\Gamma^*(\nu, t) = \frac{1}{\Gamma(\nu) t^\nu} \int_0^t e^{-x} x^{\nu-1} dx, \quad \mathcal{R}e\ \nu > 0$$

and Mittag-Leffler functions. Its definition is given by

$$E_t(\nu, a) = t^\nu e^{at} \Gamma^*(\nu, t).$$

We also write

$$E_t(\nu, a) = t^\nu \sum_{k=0}^{\infty} \frac{(at)^k}{\Gamma(k + \nu + 1)} = t^\nu E_{1,\nu+1}(at).$$

2 Riemann-Liouville fractional integral

Let ν be a real nonnegative number. Let f be piecewise continuous function on $(0, \infty)$ and integrable on any finite subinterval $[0, \infty]$. Then the integral

$$_cD_x^{-\nu} f(x) = \frac{1}{\Gamma(\nu)} \int_c^x (x - t)^{\nu-1} f(t) dt, \quad t > 0, \tag{3}$$

we call the Riemann-Liouville fractional integral of order ν. In the case $c = 0$ we denote (3) $D^{-\nu}$. It is important to note that some fractional integrals, even of such elementary functions as exponentials, lead to higher transcendental functions.

Example 1.

Suppose $f(t) = e^{at}$. Then by definition (3) we get

$$D^{-\nu} e^{at} = \frac{1}{\Gamma(\nu)} \int_0^t (t - y)^{\nu-1} e^{ay} dy.$$

If we make the substitution $x = t - y$ then we have

$$D^{-\nu} e^{at} = \frac{e^{at}}{\Gamma(\nu)} \int_0^t x^{\nu-1} e^{-ax} dx = E_t(\nu, a) = t^{\nu} E_{1,\nu+1}(at).$$

In particular, if $\nu = 1/2$,

$$D^{-1/2} e^{at} = E_t(1/2, a) = a^{-1/2} e^{at} Erf(at)^{1/2}.$$

3 Riemann-Liouville fractional derivative

The fractional derivative can be defined using the definition of the fractional integral. Suppose that $\nu = n - u$, where $0 < \nu < 1$ and n is the smallest integer greater then u . Then the fractional derivative of $f(x)$ of order u is

$$D^u f(x) = D^n [D^{-\nu} f(x)]. \tag{4}$$

Example 2.

Suppose we want to find the fractional derivative of $f(x) = x^s,\ s \geq 0$ of order ν. With respect to definition (4) we need to interchange u and ν. Let $u = n - \nu$ where $0 < u < 1$. Then we have $n = 1$ and $u = 1 - \nu$. From here we obtain

$$\begin{aligned} D^{\nu} x^{\mu} &= D^1 [D^{-(1-\nu)} x^{\mu}] = D^1 \left[\frac{\Gamma(\mu+1)}{\Gamma((\mu-\nu+1)+1)} x^{\mu-\nu+1} \right] \\ &= (\mu-\nu+1) \frac{\Gamma(\mu+1)}{(\mu-\nu+1)\Gamma(\mu-\nu+1)} x^{\mu-\nu} = \frac{\Gamma(\mu+1)}{\Gamma(\mu-\nu+1)} x^{\mu-\nu}. \end{aligned}$$

In particular, if $\nu = 1/2,\ \mu = 0$ we get

$$D^{1/2} x^0 = \frac{\Gamma(1)}{\Gamma(1/2)} = \frac{1}{\sqrt{\pi x}}.$$

4 The Laplace transform, fractional differential equations

Let us recall some basic facts about the Laplace transform.

The function $Y(p)$ of the complex variable p defined by

$$Y(p) = \mathcal{L}\{y(t)\} = \int_0^{\infty} e^{-pt} y(t) dt \tag{5}$$

is called the Laplace transform of the function $y(t)$. for the existence of integral (5) the function $y(t)$ must be of exponential order α, which means that there exist positive constants M, T such that

$$e^{-\alpha}|y(t)| \leq M, \quad t > T.$$

The fractional integral of $y(t)$ of order ν is

$$D^{-\nu}y(t) = \frac{1}{\Gamma(\nu)}\int_0^t (t-s)^{\nu-1}y(s)ds, \quad \nu > 0. \tag{6}$$

Equation (6) is actually a convolution integral. So, we obtain

$$\mathcal{L}\{D^{-\nu}y(t)\} = \frac{1}{\Gamma(\nu)}\mathcal{L}\{t^{\nu-1}\}\mathcal{L}\{y(t)\} = p^{-\nu}Y(p), \quad \nu > 0.$$

For example

$$\mathcal{L}\{D^{-\nu}t^{\mu}\} = \frac{\Gamma(\mu+1)}{p^{\mu+\nu+1}}, \quad \mathcal{L}\{D^{-\nu}e^{at}\} = \frac{1}{p^{\nu}(p-a)}$$

for $\nu > 0,\ \mu > -1$. We know that the fractional derivative of $y(t)$ of order ν is

$$D^{\nu}y(t) = D^n[D^{-u}y(t)] = D^n[D^{-(n-\nu)}y(t)],$$

where n is the smallest integer greater than $\nu > 0$ and $u = n - \nu$. Now we recall that in the integer order operations the Laplace transform of $y^{(n)}$ is given by

$$\mathcal{L}\{y^{(n)}(t)\} = p^nY(p) - \sum_{k=0}^{n-1} p^{n-k-1}y^{(k)}(0). \tag{7}$$

Suppose that the Laplace transform of $y(t)$ exists, then using (7) we obtain

$$\begin{aligned}
\mathcal{L}\{D^{\nu}y(t)\} &= \mathcal{L}\{D^n[D^{-(n-\nu)}y(t)]\} \\
&= p^n\mathcal{L}\{D^{-(n-\nu)}y(t)\} - \sum_{k=0}^{n-1} p^{n-k-1}D^k[D^{-(n-\nu)}y(t)]_{t=0} \\
&= p^n[p^{-(n-\nu)}Y(p)] - \sum_{k=0}^{n-1} p^{n-k-1}D^{k-(n-\nu)}y(0) \\
&= p^{\nu} - \sum_{k=0}^{n-1} p^{n-k-1}D^{k-n+\nu}y(0).
\end{aligned}$$

For $n = 1$ and $n = 2$, we respectively get

$$\mathcal{L}\{D^{\nu}y(t)\} = p^{\nu}Y(p) - D^{-(1-\nu)}y(0), \quad 0 < \nu \leq 1, \tag{8}$$

and

$$\mathcal{L}\{D^{\nu}y(t)\} = p^{\nu}Y(p) - pD^{-(2-\nu)}y(0) - D^{-(1-\nu)}y(0), \quad 1 < \nu \leq 2. \tag{9}$$

Example 3. Consider the fractional differential equation

$$D^{1/2}y(t) + ay(t) = 0,$$

where a is a constant. In our case $0 < \nu = 2/3 < 1$, hence we will use formula (8). Then applying the Laplace transform to both sides of the equation we get

$$\mathcal{L}\{D^{1/2}y(t)\} + a\mathcal{L}\{y(t)\} = 0,$$

which implies

$$p^{2/3}Y(p) - D^{-(1-1/2)}y(0) + aY(p) = 0.$$

The constant $D^{-1/2}y(0)$ is the value of $D^{-1/2}(t)$ at $t = 0$. Denote $C = D^{-1/2}y(0)$ then

$$Y(p) = \frac{C}{p^{1/2} + a}.$$

Applying the inverse transform with the Laplace transform of Mittag-Leffler function in two parameter (see [4] p.21) we have

$$y(t) = Ct^{-1/2}E_{\frac{1}{2},\frac{1}{2}}(-a\sqrt{t}) = C\left(\frac{1}{\sqrt{\pi t}} - e^t erfc(\sqrt{t})\right).$$

Example 4. Consider the fractional differential equation

$$D^{4/3}y(t) = 0.$$

In this case $1 < \nu = 4/3 \leq 2$. Applying formula (9) with the Laplace transform to both sides of the equation we obtain

$$p^{4/3}Y(p) - pD^{-(2-4/3)}y(0) - D^{-(1-4/3)} = 0.$$

Put $c_1 = D^{-2/3}y(0)$ and $c_2 = D^{1/3}y(0)$. Then

$$Y(p) = \frac{c_1 p}{p^{4/3}} + \frac{c_2}{p^{4/3}}.$$

Using table of the inverse Laplace transform of $Y(p)$ (see [4]) we obtain

$$y(t) = \frac{c_1}{\Gamma(1/3)}t^{-2/3} + \frac{c_2}{\Gamma(4/3)}t^{1/3}.$$

Acknowledgement

This research has been supported by the Czech Ministry of Education in the frames of MSM002160503 Research Intention MIKROSYN New Trends in Microelectronic Systems and Nanotechnologies and MSM0021630529 Research Intention Inteligent Systems in Automatization.

References

[1] HABOULD, H., MATHAI, A.: *Special Functions for Applied Scientists*, Springer, 2008.

[2] LACROIX,S.F.: *Traité Du Calcul Différential et du Calcul Intégral,* 2nd ed., 3, 1819, 409-410 (Courcier Paris).

[3] LIOUVILLE,J.: *Sur le calcul de differentielles á indices quelconques,* J.École Polytech., 13, 1832, 1-69.

[4] PODLUBNY, I.: *Fractional Differential Equation*, Academic Press, 1999.

[5] ŠMARDA, Z.: *Some properties of Fractional Integrals and Derivatives,* Proceedings of the Sixth International Conference on Soft Computing Applied in Computer and Economic Enviroments, ICSC 2008, Kunovice, 229-234, ISBN 978-80-7314-134-9 229-234.

Current address

Mgr. Hana Halfarová
Department of Mathematics
Faculty of Electrical Engineering and Communication
Brno University of Technology, Technická 8, 616 00 BRNO
e-mail: xhalfa06@stud.feec.vutbr.cz

Doc. RNDr. Zdeněk Šmarda, CSc.
Department of Mathematics
Faculty of Electrical Engineering and Communication
Brno University of Technology, Technická 8, 616 00 BRNO
e-mail: smarda@feec.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

HOW TO FIND A BOUNDED SOLUTION OF THE DISCRETE ANALOGUE OF THE EMDEN-FOWLER EQUATION

HLAVIČKOVÁ Irena, (CZ)

Abstract. This contribution studies the asymptotic behavior of solutions of a special type of discrete equations, namely, of the discrete analogue of the Emden-Fowler differential equation. Sufficient conditions guaranteeing the existence of a solution of this equation which stays between two given functions have been already described. In this paper, we deal with the problem of determining corresponding initial data generating such solution. We try to find such data with help of a numerical method which resembles the well-known bisection method used for solving nonlinear equations.

Key words and phrases. Discrete equation, bounded solutions, initial data, Emden-Fowler equation.

Mathematics Subject Classification. Primary 39A10.

1 Introduction

The well-known Emden-Fowler differential equation has been investigated from many points of view, e.g. in the book [1, Chapter VII]. The special case of it is the equation

$$u''(t) - t^{-2}u^n(t) = 0$$

with $n > 1$.
Substitution

$$t = \exp(s), \quad u(s) = Cz^{\alpha}(s),$$

where

$$\alpha = -1/(n-1) \quad \text{and} \quad C^{n-1} = -\alpha,$$

leads (under the supposition $z \neq 0$) to the equation

$$z'' + \frac{(\alpha-1)(z')^2}{z} - z' + 1 = 0.$$

We will investigate the discrete analogue of this equation, i.e. the second-order difference equation

$$\Delta^2 v(k) + \frac{(\alpha-1)(\Delta v(k))^2}{v(k)} - \Delta v(k) + 1 = 0, \tag{1}$$

where $\alpha \in \mathbb{R}$, $\alpha < 0$, $k \in \mathbb{Z}_a^\infty := \{a, a+1, \dots\}$, $a \in \mathbb{N}$, and $\Delta v(k) = v(k+1) - v(k)$.
In [4], the following theorem was proved.

Theorem 1.1 *Let numbers ν_1, ν_2, $1 < \nu_1 < 2$, $0 < \nu_2 < 1$, $1 + \nu_2 > \nu_1$, be given. Then there exists a solution $v(k)$ of equation (1), such that*

$$|v(k) - k| < k \cdot \left(\frac{1}{k}\right)^{\nu_2} \tag{2}$$

and

$$\left|\Delta v(k) - 1 - \frac{\alpha-1}{k}\right| < \left(\frac{1}{k}\right)^{\nu_1} \tag{3}$$

for k sufficiently large.

The aim of our contribution is to find this solution for fixly chosen values of α, ν_1 and ν_2, i.e. to find the values of initial conditions

$$v(a) = v^a \in \mathbb{R} \setminus \{0\}, \quad v(a+1) = v^{a+1} \in \mathbb{R} \setminus \{0\} \tag{4}$$

which generate such solution.

2 Algorithm for Finding the Initial Data

We will describe an algorithm how to find a, v^a and v^{a+1} from initial conditions (4) so that the solution generated by them satisfies conditions (2) .
From the nature of the proof of Theorem 1.1 it follows (for the details, see [4]):
I) For concrete values of α, ν_1 and ν_2, the initial value of k, i.e. the a in (4), can be specified so that the conditions (2), (3) are valid for all $k \in \mathbb{Z}_a^\infty$ (i.e. we can specify the exact meaning of "k sufficiently large" in Theorem 1.1).
II) The initial value v^a can be chosen arbitrarily between $b(a)$ and $c(a)$ where for $k \in \mathbb{Z}_a^\infty$,

$$\begin{aligned} b(k) &:= k - k \cdot \left(\tfrac{1}{k}\right)^{\nu_2} \\ c(k) &:= k + k \cdot \left(\tfrac{1}{k}\right)^{\nu_2}. \end{aligned} \tag{5}$$

More exactly, for any $v^a \in (b(a), c(a))$, there exists a solution of Eq. (1) satisfying (2) and (3). Thus, for chosen values of α, ν_1 and ν_2, we first find the corresponding value of a. This is a tedious work and it is based upon conditions which arise in the proof of Theorem 1.1.

The value of v^a we choose arbitrarily, e.g. we can put $v^a := a$. For the next considerations, take this value for fixed. Now, the solution is determined only by the value of v^{a+1}.
The method of finding v^{a+1} will be similar to the well-known bisection method for solving nonlinear equations of the form $f(x) = 0$. Let us start with an interval that certainly contains the sought "root" v^{a+1} i.e. the interval $[b(a+1), c(a+1)]$. Denote

$$v_{L,1}^{a+1} := b(a+1) \quad \text{and} \quad v_{U,1}^{a+1} := c(a+1).$$

(L as "lower", U as "upper" bound). So we have the interval $[v_{L,1}^{a+1}, v_{U,1}^{a+1}]$ and, similarly as in the bisection method, we will construct a sequence of intervals $[v_{L,i}^{a+1}, v_{U,i}^{a+1}]$, $i = 1, 2, \dots$, containing the "root" v^{a+1}. The next interval will be obtained by bisecting the previous one and choosing the correct half of it.
Denote the solutions of Eq. (1) given by the initial conditions $v(a+1) = v_{L,i}^{a+1}$ and $v(a+1) = v_{U,i}^{a+1}$ as $v = v_{L,i}(k)$ and $v = v_{U,i}(k)$, respectively.
From the proof of Theorem 1.1, it follows that

$$v_{L,1}(a+2) < b(a+2) \quad \text{and} \quad v_{U,1}(a+2) > c(a+2).$$

Now we will bisect the interval $[v_{L,1}^{a+1}, v_{U,1}^{a+1}]$. Denote its center as

$$v_1^{a+1} := \frac{v_{L,1}^{a+1} + v_{U,1}^{a+1}}{2}\,.$$

Consider the solution $v = v_1(k)$ of Eq. (1) given by the initial condition $v(a+1) = v_1^{a+1}$. There are three possibilities:
I) $b(k) < v_1(k) < c(k)$ for every $k \in \mathbb{Z}_a^\infty$. In this case $v^{a+1} = v_1^{a+1}$, we have a solution with the desired property (2) and we can stop the process.
II) There exists an $r \in \mathbb{Z}_a^\infty$ such that $b(k) < v_1(k) < c(k)$ for $k = a, \dots, r-1$, but $v_1(r) \le b(r)$. In this case we set

$$v_{L,2}^{a+1} := v_1^{a+1}, \quad v_{U,2}^{a+1} := v_{U,1}^{a+1}.$$

III) There exists an $s \in \mathbb{Z}_a^\infty$ such that $b(k) < v_1(k) < c(k)$ for $k = a, \dots, s-1$, but $v_1(s) \ge c(s)$. This time we change the upper bound of the interval:

$$v_{U,2}^{a+1} := v_1^{a+1}, \quad v_{L,2}^{a+1} := v_{L,1}^{a+1}.$$

Now, either we have the desired v^{a+1}, or we have a new interval $[v_{L,2}^{a+1}, v_{U,2}^{a+1}]$ with the property that the solution $v = v_{L,2}(k)$ exceeds the lower bound $b(r)$ for some $r \in \mathbb{Z}_a^\infty$, meanwhile the solution $v = v_{U,2}(k)$ exceeds the upper bound $c(s)$ for some $s \in \mathbb{Z}_a^\infty$. Such interval has to contain a point v^{a+1} for which the corresponding solution $v = v^{a+1}(k)$ stays between $b(k)$ and $c(k)$.
Further, we will proceed inductively. Continuing this process, either we get the sought initial point v^{a+1} in a finite number of steps, or we get infinite sequences $\{v_{L,i}^{a+1}\}_{i=1}^\infty$, $\{v_{U,i}^{a+1}\}_{i=1}^\infty$ and $\{v_i^{a+1}\}_{i=1}^\infty$. These sequences are obviously convergent as $\{v_{L,i}^{a+1}\}_{i=1}^\infty$ is a nondecreasing sequence bounded from above by $c(a+1)$, $\{v_{U,i}^{a+1}\}_{i=1}^\infty$ is a nonincreasing sequence bounded from below by $b(a+1)$ and $v_{L,i}^{a+1} < v_i^{a+1} < v_{U,i}^{a+1}$ for every $i \in \mathbb{N}$. In this case, $v^{a+1} = \lim_{i\to\infty} v_i^{a+1}$.

3 Practical Implementation of the Algorithm

Programming the above described method, we are limited by the possibilities of computers. In the ideal case, we would bisect the intervals until either we find a solution with property(2), or the length of the interval $[v_{L,i}^{a+1}, v_{U,i}^{a+1}]$ is less than some chosen $\varepsilon > 0$. But, practically, for given initial conditions (4), we can compute the values of the corresponding solution of Eq. (1) for $k = a, a+1, \dots,$ but it is clear that it is impossible to compute to infinity. We have to stop sometimes. Thus, given a fixed $n \in \mathbb{Z}_a^\infty$, we are able to find the value $\tilde{v}^{a+1}$ such that the corresponding solution satisfies the conditions (2) and (3) for $k = a, \dots, n$ but not necessarily for all $k \in \mathbb{Z}_a^\infty$.

We present a program written in Maple for solving the above described problem.
As one can easily see, Eq. (1) can be rewritten as

$$v(k+2) = 3v(k+1) - 2v(k) - \frac{(\alpha-1)(v(k+1)-v(k))^2}{v(k)} - 1.$$

Thus, the function `f` in the program below serves for computing the next member of solution of Eq. (1) when we know the two previous members $v(k)$ and $v(k+1)$, in the program denoted as `u1` and `u2`, respectively.
Further, the function `b` and `c` are defined by formulas (5).
The main function, which finds the appropriate value of the initial condition v^{a+1}, is named `calculate`.

```
 > f:=(u1,u2,alpha)->3*u2-2*u1 -(alpha-1)*(u2-u1)^2/u1-1:
> b:=(k,alpha,nu2)->k-k*(1/k)^nu2:
> c:=(k,alpha,nu2)->k+k*(1/k)^nu2:
>
> calculate:=proc(alpha,nu2,a,n)
> local k, u1, u2, u1next, u2next, success, lower, upper, centre, u1a:
> lower:=b(a,alpha,nu2):
> upper:=c(a,alpha,nu2):
> u1a:=(b(a,alpha,nu2)+c(a,alpha,nu2))/2;
> success:=false:
> while (not success) do
>   success:=true:
>   centre:=(lower+upper)/2:
>   u1next:=u1a:
>   u2next:=centre:
>   printf("-------------------------\n");
>   printf("k b(k) v(k) c(k)\n");
>   printf("%d %.5f %.5f %.5f\n",a,b(a,alpha,nu2),u1next,c(a,alpha,nu2));
>   printf("%d %.5f %.5f %.5f\n",a+1,b(a+1,alpha,nu2),u2next,c(a+1,alpha,nu2));
>   for k from a to n do
>     u1:=u1next:
>     u2:=u2next:
```

```
>     u2next:=f(u1,u2,alpha):
>     u1next:=u2:
>     printf("%d %.5f %.5f %.5f\n",k+2,b(k+2,alpha,nu2),u2next,c(k+2,alpha,nu2));
>     if (u2next<b(k+2,alpha,nu2)) then
>       lower:=centre:
>       success:=false:
>       break:
>     elif (u2next>c(k+2,alpha,nu2)) then
>       upper:=centre:
>       success:=false:
>       break:
>     end if:
>   end do:
> end do:
> return centre:
> end proc:
```

4 Numerical Experiment

Example 4.1 *Consider Eq.* (1) *with*

$$\alpha = -1, \quad \nu_1 = \frac{3}{2}, \quad \nu_2 = \frac{3}{4}.$$

In this case, one can show that "k sufficiently large" in Theorem 1.1 means $k \geq 1078$*. Hence, we can choose, e.g.,* $a = 1080$ *and with help of function* `calculate` *we can find a solution of Eq.* (1) *which satisfies* (2) *for* $k = a, \ldots, n+2$ *where, e.g.,* $n = 1100$*:*

```
> calculate(-1.0,0.75,1080,1100);
```

Calling this function results in the following printout:

```
 --------------------------
k    b(k)       v(k)       c(k)
1080 1074.26734 1080.00000 1085.73266
1081 1075.26602 1080.00000 1086.73398
1082 1076.26469 1079.00000 1087.73531
1083 1077.26337 1076.00185 1088.73663
--------------------------
k    b(k)       v(k)       c(k)
1080 1074.26734 1080.00000 1085.73266
1081 1075.26602 1082.86633 1086.73398
1082 1076.26469 1087.61420 1087.73531
1083 1077.26337 1096.15158 1088.73663
--------------------------
...
```

```
-------------------------
k     b(k)       v(k)       c(k)
1080 1074.26734 1080.00000 1085.73266
1081 1075.26602 1080.99816 1086.73398
1082 1076.26469 1081.99631 1087.73531
⋮
1100 1094.24099 1099.50016 1105.75901
1101 1095.23968 1100.03350 1106.76032
1102 1096.23837 1100.10071 1107.76163
```

That means that at the first attempt, starting at $v(1080) = 1080, v(1081) = 1080$ *the solution exceeded the lower bound* $b(1083)$*, at the second attempt, it exceeded* $c(1083)$*, etc. Finally, the value* $v^{a+1} \doteq 1080.99816$ *was found for which the solution stays in the given bounds for* $k = 1080, \ldots, 1102$.

Acknowledgement

This research was supported by the Grant P201/10/1032 of the Czech Grant Agency (Prague).

References

[1] BELLMAN, R.: *Stability theory of differential equations (Internat. Ser. in Pure and Appl. Math.)*. New York-London: McGraw-Hill Book Company, XIII, 1953.

[2] DIBLÍK J., RŮŽIČKOVÁ I.: *Compulsory asymptotic behavior of solutions of two-dimensional systems of difference equations*. In Difference Equations and Discrete Dynamical Systems, Proceedings of the 9th International Conference, University of Southern California, Los Angeles, California, USA, 2–7 August 2004. World Scientific Publishing Co. Pte. Ltd, Singapore, pp. 35–49, 2005.

[3] DIBLÍK, J., ŠMARDA, Z., RŮŽIČKOVÁ, M.: *Bounded solutions of systems of dynamic equations on time scales*. In Communications of the Laufen colloquium on science, No. 4(1), pp. 1-12, 2008.

[4] DIBLÍK, J., HLAVIČKOVÁ, I.: *Asymptotic properties of solutions of the discrete analogue of the Emden-Fowler equation*. In Advances in discrete dynamical systems. Advances studies in pure mathematics. Kyoto, ISDEA, pp. 23 - 32, 2009.

[5] DIBLÍK, J., RŮŽIČKOVÁ, M., ŠMARDA, Z.: *Ważewski's method for systems of dynamic equations on time scales*. In Nonlinear Analysis: Theory, Methods & Applications, Vol. 71, pp. 1124-1131, 2009.

Current address

Irena Hlavičková, Mgr., PhD.
Department of Mathematics

Faculty of Electrical Engineering and Communication
Brno University of Technology
Technická 8
616 00 Brno
Czech Republic
tel.: +420-541143152
email: hlavicka@feec.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

THE ALLEE EFFECTS IN POPULATION AND ECONOMIC GROWTH MODELS

MUREŞAN Anton S., (RO)

Abstract. In this paper we define the Allee effects and prove how they give some influence in population and economic growth models. In discrete growth population models we have a family of unimodal maps depending on one or more parameters such that a difference equation holds. The "function" that appears in this equation is positive, the values at the ends of interval are zero and the derivative of this function is continuous and has only one root for all values of the argument.

Key words and phrases. Allee effects, Ricker population model, stabilization, profit rate, organic composition of the capital, exploitation rate

Mathematics Subject Classification. 34K05, 34K15

1 Introduction

In discrete growth population models we have a family of unimodal maps (that has only one critical point) depending on one or more parameters such that the following equation

$$x_{n+1} = f(x_n)$$

holds. Here, the "function" f is continuous on the interval $[0, +\infty)$ and has the following properties:

$$f(x) \geq 0 \quad \text{for} \quad \text{all} \quad x \geq 0,$$
$$f(0) = f(\infty) = 0,$$
$$f' \quad \text{is} \quad \text{continuous} \quad \text{and} \quad \text{has} \quad \text{only} \quad \text{one} \quad \text{root}.$$

Therefore the function f attains its maximum at a unique positive critical point x_0 and thus has an upper limit.

We have $f(x) < f(x_0)$ for all $x \geq 0$ and $x \neq x_0$.

If $f'(0) < 1$ the origin is a locally asymptotically stable fixed point, and if $f'(0) > 1$, is unstable.

Remark 1.1 *In some models (see for example the discrete Ricker population model [10]) we have, moreover, that* $f'(0) = 0$, *and so the origin is a trivial fixed point and it is locally asymptotically stable.*

Remark 1.2 *If* $f'(0) = 1$ *and* $f"(0) \neq 0$ *then the origin is unstable (see [5], [6]).*

In the classical Ricker model, a density-dependent survival function is assumed, but the birth or grove rate is assumed to be density-independent.

While this is a valid assumption in many ecological situations, there are many circumstances which lead to nonconstant density-dependent birth or grove functions, such as species encountering possible mating limitation inbreeding depression, failure to satiate predators and lack of cooperative feeding, at a lower population level. Such a phenomenon has been observed and is called an Allee effect [10].

A typical example came from experiments in which females are forcibly mated to a particular male, which leads to lower fertilization rates, in particular, when the population size is small [10].

To incorporate Allee effects into the population growth we assume that the birth rate is proportional to the population size as the population level is low, but is saturated approximately to be constant when the population size is sufficiently large.

2 The discrete Ricker population model

Let x_n be a non-overlapped population at generation n. We suppose that the dynamics of the population is governed by the discrete equation

$$x_{n+1} = b(x_n)s(x_n)x_n \tag{1}$$

where b is the per-capita birth or growth function and s is the survival function.

In the relation (1) the $b(x_n)$ is the value of growth function and $s(x_n)$ is the value of survival function for generation n respectively.

As an example of value for the survival function we can consider

$$s(x_n) = e^{-\mu - kx_n}, \tag{2}$$

where $\mu > 0$ is density independent death rate and $k > 0$ is carrying capacity parameter.

As an example of value for the growth function we can consider

$$b(x_n) = \frac{cx_n}{\theta + x_n}, \tag{3}$$

where $c > 0$, and $\theta \geq 0$ are constants.

Remark 2.1 *The constant c measures the maximal reproduction or growth rate and the ratio c/θ measures the relative growth rate as the population size is small.*

After some transformations ($r = ce^{-\mu}$ and rescaling by $y = kx_n$, and denoting $\alpha = k\theta$) the equation

$$x_{n+1} = \frac{cx_n}{\theta + x_n} x_n e^{-\mu - kx_n} \tag{4}$$

becomes

$$x_{n+1} = \frac{rx_n^2}{\alpha + x_n} e^{-x_n} := f(x_n). \tag{5}$$

We have that the function f has similar property as the nonlinear unimodal function in classical Ricker model, that is, $f(x) \geq 0$ for all $x \geq 0$, $f(0) = f(\infty) = 0$, there is a unique positive point x_0 such that $f(x) < f(x_0)$ for all $x \geq 0$ and $x \neq x_0$.

Moreover, since $f'(0) = 0$, the origin is a trivial fixed point.

We have the following existence result

Theorem 2.1 *For the model (5) there exist*

- two positive fixed points if $r > P_1(\alpha)$,
- a unique positive fixed point if $r = P_1(\alpha)$,
- no positive fixed point if $r < P_1(\alpha)$,
where P_1 is the function defined by

$$P_1(\alpha) = \frac{\sqrt{\alpha^2 + 4\alpha} + \alpha + 2}{2} e^{\frac{\sqrt{\alpha^2+4\alpha} - \alpha}{2}}. \tag{6}$$

Now we have the following stability results for positive fixed points

Theorem 2.2 *If $r = P_1(\alpha)$ then there exists a unique positive fixed point $x = \frac{\sqrt{\alpha^2+4\alpha}-\alpha}{2}$ and it is unstable. If $r > P_1(\alpha)$ then there exist two positive fixed points $x_1 < x_2$. The positive fixed point x_1 is unstable, and x_2 is locally asymptotically stable if $r < P_2(\alpha)$ and is unstable if $r > P_2(\alpha)$, where P_2 is the function defined by*

$$P_2(\alpha) = \frac{2 + \alpha + \sqrt{\alpha^2 + 8\alpha + 4}}{2 - \alpha + \sqrt{\alpha^2 + 8\alpha + 4}} e^{\frac{2-\alpha+\sqrt{\alpha^2+8\alpha+4}}{2}}. \tag{7}$$

Theorem 2.3 *If there are no positive fixed points then the trivial solution $x = 0$ is globally asymptotically stable. If there exist two positive fixed points $x_1 < x_2$, then there exists a unique point $x^* > x_0$ such that $x_1 = f(x^*)$, where x_0 is the critical point of the function f.*

3 An economical model with Allee effect

In what follows we consider the economic theory in which the profit rate r is a function of the exploitation rate e and the organic composition of the capital k.

This organic composition of the capital k is defined as the ration of the constant capital c to variable capital v. The variable capital v is defined as advance to labour, that is, total wage payment, or heuristically, $v = wL$, where w is wages and L is labour employed. So, we get the relationship

$$k = \frac{c}{v}. \tag{8}$$

The profit rate is defined as

$$r = \frac{s}{v + c}, \tag{9}$$

where s is the surplus and $v + c$ is the total advance.

The surplus s is the amount of total output produced above total advance, or

$$s = y - (v + c), \tag{10}$$

where y is the total output.

The exploitation rate e is the ration of surplus to variable capital, that is

$$e = \frac{s}{v} \tag{11}$$

(the surplus produced for every dollar/euro spent on labour). From the relationship (9) we obtain the equation

$$r = \frac{e}{1 + k}. \tag{12}$$

Let r_n be the profit rate at the time unit n.

We assume that the exploitation rate and the organic composition of the capital at time $n + 1$ depend on the profit rate at time n, that is

$$e_{n+1} = E(r_n), \qquad k_{n+1} = K(r_n) \tag{13}$$

Then the profit rate at time $n + 1$ depends on the profit rate at time n, that is

$$r_{n+1} = \frac{E(r_n)}{1 + K(r_n)}. \tag{14}$$

Remark 3.1 *In some suitable assumptions we consider that the functions E and K are given by*

$$E(r_n) = \frac{ar_n}{1 + r_n^2}, \tag{15}$$

and

$$K(r_n) = \frac{e^{r_n - b}}{(r_n + d)^2}, \tag{16}$$

where $a > 0, b > 0, d > 0$ are constants.

By using the equations (15) and (16) the equation (14), for the profit rate, becomes

$$r_{n+1} = \frac{a r_n (r_n + d)^2}{(1 + r_n^2)[(r_n + d)^2 + e^{r_n - b}]} := f(r_n). \tag{17}$$

We have the following result about the existence of positive fixed points:

Theorem 3.1 *For the model of the profit rate (17) there are two positive fixed points if $a > a_c$, a unique positive fixed point if $a = a_c$ and no positive fixed points if $a < a_c$, where a_c is a real value depending on d and b.*

Proof. To find the positive fixed point of (17) we consider the equation

$$\frac{a r (r + d)^2}{(1 + r^2)[(r + d)^2 + e^{r - b}]} = 1, \tag{18}$$

or equivalently

$$\ln(1 + r^2) + \ln[(r + d)^2 + e^{r-b}] - 2\ln(r + d) = \ln a. \tag{19}$$

Let g be the function defined by

$$g(r) = \ln(1 + r^2) + \ln[(r + d)^2 + e^{r-b}] - 2\ln(r + d).$$

We have

$$g'(r) = \frac{2r}{1 + r^2} + \frac{2(r + d) + e^{r-b}}{(r + d)^2 + e^{r-b}} - \frac{2}{r + d}. \tag{20}$$

For the equation $g'(r) = 0$ we get a unique positive critical point $r_{d,b} = r(d, b)$. We can observe that $g(r) > 0$ for all $r \geq 0$. Let a_c be the value of a for which we have $g(r_{d,b}) = \ln a_c$, that is a_c is the value for which equation (19) has a unique positive solution. If $g(r_{d,b}) < \ln a_c$ then there exist two positive solutions of the equation (19), and if $g(r_{d,b}) > \ln a_c$ then the equation (19) has no positive solutions. Corresponding to these situations we have three cases for the conclusion of above theorem.

Relative to the stability of each positive fixed point, by using the derivatives f' and f", we can obtain the following situations:

(1) Suppose that the equation (17) has only a positive fixed point $r = r_{d,b}$. Then, because $f'(r) = 1 - rg'(r)$, we have $f'(r_{d,b}) = 1$. If $f"(r_{d,b}) < 0$ it follows that $r = r_{d,b}$ is unstable (more precisely $r = r_{d,b}$ is semi-stable from the right side and is unstable from the left side if the

initial value r_0 belongs to the interval $(r_{d,b}, r_3)$, where $f(r_3) = r_1$ and $r_3 > c$, where c is the critical point of f).

(2) Suppose that the equation (17) has two positive fixed points r_1 and r_2 for the profit rate, such that $r_1 < r_2$. We have $r_1 g'(r_1) < 0$ and therefore r_1 is always unstable. For r_2 we have $g'(r_2) > 0$, and therefore $r_2 g'(r_2) > 0$. So we can consider two cases:

a) If $0 < r_2 g'(r_2) < 2$, then r_2 is locally asymptotic stable;

b) If $r_2 g'(r_2) > 2$ then r_2 is unstable.

Now, we give the following

Definition 3.1 *The model of profit rate given by relationship (17) is under an Allee effect whenever the function f has three fixed points, 0 and two positive fixed points, r_1 and r_2 such that $r_1 < r_2$ and 0 is globally attracting in $[0, r_1)$, r_1 is repellers and a positive attractor exists in the interval $(r_1, \max I)$, where $I = f([0, r_2])$.*

From this definition, it results that "function" f of profit rate model (17) belongs to a family of unimodular maps depending on the parameter a, where a measures the relative growth of the exploitation rate when the profit rate is zero. Moreover, the model predicts a period-doubling bifurcation scenario as the parameter a increases.

Remark 3.2 *In the population growths the Allee effects occur when individuals benefits from the presence of con-specifics, and as a result suffer a decline in some component of fitness when populations become small or sparse. Strong Allee effects can lead to threshold population densities, below which the population growth is negative, making extinction likely.*

Berec and all [1] make a distinction between **component Allee effects** which means a reduction in a component fitness with decreasing population density caused by a single mechanism and **demographic Allee effects,** which means that the total reduction of fitness resulting from all the different component Allee effects and respectively density dependent mechanisms affecting the population.

However, if the negatively density dependent mechanism are strong enough, it is also possible that the population has a component Allee effect but not a demographic Allee effect.

References

[1] BEREC, L., ANGULO, E., COURCHAMP, F., *Multiple Allee effects and population management,* Trends Ecol. Evol., 22(2007), 185-191

[2] CELIK, C., MERDAN, H, DUMAN, O., AKIN, O., *Allee effects on population dynamics with delay,* Chaos, Solitons and Fractals, 37 (2008), 65-74

[3] DUMAN, O., MERDAN, H. *Stability of continuous population model involving predation and Allee effect,* Chaos, Solitons and Fractals, 41(2009), 1218-1222

[4] ELAYDI, S., *An introduction to difference equations, Third ed., Springer, New York, NY, 2005*

[5] ELAYDI, S., *Discrete Chaos*, Second ed. Chapman & Hall/CRC, Boca Raton, Fl. 2008

[6] ESKOLA, H.T.M., PARVINEN K., *The Allee effect in mechanistic models based on inter-individual processes,* Bull. Math. Biology, 72(2010), 184-207

[7] HILKER, F.M., *Population collapse to extinction: the catastrophic combination of parasitism and Allee effect,* J. Biological Dynamics, 4, 1(2010), 86-101

[8] LI, J., SONG, B. WANG, X., *An extended discrete Ricker population modl with Allee effects,* J. Diff. Eq. Appl., 13, 4(2007), 309-321

[9] LOPEZ-RUIZ, R., FOURNIER-PRUNARET, D., *Indirect Allee effect, bistability and chaotic oscillations in a predator-prey discrete model of logistic type,* Chaos, Solitons and Fractals, 24(2005), 85-101

[10] LUIS, R., OLIVEIRA, H., *A dynamical system for a long term economical model*, Preprint, Proc., Conf., ICDEA 2007, Lisbon, 2007

[11] LUIS, R., ELAYDI, S., OLIVEIRA, H., *An economic model with Allee effect*, J. Diff. Eq. Appl., 15, 8-9(2009), 877-894

[12] MUREŞAN, A.S., *The Solow-Swan growth model with AK Technology and delayed population change*, ACAM, 19, (2010), no. 2, 135-143

[13] SUN, Y.G., SAKER, S.H., *Existence of positive periodic solutions on nonlinear discrete model exhibiting the Allee effect,* Appl. Math. Comp., 168(2005), 1086-1097

[14] WANG, M-H., KOT, M., NEUBERT, M.G., *Integrodifferential equations, Allee effects, and invasions* , J. Math. Biol., 44(2002), 150-168

[15] ZHOU, S.R., LIU, Y.F., WANG, G., *The stability of predator-prey systems subject to the Allee effects,* Theor. Population Biol. 67(2005), 23-31

[16] YAN, J., ZHAO, A., YAN, W., *Existence and global attractivity of periodic solution for an impulsive delay differential equation with Allee effect*, J. Math. Anal. Appl., 309(2005), 489-504

Current address

Anton S. MUREŞAN
Department of Statistics, Forecasting and Mathematics,
Faculty of Economics and Business Administration
Babeş-Bolyai University of Cluj-Napoca, 58-60 T. Mihali Street, Cluj-Napoca
e-mail: anton.muresan@econ.ubbcluj.ro

Journal of Applied Mathematics

volume IV (2011), number I

SOME PROPERTIES OF THE SOLUTIONS OF A FUNCTIONAL-DIFFERENTIAL EQUATION

MUREŞAN Viorica, (RO)

Abstract. In this paper we establish existence, uniqueness, data dependence and comparison results for the solutions of a functional-differential equation.

Key words and phrases. functional-differential equations, fixed points, Picard operators.

Mathematics Subject Classification. 34K05, 34K15, 47H10

1 Introduction

Many problems from physics, chemistry, astronomy, biology, engineering, economics, social sciences lead to mathematical models described by functional-differential equations ([4], [12],[13],[17], [18]).

The theory of functional-integral equations has developed very much in the last fifty years. Many monographs appeared: Bellman and Cooke [2] (1963), Elsgoltz and Norkin [5] (1971), Bernfeld and Lakshmikantham [3] (1974), Hale [8] (1977), Azbelev, Maksimov and Rahmatulina [1] (1991), Corduneanu [4] (1991), Gopalsamy [6] (1992), Hale and Verduyn Lunel [9] (1993), Guo and Lakshmikantham [7] (1996), such as a large number of papers.

The differential equations with linear modification of the argument are a special class of functional-differential equations. The pantograph equation

$$y'(x) = a\ y(\lambda x) + y(x),\ x > 0,\ \lambda > 0,\ \lambda \neq 1$$

and its generalizations have been studied in many papers ([11], [12], [14]).

The aim of this paper is to study the following n-th order nonlinear differential equation with linear modification of the argument:

$$y^{(n)} = f(x, y(x), y(\lambda x)),\quad x \in [0,T],\ 0 < \lambda < 1,$$

where $f \in C([0,T] \times X^2, X)$, and $(X, ||.||)$ is a Banach space..

We apply Picard and weakly Picard operators'technique due to I.A. Rus (see [20] and [21]).

We obtain existence and uniqueness results for the solution of a Cauchy problem with respect to this equation. We consider the solution set of this equation and we give some data dependence and comparison results.

2 Basic results from Picard and weakly Picard operators' theory

Let (X,d) be a metric space and $A : X \longrightarrow X$ an operator. We shall use the following notations:

$P(X) := \{Y \subseteq X | Y \neq \emptyset\}$;

$F_A := \{x \in X | A(x) = x\}$ - the fixed point set of A;

$I(A) := \{Y \in P(X) | A(Y) \subseteq Y\}$;

$A^0 := 1_X, A^1 := A, \ldots, A^{n+1} := A \circ A^n, \ldots, n \in \mathbb{N}$.

Following I.A. Rus [21], we present the following notions and results:

Definition 2.1 *A is a Picard operator if there exists $x^* \in X$ such that:*

1) $F_A = \{x^*\}$;

2) the successive approximation sequence $(A^n(x_0))_{n\in\mathbb{N}}$ converges to x^*, for all $x_0 \in X$.

Definition 2.2 *A is a weakly Picard operator if the sequence $(A^n(x_0))_{n\in\mathbb{N}}$ converges for all $x_0 \in X$ and its limit (which may depend on x_0) is a fixed point of A.*

If A is a weakly Picard operator then we consider the operator

$$A^\infty : X \longrightarrow X, \text{ defined by } A^\infty(x) := \lim_{n\to\infty} A^n(x).$$

We remark that $A^\infty(X) = F_A$.

Definition 2.3 *Let A be a weakly Picard operator and $c > 0$. A is a c-weakly Picard operator if*

$$d(x, A^\infty(x)) \leq c\, d(x, A(x)), \quad \text{for all} \quad x \in X.$$

We have

Theorem 2.1 *(data dependence theorem) Let (X,d) be a metric space and $A : X \longrightarrow X$, $B : X \to X$ two operators. We suppose that:*

(i) A is $\alpha-$contraction and $F_A = \{x_A^*\}$;

(ii) $F_B \neq \Phi$;

(iii) there exists $\delta > 0$ such that

$$d(A(x), B(x)) \leq \delta, \text{ for all } x \in X.$$

Then

$$d(x_A^*, x_B^*) \leq \frac{\delta}{1-\alpha}, \text{ for all } x_B^* \in F_B.$$

Theorem 2.2 *(characterization theorem) Let (X, d) be a metric space and $A : X \longrightarrow X$ an operator. A is a weakly Picard operator (a c-weakly Picard operator) if and only if there exists a partition of X, $X = \bigcup\limits_{\mu\in\Lambda} X_\mu$, such that:*

(a) $X_\mu \in I(A)$, for all $\mu \in \Lambda$;

(b) $A|_{X_\mu} : X_\mu \longrightarrow X_\mu$ is a Picard (a c-Picard) operator, for all $\mu \in \Lambda$.

Theorem 2.3 *Let (X, d) be a metric space and $A_i : X \longrightarrow X$, $i = 1, 2$. We suppose that*

(i) the operator A_i is c_i-weakly Picard operator, $i \in \{1, 2\}$;

(ii) there exists $\eta > 0$ such that

$$d(A_1(x), A_2(x)) \leq \eta, \quad \text{for all} \quad x \in X.$$

Then

$$H(F_{A_1}, F_{A_2}) \leq \eta \max\{c_1, c_2\},$$

where H stands for Pompeiu-Hansdorff functional.

Theorem 2.4 *Let $(X, d, \leq)$ be an ordered metric space and $A : X \longrightarrow X$, such that*

(i) A is increasing;

(ii) A is a weakly Picard operator

Then the operator A^∞ is increasing.

Theorem 2.5 *(abstract Gronwall lemma) Let $(X, d, \leq)$ be an ordered metric space and $A : X \longrightarrow X$ an operator such that:*
(i) A is a Picard operator;
(ii) A is increasing.
Then:
(a) $x \leq A(x)$ implies $x \leq x^*_A$;
(b) $x \geq A(x)$ implies $x \geq x^*_A$,
where $F_A = \{x^*_A\}$.

Theorem 2.6 *Let $(X, d, \leq)$ be an ordered metric space and $A, B, C : X \longrightarrow X$ be such that:*

(i) $A \leq B \leq C$;

(ii) the operators A, B, C are weakly Picard operators;

(iii) the operator B is increasing.

Then $x \leq y \leq z$ implies $A^\infty(x) \leq B^\infty(y) \leq C^\infty(z)$.

3 A Cauchy problem

Let $(X, ||.||)$ be a Banach space. On $C([0,T], X)$ we consider the following Bielecki norm:

$$\|y\|_B := \max_{x\in[0,T]}(\|y(x)\|e^{-\tau x}), \quad \tau > 0.$$

Let the following Cauchy problem be:

$$y^{(n)}(x) = f(x, y(x), y(\lambda x)),\ x \in [0,T],\ 0 < \lambda < 1, \tag{3.1}$$

$$y^{(k)}(0) = c_k, \quad k = \overline{0, n-1}, \tag{3.2}$$

where $f \in C([0,T] \times X^2, X)$ and $c_k \in X$, $k = \overline{0, n-1}$.

The above problem is equivalent with the following functional-integral equation:

$$y(x) = \frac{1}{(n-1)!}\int_0^x (x-s)^{n-1} f(s, y(s), y(\lambda s))\ ds + \frac{x^{n-1}}{(n-1)!}c_{n-1} + ...+$$

$$+\frac{x}{1!}c_1 + c_0,\ x \in [0,T]. \tag{3.3}$$

We have

Theorem 3.1 *We suppose that:*
(i) $f \in C([0,T] \times X^2, X)$;
(ii) there exists $L_f > 0$ such that

$$\|f(s, u_1, v_1) - f(s, u_2, v_2)\| \leq L_f(\|u_1 - u_2\| + \|v_1 - v_2\|),$$

for all $s \in [0,T]$ and $u_1, v_1, u_2, v_2 \in X$.
Then the Cauchy problem (3.1)+(3.2) has in $C([0,b], X)$ a unique solution .

Proof. Consider the operator $A_f : C([0,T], X) \to C([0,T], X)$, defined by

$$A_f(y)(x) := \frac{1}{(n-1)!}\int_0^x (x-s)^{n-1} f(s, y(s), y(\lambda s))\ ds + \frac{x^{n-1}}{(n-1)!}c_{n-1}+$$

$$+... + \frac{x}{1!}c_1 + c_0.$$

So, we obtain the fixed point problem $y = A(y)$. We have

$$\|A_f(y)(x) - A_f(z)(x)\| \leq$$

$$\leq \frac{1}{(n-1)!}L_f T^{n-1}\int_0^x (\|y(s) - z(s)\|e^{-\tau s}e^{\tau s}+$$

$$+\|y(\lambda s) - z(\lambda s)\|e^{-\tau\lambda s}e^{\tau\lambda s})ds \leq$$

$$\leq \frac{1}{(n-1)!}\frac{L_f T^{n-1}}{\tau}\left(1 + \frac{1}{\lambda}\right)\|y - z\|_B e^{\tau x},\ \text{for all } x \in [0,T].$$

So,

$$\|A_f(y)(x) - A_f(z)(x)\| e^{-\tau x} \leq$$

$$\leq \frac{1}{(n-1)!} \frac{L_f T^{n-1}}{\tau} \left(1 + \frac{1}{\lambda}\right) \|y - z\|_B,$$

for all $x \in [0, T]$.

It follows that

$$\|A_f(y) - A_f(z)\|_B \leq$$

$$\leq \frac{1}{(n-1)!} \frac{L_f T^{n-1}}{\tau} \left(1 + \frac{1}{\lambda}\right) \|y - z\|_B. \text{ for all } y, z \in C([0, T], X).$$

By choosing τ large enough, we have that A_f is a contraction. So, from Contraction principle it follows that A_f is a Picard operator and $F_{A_f} = \{y^*_{A_f}\}$.□

Remark 3.1 $A_f(C^n([0, T], X)) \subseteq C^n([0, T], X)$ and $A_f|_{C^n([0,T],X)}$ is a Picard operator. Obviously, $y^*_{A_f} \in C^n([0, T], X)$.

Together with the problem (3.1)+(3.2) we consider the problem

$$y^{(n)}(x) = g(x, y(x), y(\lambda x)),\ x \in [0, T],\ 0 < \lambda < 1, \quad (3.4)$$

$$y^{(k)}(0) = c_k, \quad k = \overline{0, n-1}, \qquad (3.5)$$

where $g \in C([0, T] \times X^2, X)$ and λ, $c_k \in X$, $k = \overline{0, n-1}$.

We have

Theorem 3.2 *We suppose that:*

(i) f, λ, c_k, $k = \overline{0, n-1}$ are as in Theorem 3.1;

(ii) $g \in C([0, T] \times X^2, X)$ is such that the problem (3.4)+(3.5) has at least a solution $y^*_{A_g}$;

(iii) there exists $\eta > 0$, such that

$$\|f(s, u, v) - g(s, u, v)\| \leq \eta, \text{ for all } s \in [0, T] \text{ and } u, v \in X.$$

Then

$$\begin{aligned} d_{\|.\|}(y^*_{A_f}, y^*_{A_g}) &\leq \frac{1}{(n-1)!} \frac{\eta T^n}{1 - \frac{1}{(n-1)!} \frac{L_f T^{n-1}}{\tau} \left(1 + \frac{1}{\lambda}\right)}, \\ \text{for all } y^*_{A_g} &\in F_{A_g}. \end{aligned}$$

Proof. We have

$$||A_f(y) - A_g(y)|| \leq \frac{1}{(n-1)!} \eta T^n, \text{ for all } y \in C([0, T], X).$$

We apply data dependence theorem (Theorem 2.1).□

4 The solution set of the equation (3.1)

Consider the equation (3.1). By successive integrations and by using the mathematical induction method, we get the following functional −integral equation:

$$y(x) = \frac{1}{(n-1)!}\int_0^x (x-s)^{n-1} f(s, y(s), y(\lambda s))\, ds + \frac{x^{n-1}}{(n-1)!} y^{(n-1)}(0) +$$

$$+ ... + \frac{x}{1!} y'(0) + y(0),\ x \in [0, T]. \qquad (4.1)$$

This equation can be written as $y = B_f(y)$, where

$$B_f : C^n([0, T], X) \to C^n([0, T], X),$$

$$B_f(y)(x) := \frac{1}{(n-1)!}\int_0^x (x-s)^{n-1} f(s, y(s), y(\lambda s))\, ds +$$

$$+ \frac{x^{n-1}}{(n-1)!} y^{(n-1)}(0) + ... + \frac{x}{1!} y'(0) + y(0),\ x \in [0, T].$$

It is obvious that the solution set of the equation (4.1) is the fixed point set F_{B_f} of the operator B_f. Let $a = (a_0, a_1, ..., a_{n-1}) \in X \times X \times ... \times X$ be, and

$$X_a := \{y \in C^n([0, T], X) / y(0) = a_0, y'(0) = a_1, ..., y^{(n-1)}(0) = a_{n-1}\}.$$

We remark that

$$C^n([0, T], X) = \cup_{a \in X \times X \times ... \times X} X_a \text{ is a partition of } C^n([0, T], X),$$

and $B_f|_{X_a}$ is a Picard operator. From the characterization theorem of the weakly Picard operators (Theorem 2.2), we obtain that B_f is a weakly Picard operator.

Now, we consider the following integral equations:

$$y(x) = \frac{1}{(n-1)!}\int_0^x (x-s)^{n-1} f_i(s, y(s), y(\lambda s))\, ds + \frac{x^{n-1}}{(n-1)!} y^{(n-1)}(0) + ... +$$

$$+ \frac{x}{1!} y'(0) + y(0),\ x \in [0, T],\ i \in \{1, 2\}, .$$

and B_{f_i}, $i \in \{1, 2\}$, the corresponding operators to these equations.
We have

Theorem 4.1 *We suppose that:*
(i) $f_i \in C([0, T] \times X^2, X)$, $i \in \{1, 2\}$;
(ii) there exists $L_{f_i} > 0$ *such that*

$$\|f_i(s, u_1, v_1) - f_i(s, u_2, v_2)\| \le L_{f_i}(\|u_1 - u_2\| + \|v_1 - v_2\|),$$

for all $s \in [0, T]$ *and* $u_1, v_1, u_2, v_2 \in X$, $i \in \{1, 2\}$;

(iii) there exists $\eta > 0$*, such that*

$$\|f_1(s,u,v) - f_2(s,u,v)\| \le \eta, \text{ for all } s \in [0,T] \text{ and all } u, v \in X.$$

Then

$$H_{\|.\|_B}(F_{B_{f_1}}, F_{B_{f_2}}) \le \frac{\eta T^n}{(n-1)!} \max\{c_{f_1}, c_{f_2}\},$$

where $c_{f_i} = \left(1 - \frac{1}{(n-1)!}\frac{L_{f_i}T^{n-1}}{\tau}\left(1 + \frac{1}{\lambda}\right)\right)^{-1}$, $i \in \{1,2\}$, *for suitable* $\tau > 0$.

Proof We have

$$\|B_{f_1}(y)(x) - B_{f_2}(y)(x)\| \le \frac{\eta T^n}{(n-1)!}, \text{ for all } x \in [0,T] \text{ and all } y \in X.$$

It follows that

$$\|B_{f_1}(y) - B_{f_2}(y)\|_B \le \frac{\eta T^n}{(n-1)!}, \text{ for all } y \in X.$$

So, we apply Theorem 2.3. □

5 Comparison results

In what follows we consider that $(X, ||.||, \le)$ is an ordered Banach space.

Consider the equation (4.1) .,

We have:

Theorem 5.1 *We suppose that:*

(i) $f \in C([0,T] \times X^2, X)$*;*

(ii) there exists $L_f > 0$ *such that*

$$\|f(s,u_1,v_1) - f(s,u_2,v_2)\| \le L_f(\|u_1 - u_2\| + \|v_1 - v_2\|),$$

for all $s \in [0,T]$ *and* $u_1, v_1, u_2, v_2 \in X$*;*

(iii) $f(s,.,.)$ *is increasing, for all* $s \in [0,T]$.

Let y *and* z *be two solutions of (4.1). Then*

$$y(0) \le z(0), y'(0) \le z'(0), ..., y^{(n-1)}(0) \le z^{(n-1)}(0) \text{ imply}$$

$$y(x) \le z(x), \text{ for all } x \in [0,T].$$

Proof The operator B_f corresponding to the above equation is a weakly Picard operator. For $a_0 \in X$, we define $\widetilde{a_0} : [0,T] \to X$, $\widetilde{a_0}\ (x) := a_0$, for all $x \in [0,T]$. We have $y = B_f^\infty(\widetilde{y(0)})$ and $z = B_f^\infty(\widetilde{z(0)})$. From the monotonicity of the operator B_f^∞ (Theorem 2.4), we have that $y(0) \le z(0), y'(0) \le z'(0), ..., y^{(n-1)}(0) \le z^{(n-1)}(0)$ imply $y(x) \le z(x)$, for all $x \in [0,T]$. □

Theorem 5.2 *We suppose that all the conditions in Theorem 3.1 are satisfied. Let y be a subsolution of the equation (3.3), and $y^*_{A_f}$ the unique solution of the problem (3.1)+(3.2). Then $y \le y^*_{A_f}$.*

Proof We have $y \le A_f(y)$.So, the proof follows from Theorem 2.5. $\square$

Now, we consider the following equations:

$$y(x) = \frac{1}{(n-1)!}\int_0^x (x-s)^{n-1} f_i(s, y(s), y(\lambda s))\, ds + \frac{x^{n-1}}{(n-1)!} y^{(n-1)}(0) + \ldots +$$

$$+\frac{x}{1!}y'(0) + y(0),\ x \in [0,T],\ i \in \{1,2,3\},$$

and B_{f_i}, $i \in \{1,2,3\}$ the corresponding operators to these equations.
We have

Theorem 5.3 *We suppose that:*
(i) $f_i \in C([0,T] \times X^2, X)$, $i \in \{1,2,3\}$;
(ii) there exists $L_{f_i} > 0$ such that

$$\|f_i(s, u_1, v_1) - f_i(s, u_2, v_2)\| \le L_{f_i}(\|u_1 - u_2\| + \|v_1 - v_2\|),$$

for all $s \in [0,T]$ and all $u_1, v_1, u_2, v_2 \in X$, $i \in \{1,2,3\}$.
(iii) $f_2(s, ., .)$ is increasing, for all $s \in [0,T]$;
(iv) $f_1 \le f_2 \le f_3$.
If y_i , $i \in \{1,2,3\}$.are corresponding solutions of the above equations, then

$$y_1(0) \le y_2(0) \le y_3(0),\ y_1'(0) \le y_2'(0) \le y_3'(0), \ldots,$$

$$, \ldots, y_1^{(n-1)}(0) \le y_2^{(n-1)}(0) \le y_3^{(n-1)}(0) \text{ imply } y_1 \le y_2 \le y_3.$$

Proof Let B_{f_i}, $i \in \{1,2,3\}$ the corresponding operators to these equations. We remark that $y_i = B^\infty_{f_i}(\widetilde{y_i(0)})$, $i \in \{1,2,3\}$. So, the proof follows from Theorem 2.6.$\square$

6 Numerical example

On $\mathbb{R}$, we consider the following Cauchy problem:

$$y'''(x) = x + \sin y(x) + \cos y(\frac{x}{3}),\quad x \in [0,6]$$

$$y(0) = 4,\quad y'(0) = 5,\quad y''(0) = 2,$$

that is equivalent with the following integral equation:

$$y(x) = \frac{1}{2!}\int_0^x (x-s)^2 (s + \frac{1}{4}\sin y(s) + \frac{1}{8}\cos y(\frac{s}{3}))ds + x^2 + 5x + 4.$$

Here $L_f = \frac{1}{4}$. For $\tau > 18$ the operator A_f,corresponding to this problem, is a contraction. So, we have that the above problem has in $C^3[0,6]$ a unique solution and this solution can be obtained by the successive approximation method, starting from any element of $C[0,6]$.

References

[1] N.V. AZBELEV, V.P. MAKSIMOV, L.F. RAHMATULINA, *Introduction to functional differential equations theory*, MIR, Moscow, 1991 (in Russian)

[2] R.E. BELLMAN, K.L. COOKE, *Differential difference equations*, Acad. Press, New York, 1963

[3] S.R. BERNFELD, V. LAKSHMIKANTHAM, *An introduction to nonlinear boundary value problems*, Acad.Press, New York, 1974

[4] C. CORDUNEANU, *Integral equations and applications,* Cambridge University Press, 1991

[5] L.F. ELSGOLTZ, S.B. NORKIN, *Introduction to the theory of differential equations with deviating arguments,* MIR, Moscow, 1971 (in Russian)

[6] K. GOPALSAMY, *Stability and oscillations in delay differential equations of population dynamics*, Kluwer Academic Publishers, 1992

[7] D. GUO, V. LAKSHMIKANTHAM, X. LIU, *Nonlinear integral equations in abstracts spaces,* Kluwer, Dordrecht, 1996

[8] J.K. HALE, *Theory of functional differential equations,* Springer Verlag, 1977

[9] J.K. HALE, S.M. SJOERD, VERDUYN LUNEL, *Introduction to functional differential equations,* Springer Verlag, New York, 1993

[10] M.W. HIRSCH, C.C. PUGH, *Stable manifolds and hyperbolic sets*, Proc. Symp. in Pure Math., 14(1970), 133-163

[11] A. ISERLES, *On the generalized pantograph functional-differential equatio*n, European J. Appl. Math. (1992), 1-38

[12] T. KATO, J.B. McLEOD, *The functional-differential equation* $y'(x) = ay(\lambda x) + by(x)$, Bull. Amer. Math. Soc., 77, 6(1971), 891-937

[13] V. LAKSHMIKANTHAM (ed), *Trends in the theory and practice of nonlinear differential equations*, Marcel Dekker, New York, 1984

[14] V. MUREŞAN, *Differential equations with affine modification of the argument*, Transilvania Press, Cluj-Napoca, 1997 (in Romanian)

[15] V. MUREŞAN, *Functional-integral equations*, Ed. Mediamira, Cluj-Napoca, 2003

[16] V. MUREŞAN, *A functional-integral equation with linear modification of the argument, via weakly Picard operators*, International Conference on Nonlinear Operators, Differential Equations and Applications (ICNODEA), July 4-8, 2007, Cluj-Napoca, Romania, Fixed Point Theory, 9(2008), No.1, 189-197

[17] J.R. OCKENDON, *Differential equations and industry*, The Math. Scientist. 5(1980), No.1, 1-12

[18] W. POGORZELSKI, *Integral equations and their applications*, Pergamon Press, 1966

[19] R. PRECUP, *Integral and nonlinear equations*, Babes-Bolyai University, Cluj-Napoca, 1993 (in Romanian)

[20] I. A. RUS, *Picard operators and applications*, Scientia Math. Japonicae*,* 58(2003), No.1, 191-219

[21] I. A. RUS, *Some nonlinear functional differential and integral equations, via weakly Picard operator theory: a survey*, Carpathian J. Math., 26(2010), No.2, 230-258

Current address

Viorica MUREŞAN
Department of Mathematics, Faculty of Automation and Computer Sciences
Technical University of Cluj-Napoca
28 Memorandumului Street, 400114 Cluj-Napoca
e-mail: vmuresan@math.utcluj.ro

Journal of Applied Mathematics

volume IV (2011), number I

BISTABLE EQUATION
A NUMERICAL MODEL OF LEVEL-SET SEGMENTATION IN IMAGE PROCESSING

OTTA Josef, (CZ)

Abstract. In this work we aim to the quasilinear bistable equation

$$\begin{cases} u_t(x,t) = \varepsilon^p \Delta_p u(x,t) - W'(u(x,t)), \quad x \in \Omega, \quad t > 0, \\ u(x,t) = 0, \quad \text{for } x \in \partial\Omega, \ t > 0; \quad u(x,0) = u_0(x). \end{cases}$$

where Δ_p stands for the p-Laplacian, $1 < p < \infty$ and W is a double-well potential (a typical choice is $W(s) = (1-s^2)^2$). We study the right choice of equation parameters to get a suitable tool for level-set image segmentation. The overview of used numerical schemes is presented with a discussion of the related choice of discretization-step size. Some numerical examples illustrating the behavior this model are provided.

Key words and phrases. Quasilinear differential equations, p-Laplacian, bi-stable equation, initial-boundary value problem, finite diference method.

Mathematics Subject Classification. Primary 35J20, 35K92 Secondary 65M06

1 Introduction

In this work, we study the bistable equation as a tool in the image processing. The localization of objects in the image is an important task of nowadays. Let us remark for example the localization of tumors in CT or MRI scans, car tracking by traffic cameras or even fingerprint recognition. There exist various approaches to to treat this problem. Let us mention the active-contour methods, discontinuity based methods [4] and global PDE methods [5] [2], [3] which will be our point of interest.

A grayscale image can be interpreted as a real-valued function $u_0 : \Omega \subset \mathbb{R}^2 \to [-1, 1]$ where the interval $[-1, 1]$ represents normalized grayscale shades such that the value -1 corresponds to black and 1 to white color. We usually consider $\Omega = [0, a] \times [0, b]$, $a, b > 0$.

The level set segmentation is based on the idea that the pixels with higher values than given threshold belong to one region. If the values are smaller, they corresponds to the second one. The aim of the segmentation is to process the initial image u_0 in a way that the final image, represented by the function u, has partially or fully segmented regions, i.e. the level sets are clustered and eventual noise inside of these regions is removed. For simplicity, we put the threshold value $\equiv 0$. Then we can process initial image u_0 as a numerical approximation of the solution u at some time T, $T > 0$ of the quasilinear problem

$$\begin{cases} u_t(x,t) = \varepsilon^p \Delta_p u(x,t) - W'(u(x,t)), \quad x \in \Omega,\ t > 0, \\ \dfrac{\partial u(x,t)}{\partial \nu} = 0 \quad \text{for } x \text{ on } \partial\Omega,\ t > 0;\ u(x,0) = u_0(x) \end{cases} \tag{1}$$

where Δ_p stands for p-Laplacian operator with $1 < p < \infty$, $\Delta_p u = \operatorname{div}(|\nabla u|^{p-2}\nabla u)$; $\nabla u = (\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_n})$; $W(u) = |1 - u^2|^\alpha$ with $\alpha > 1$ is a symmetric double-well potential; and $\varepsilon > 0$; ν denotes the outer normal at the boundary $\partial\Omega$.

Note that the equation (1) is a dynamical system driven by the reduction of the total energy

$$\mathcal{J}(u(x,t)) := \int_\Omega \frac{\varepsilon^p}{p} |\nabla u(x,t)|^p + W(u(x,t))\, \mathrm{d}x. \tag{2}$$

Indeed, the energy is a decreasing function in the time variable for given u. Using the Green identity, boundary conditions and the fact that u solves (1), we obtain that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{J}(u(x,t)) &= \int_\Omega \varepsilon^p |\nabla u(x,t)|^{p-1} \operatorname{sign}(u(x,t)) \nabla u_t(x,t) + W'(u(x,t)) u_t(x,t)\, \mathrm{d}x \\ &= \int_\Omega \big(-\varepsilon^p \Delta_p u(x,t) + W'(u(x,t))\big) u_t(x,t)\, \mathrm{d}x = -\int_\Omega u_t^2(x,t)\, \mathrm{d}x \le 0. \end{aligned}$$

The right-hand side in the equation (1) consists of two terms which counteract. The diffusive p-Laplacian Δ_p smoothens the initial image u_0 whereas the nonlinearity W' forces the image to the values ± 1. Moreover, the p-Laplacian with $p < 2$ enhances diffusion speed for small gradients (corresponding for example to small, noisy perturbations in the image) and preserves the edges. The p-Laplacian with $p > 2$ has opposite behavior. It models fast diffusion which vanishes high gradients quickly and causes fast smoothening of the edges.

In this paper, we utilize the results from the work of Drábek *et al* [6] who studied the stationary case of (1) in the form

$$\begin{cases} (\varepsilon^p |u'(x)|^{p-2} u'(x))' - W'(u(x)) = 0, \quad x \in (0,1), \\ u'(0) = u'(1) = 0. \end{cases} \tag{3}$$

They proved that for $p > \alpha$ there exist dead-core solutions of (3) connecting values ± 1. This fact will help us to study the influence of parameters p, α and ε on the solution and propose their suitable choice which provides reasonable model for the image segmentation. Our investigation is mainly based on the observation of the one-dimensional problem (3). The choice of parameters is also based on presented various numerical tests and experiments.

2 Difference Schemes

In this section, we introduce numerical schemes for the approximation of (1) which has been used in this work. For simplicity, let us restrict our attention to $\Omega = (0,1)$ to simplify our notation. The scheme for $\Omega \subset \mathbb{R}^2$ will be introduced later.

Let N be a positive integer and $\{x_i\}_{i=0}^N \subset \Omega$ be a uniform mesh such that $x_i = ih$, $h = \frac{1}{N}$. Symbol $\delta T = \frac{T}{M}$ denotes the time step size, T is the final time and M is the number of time steps. Then u_i^k stands for the approximation of $u(x_i, k\delta T)$ and $u^k = \{u_i^k\}_{i=0}^N$.

We introduce three types of difference schemes – explicit (Ex), implicit (Im) and semi-implicit (sI) time discretization with the combination of the approximation of spatial derivatives by central differences:

$$\frac{u_i^{k+1} - u_i^k}{\delta T} = \varepsilon^p (\tilde{\Delta}_p u^k)_i - W'(u_i^k), \qquad i = 1, \dots, N-1;\ k = 0, \dots, M-1 \tag{Ex}$$

$$\frac{u_i^{k+1} - u_i^k}{\delta T} = \varepsilon^p (\tilde{\Delta}_p u^{k+1})_i - W'(u_i^{k+1}), \qquad i = 1, \dots, N-1;\ k = 0, \dots, M-1 \tag{Im}$$

$$\frac{u_i^{k+1} - u_i^k}{\delta T} = \varepsilon^p (\tilde{\Delta}_p u^{k+1})_i - W'(u_i^k), \qquad i = 1, \dots, N-1;\ k = 0, \dots, M-1 \tag{sI}$$

The Neumann boundary conditions are treated by $u_0^k = u_1^k$ and $u_{N-1}^k = u_N^k$. The operator $\tilde{\Delta}_p$ stands for the approximation of the p-Laplacian Δ_p. For $\Omega = (0,1)$, we obtain $\Delta_p(u(x)) = (p-1)|u'(x)|^{p-2}u''(x))$ (here, $(\cdot)'$ denotes the spatial derivative). Then $\tilde{\Delta}_p$ can be written in the following form

$$\tilde{\Delta}_p(u^k) = \operatorname{diag}(|Du^k|^{p-2}) \cdot Lu^k$$

where D denotes the differentiation matrix approximating the first derivative by central differences and the matrix L stands for the difference approximation of the linear Laplace operator computed on the 3-point stencil. The explicit method (Ex) gives a straightforward recurrence scheme for the approximation of the solution to (1). On the other hand, every step in the implicit scheme (Im) corresponds to the solution of the nonlinear equation for u^{k+1} with given u^k. Since one can suppose that choosing δT sufficiently small provides small change from u^k to u^{k+1}, we treat this problem by Polak-Ribiére nonlinear conjugate gradient method. The semi-implicit scheme (sI) does not have almost any advantage in comparison with (Im) for general $p \neq 2$. However, the scheme (sI) for $p = 2$ reduces to

$$\frac{u^{k+1} - u^k}{\delta T} = \varepsilon^p L u^{k+1} - W'(u^k), \quad K = 0, \dots, M-1$$

which can be formally rewritten as

$$u^{k+1} = (I - \delta T \varepsilon^2 L)^{-1} u^k - W'(u^k)\delta T. \tag{4}$$

The solution of (4) can be computed by LU-factorization of $I - \delta T \varepsilon^2 L$ because this matrix is strictly diagonally dominant. Scheme then forms recurrently defined approximation of the solution to (1) which can be solved directly. The time demand of every iteration step is $\frac{8}{3}$ times higher than the explicit scheme (Ex), [4]. However, this method is unconditionally stable (this fact has been proven in [5]) and it does not need any restriction of the time-step size δT for fine spatial meshes as (Ex).

Time demands and the stability of schemes will be discussed in Section 4.

3 Transition Length

Let us remark that for the problem (1) on $\Omega = (0,1)$ with $p > \alpha$ and ε sufficiently small, there exist dead-core solutions of (3) which connect the values $u = \pm 1$. In the applications it could be useful to know *the transition length* – the length of the smallest interval on which the solution connects the values $u = -1$ and $u = +1$. With this knowledge, one can choose the discretization step h to obtain suitable approximation of these transitions with preserving reasonable number of mesh points.

Let us remark that the equation in (3) has *the first integral*. We rewrite the equation to the system of equations of the first order in the form

$$\begin{cases} u'(x) = \varphi_{p'}(v(x)), \\ v'(x) = \varepsilon^{-p} W'(u(x)) \end{cases} \tag{5}$$

where $\varphi_p(s) := |s|^{p-2}s$ for $s \neq 0$ and $\varphi_p(0) := 0$; p and p' are conjugate exponents, i.e. $\frac{1}{p}+\frac{1}{p'} = 1$. If we multiply the first equation in (5) by v' and the second term by u', sum up these equations and integrate them from x_0 to x, we get the first integral in the form:

$$\frac{\varepsilon^p}{p'}|v(x)|^{p'} - W(u(x)) = \frac{\varepsilon^p}{p'}|v(x_0)|^{p'} - W(u(x_0)) = C \quad \forall x \in \mathbb{R}. \tag{6}$$

Suppose that (u, v) is a solution of (5) with initial conditions $u(x_0) = 0,\ v(x_0) = v_c$. We want to find such v_c that related solution u touches the value $u = 1$ with the derivative equal to zero at some point x_c, i.e. $(u(x_c), v(x_c)) = (1, 0)$. Due to the symmetry of the equation, $(u(-x_c), v(-x_c)) = (-1, 0)$ holds and $l_T := 2x_c$ is *the transition length*.

Let us investigate the dependence of the transition length on the equation parameters. If we substitute $(u(x_c), v(x_c)) = (1, 0)$ to (6), we get

$$\frac{\varepsilon^p}{p'}|v(x)|^{p'} - W(u(x)) = \frac{\varepsilon^p}{p'}|v_c|^{p'} - W(0) = \frac{\varepsilon^p}{p'}|v(x_c)|^{p'} - W(u(x_c)) = 0 \tag{7}$$

due to $W(1) = 0$.

Moreover, using $|v(x)|^{p'} = |\varphi_p(u'(x))|^{p'} = |u'(x)|^p$, one can rewrite (7) to the form

$$\varepsilon|u'(x)| = \sqrt[p]{p'W(u(x))}.$$

If we consider $x \in [0, x_c]$ where $u'(x)$ is nonnegative, we can separate variables and integrate arisen identity from 0 to x_c to get

$$x_c = \frac{\varepsilon}{p'^{1/p}} \int_0^1 \frac{\mathrm{d}s}{\sqrt[p]{W(s)}} = \frac{\varepsilon}{p'^{1/p}} \int_0^1 \frac{\mathrm{d}s}{(1-s^2)^{\frac{\alpha}{p}}} = \frac{\varepsilon}{2}\sqrt[p]{1-\frac{1}{p}}B\left(\frac{1}{2}, 1-\frac{\alpha}{p}\right)$$

where B is the *Beta function*. The transition length l_T is then equal to

$$l_T = \varepsilon\sqrt[p]{1-\frac{1}{p}}B\left(\frac{1}{2}, 1-\frac{\alpha}{p}\right). \tag{8}$$

Remark 3.1 *Note that l_T is finite only for $p > \alpha$. The property of connecting values ± 1 on finite interval and the existence of dead cores promise a reasonable extension (improvement) of classical models ($p = \alpha = 2$) mentioned in [1].*

We can employ (8) for $p > \alpha$ to determine the number of equidistributed mesh points $\{x_i\}_{i=1}^N$ which guarantees that the transition length l_T is discretized sufficiently. Suppose that the transition is approximated by n_T points. Then the total number of mesh points is determined as the least integer N such that

$$N \geq \frac{n_T}{l_T} \tag{9}$$

is satisfied.

One can take advantage of the knowledge of the transition length in the image segmentation. Consider normalized grayscale shades to the interval $[-1, 1]$ with values $+1$ and -1 corresponding to the white and black color. Let us note that grayscale pictures are usually represented by 256 different shades of gray (8-bit grayscale level). But this means that every value in the interval $(\frac{127}{128}, 1]$ or in $[-1, -\frac{127}{128})$ is represented by one intensity of gray. So let us modify the transition length used in the processing of the images, analogously to (8), as the length of the smallest interval l_I such that the solution connects values $-1 + \delta$ and $1 - \delta$ with some $\delta > 0$:

$$l_I = \frac{2\varepsilon}{p'^{1/p}} \int_0^{1-\delta} \frac{\mathrm{d}s}{\sqrt[p]{W(s)}} = \frac{2\varepsilon}{p'^{1/p}} \int_0^{1-\delta} \frac{\mathrm{d}s}{(1-s^2)^{\frac{\alpha}{p}}} = \varepsilon \sqrt[p]{1 - \frac{1}{p}} B_{(1-\delta)^2}\left(\frac{1}{2}, 1 - \frac{\alpha}{p}\right) \tag{10}$$

where B_s with $|s| \leq 1$ is the *Incomplete Beta function*. Note that if $s = 1$ then B_1 reduces to standard Beta function.

If we want to apply the model (1) in the image segmentation, we are interested in the convenient choice of l_I which allows to approximate the transition from black to white colors by n_T points (pixels). For given spatial discretization by N nodes and given n_T, we can determine the equation parameters p, α and ε analogously to (9) as

$$N \geq \frac{n_T}{l_I} \tag{11}$$

where l_I is defined by (10) and we usually take $\delta = 1/128$.

4 Numerical Tests

In this section, we present numerical tests which reveals some differences between numerical schemes (Ex), (Im) and (sI).

The main advantage of the explicit scheme is in its computational simplicity which allows to compute single iteration quickly. But the main problem of this method is in the restriction of the time-step size δT for given spatial discretization. The explicit scheme is conditionally stable even for the linear diffusion equation

$$u_t(x,t) = \varepsilon^2 u_{xx}(x,t)$$

where the stability condition is provided by $\delta T \leq \frac{h^2}{2\varepsilon^2}$, cf. [4]. Unfortunately, it would be a tedious work to show the analogy of stability condition for (Ex) with $p \neq 2$. However, we performed numerical tests which revealed some analogy of this condition for $p \in (2, 10)$ and $\alpha \in (1.25, 10)$. If the time step δT satisfies

$$\delta T \leq \frac{1}{20\varepsilon^p N^p} \tag{12}$$

then the scheme (Ex) is not hampered by the instability. But let us point out that this condition is quite restrictive for higher values of ε and finer meshes in the spatial variable. Consider the discretization of $\Omega = (0, 1)$ by 800 meshpoints and $\varepsilon = 0.05$. Then i.e. for $p = 2$ we get $\delta T = 3.13 \cdot 10^{-5}$ and for $p = 5$ we obtain $\delta T = 4.88 \cdot 10^{-10}$ which is very restrictive for the computation on longer time intervals. The analogy of this condition for $p < 2$ has been tested for $p = 1.5$ and $\varepsilon = 0.05$ in the form

$$\delta T \leq 10^{-\frac{N}{200}-2} \tag{13}$$

which is even more restrictive than (12).

Let us point out that the implicit scheme (Im) does not require any refinement of the time step δT if the number of spatial nodes is increased. On the other hand, the computation of every time layer u^k is time consuming in comparison with the scheme (Ex). Therefore we also introduced semi-implicit scheme which can be employed effectively for $p = 2$, cf. (4). This method is unconditionally stable as (Im) and the computation of u^k is not time consuming.

The illustration of time demands and the time-step size in dependence on the number of mesh points is illustrated in Figure 1. The value of δT for the explicit scheme has been chosen as $\delta T = \min(\delta_m, \delta T_1)$ where δT_1 satisfies (12) or (13); the choice of the value δ_m follows Section 3 ($\delta_m = 5 \cdot 10^{-3}$ or 10^{-3}). The value of δT for (Im) and (sI) has been chosen as $\delta T = 10^{-2}$.

5 Application

In this section, we aim to the application of (1) in the image processing. Let us consider a grayscale picture with $N_x \times N_y$ pixels. Then we represent it as a function u_0 defined on the rectangle $\Omega = (0, 1) \times (0, N_y/N_x)$ with values rescaled into the interval $[-1, 1]$. Then we define $\tilde{\Omega} = \{x_{ij} : x_{i,j} = (\frac{i}{N_x-1}, \frac{j}{N_x-1}), i = 0, \ldots, N_x - 1, j = 0, \ldots, N_y - 1\}$ to be the discretization of Ω. Before applying the model (1), we need the numerical approximation of the operator p-Laplacian on $\tilde{\Omega}$. Since Δ_p can be rewritten as

$$\Delta_p(u) := \begin{cases} \frac{\partial}{\partial x}\left[(u_x^2 + u_y^2)^{\frac{p-2}{2}} u_x\right] + \frac{\partial}{\partial y}\left[(u_x^2 + u_y^2)^{\frac{p-2}{2}} u_y\right] & \text{for } |\nabla u| \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

we can differentiate the terms inside square brackets and approximate arisen formula as:

$$\tilde{\Delta}_p(u) := \begin{cases} (D_x \cdot G(u)) \cdot D_x u + (D_y \cdot G(u)) \cdot D_y u + G(u) \cdot Lu & \text{for } |\nabla u| > \texttt{tol}, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

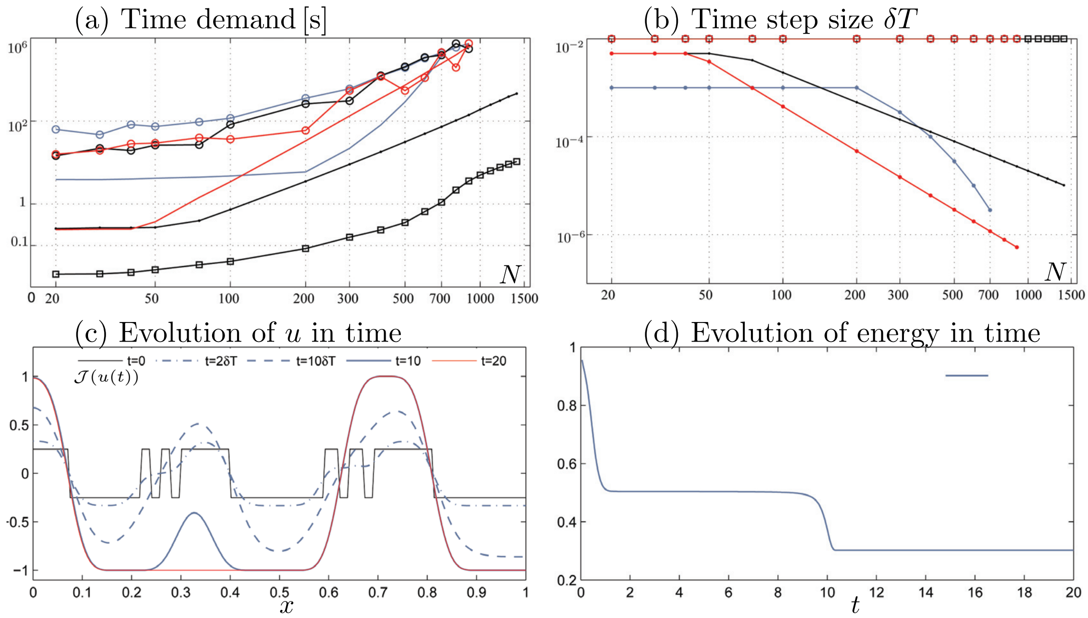


Figure 1: The time demand (panel (a)) and the time-step size δT (panel (b)) in dependence on the number of mesh points N for the approximation of (1) from $t = 0$ to T=20. The explicit method (Ex) is represented by solid line, the implicit method (Im) is displayed by lines with circles and the semi-implicit method (for $p = 2$ only) by line with squares. Graphs for the solution of (1) with $\alpha = 1.5$, $\varepsilon = 0.05$ and $p = 1.5$ are represented by blue lines; $p = 2$ by black lines and $p = 3$ by red lines. Panel (c) illustrates the evolution of the solution u for $p = 2$, $\alpha = 1.5$, $\varepsilon = 0.05$ with $N = 200$ at various values of t. Corresponding energy decrease is presented in panel (d).

where D_x (D_y) denotes the differentiation matrix approximating the partial derivative with respect to the $x(y)$-variable on $\tilde{\Omega}$ by central differences and the matrix L stands for the difference approximation of the linear Laplace operator on $\tilde{\Omega}$ computed by the standard 5-point stencil, cf. [4]. Diagonal matrix $G(u)$ is defined as $G(u) = \mathrm{diag}\big([(D_x u)^2 + (D_y u)^2]^{\frac{p-2}{2}}\big)$ and the tolerance `tol` is usually taken as 10^{-4}. The homogenous Neumann boundary conditions are treated by reflection method, cf. [4].

Let us remark that we have determined the dependence of the transition length on the values of equation parameters by (11) for one dimensional case. We will use analogous identity for Ω in the form

$$\sqrt{N_x^2 + N_y^2} \geq \frac{n_T}{l_I}$$

where n_T corresponds to the desired number of pixels approximating transition length l_I which is defined by the same formula as in one-dimensional case, cf. (10).

Now, we are ready to test the model (1) on grayscale images. At the beginning, we compare the results with various values of equation parameters. This observation gives us suitable choice of p, α and ε which guarantee reasonable image segmentation. In Figure 2, we present grayscale picture (100×100 pixels) with added white noise. This noised image has been used as the initial condition u_0 in (1). We fixed the value $\alpha = 2$ and vary diffusion parameter p and ε to satisfy (10) with $n_T = 1,\ 3,\ 5$.

One can observe that solutions with smaller value of p preserve the edges of original image without noise. On the other hand, the images with greater value of p have edges smoothened. If the number n_T is increased, the transition length is prolonged which allows to remove unwanted noise sufficiently. However larger values of n_T also cause blurring of the edges, so one has to be careful in choosing suitable equation parameters.

Moreover, in Figure 3, we present some examples of the image segmentation of grayscale pictures with added white noise (values in the interval $[-\frac{3}{4}, \frac{3}{4}]$). The images has been processed by (1) with $p = 1.1$ $\alpha = 2$ and $n_T = 5$. One can observe that the edges in the axis directions are completely preserved. On the other hand, the diagonal edges are not smooth as one could

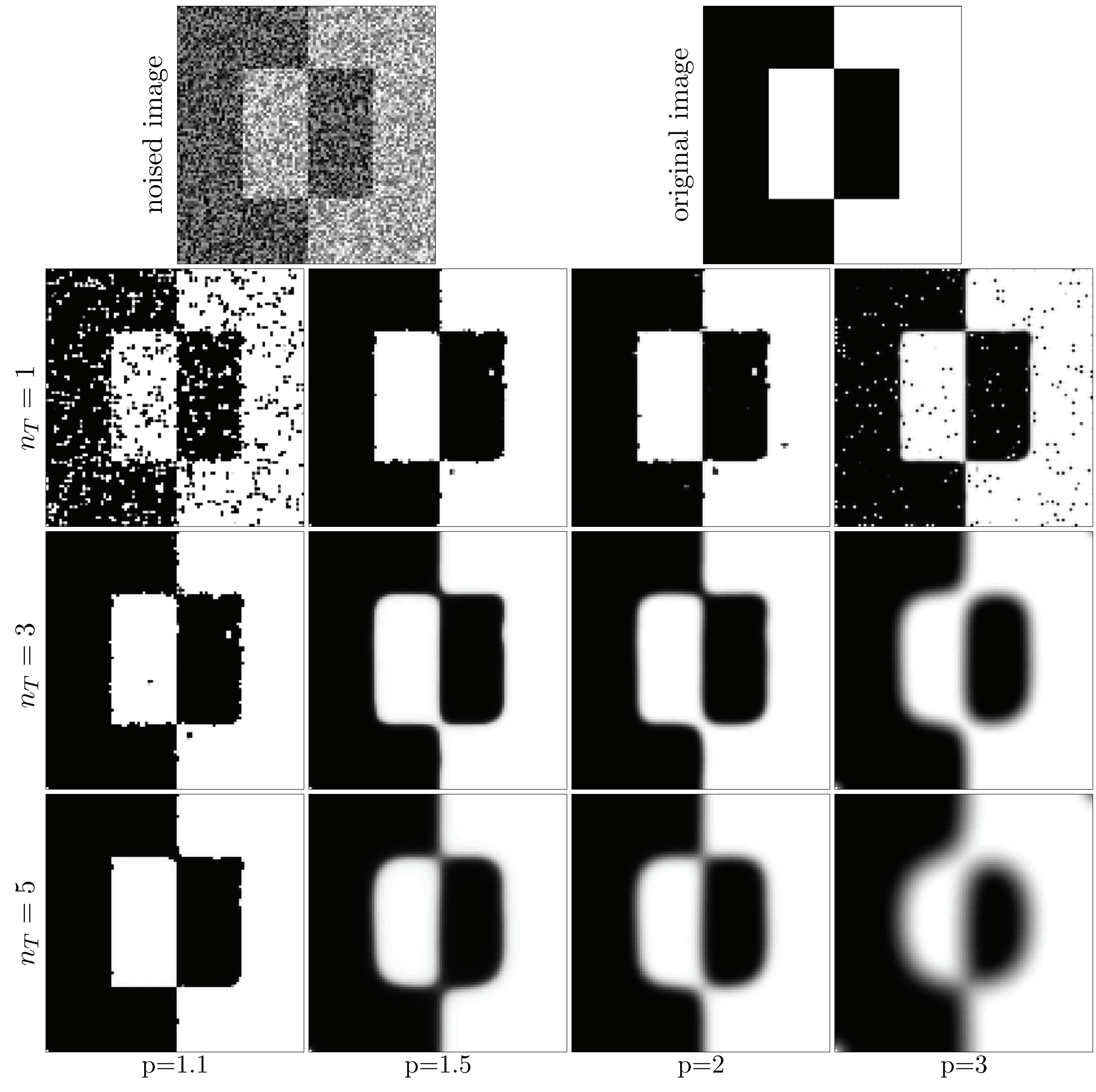


Figure 2: Segmentation of noised pictures (100×100 pixel) with added white noise by (1) with $\alpha = 2$; ε chosen accordingly to (11) with $n_T = 1, 3, 5$ and $p = 1.1,\ 1.5,\ p = 2,\ 3$. The figure presents solutions at the time $T = 2$.

original image
noised image
segmented image

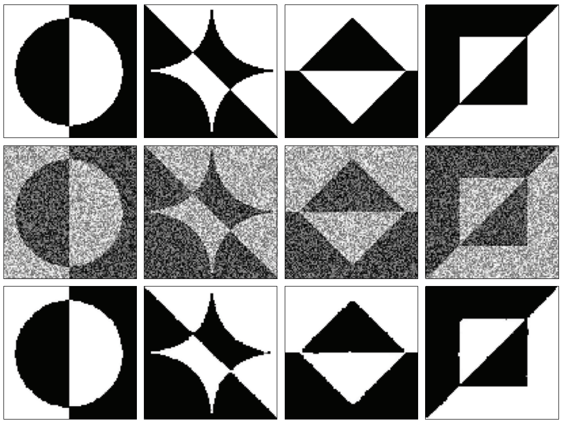
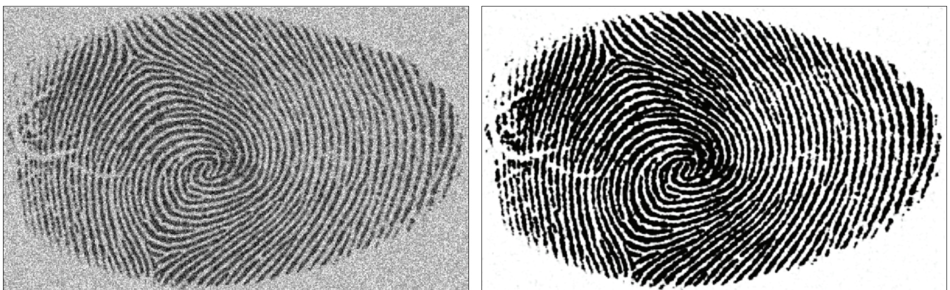


Figure 3: Segmentation of noised pictures (100 × 100 pixel) by (1) with $\alpha = 2$, $p = 1.1$ and $\varepsilon = 2.695 \cdot 10^{-3}$ (corresponds to $n_T = 5$). Segmented pictures are represented by solutions at the time $T = 2$.

expect. This behavior is caused by the choice of the discretiazation of the p-Laplacian which uses standard 5-point stencil. One can obtain better results using the approximation of partial derivatives and the Laplacian by more sophisticated schemes, for example the difference schemes using the 9-point stencil, cf. [4].

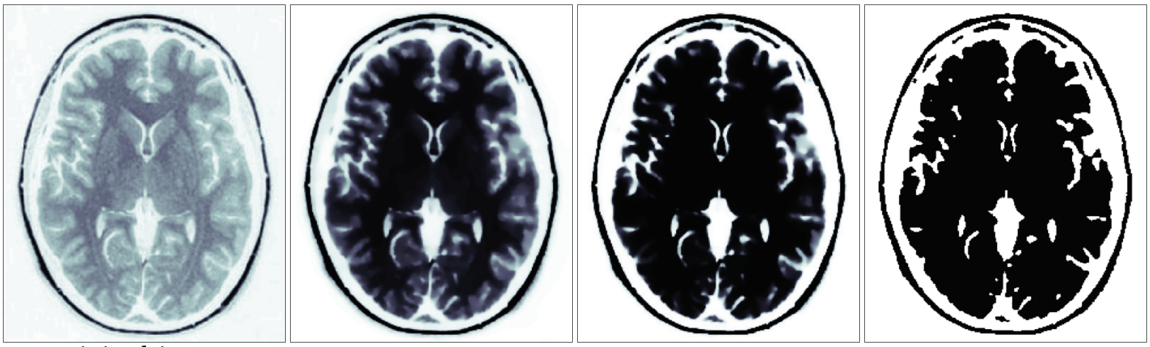
Figure 5 illustrates the usage of the equation to the preprocessing of the fingerprint scans. One can observe that the noise is removed whereas the epidermal ridges are preserved. The results has been attained using $p = 2$ and $\alpha = 2$ together with $n_T = 2$. Note that the modified image has some extra connections of ridges. The solution has been computed by the semi-implicit method (sI) with $\delta T = 0.001$.



noised image segmented image

Figure 4: Segmentation of noised fingerprint image (365 × 600 pixels) by (1) with $\alpha = 2$, $p = 2$ and $\varepsilon = 2.3756 \cdot 10^{-3}$ corresponding to $n_T = 2$. Segmented picture is represented by solution at the time $T = 0.9$.

As another application, we present an example of processing of the MRI scan of the human brain. The most used method for the segmentation of the MRI scans is the active contour method, cf. [4]. However our model gives also promising results. One can observe these results in Figure 5 where we used the model (1) with $p = 1.1$, $\alpha = 2$ and $n_T = 2$. The processing has been performed by (Ex) with $\delta T = 0.001$.



original image $T = 0.5$ $T = 0.8$ $T = 2$

Figure 5: Processing of brain MRI scan (256×315 pixels) by (1) with $\alpha = 2$, $p = 1.1$ and $\varepsilon = 3.4224 \cdot 10^{-4}$ corresponding to $n_T = 2$. Segmented image is presented at times $T = 0.5$, 0.8 and $T = 2$.

6 Summary

Quasilinear bistable equation can be used as a tool for the level-set-segmentation problems. The main advantage of this method is in its possibility to control the interaction of the smoothening diffusion term Δ_p and the double-well potential which forces the solution to the values ± 1 easily by varying the parameter ε. The suitable choice of this equation parameters is presented in this paper. We also propose an approach how to numerically treat the problem (1). But there exist a lot of open problems which could be solved further.

For example, using adaptive-mesh technique, similar to [2], one can get faster segmentation using coarsening-mesh strategy. The idea based on the coarsening of mesh points where the solution is flat could give promising speed up of the computation. Also the investigation of the equation with multiple-well potential could be used in more complex segmentation problems as for example complete segmentation of CT scans where usually 6 shades of gray are used [**?**].

Acknowledgement

The author gratefully acknowledges the support by the Ministry of Education, Youth and Sports of the Czech Republic, grant KONTAKT no. ME10093.

References

[1] ALLEN, S. M., CAHN, J. W., *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening.* In Acta Metallurgica, Vol 27, Issue 6, June 1979, p. 1085-1095.

[2] BÄNSCH, E., MIKULA, K., *A coarsening finite element strategy in image selective smoothing.* In Computing and Visualization in Science, 1(1):5361, 1997.

[3] BENEŠ, M., CHALUPECKÝ, V., MIKULA, K., *Geometrical image segmentation by the AllenCahn equation.* In Applied Numerical Mathematics, Vol 51, Issues 2-3, November 2004, p. 187-205.

[4] AUBERT, G., KORNPROBST, P.*Mathematical problems in image processing: partial differ- ential equations and the calculus of variations*. Springer-Verlag, 2006.

[5] CHOI, J. W., LEE, H. G., JEONG, D., KIM, J., *An unconditionally gradient stable numerical method for solving the AllenCahn equation*. In Physica A: Statistical Mechanics and its Applications, Vol 388, Issue 9, May 2009, p. 1791-1803.

[6] DRÁBEK, P., MANÁSEVICH, R. F., TAKÁČ, P., *Stationary Solutions for a Quasilinear Model for Phase Transitions in One Space Dimension*. To Appear

[7] WHEELER, A. A., BOETTINGER, W. J., McFADDEN, G. B., *Phase-field model for isothermal phase transitions in binary alloys*. In Phys. Rev. A, Vol. 45, Issue, 10, 1992, p. 7424-7439.

Current address

Josef Otta, Ing.
Department of Mathematics
University of West Bohemia in Pilsen
300 00 Plzeň
Czech Republic
e-mail: jotta@kma.zcu.cz

Journal of Applied Mathematics

volume IV (2011), number I

ASYMPTOTIC PROPERTIES OF A SYSTEM OF LINEAR DIFFERENTIAL EQUATIONS OF SECOND ORDER WITH CONSTANT COEFFICIENTS AND CONSTANT DELAY

SVOBODA Zdeněk, (CZ)

Abstract. Investigation of structure of the linear system of differential equations of the second order with constant delay and constant matrix is based on the concepts of the delayed matrix cosine and delayed matrix sine. These concepts are analog with delayed exponential of matrix and the linear system of differential equations. The purpose of this contribution is to describe asymptotic properties of the delayed matrix cosine and delayed matrix sine by asymptotic properties of the delayed exponential of matrix. This problem is solved in a special case.

Key words and phrases. delayed equation, exponential matrix,

Mathematics Subject Classification. Primary 34K06; Secondary 34K25.

1 Introduction

We will discuss the asymptotic properties of the solutions of the system of linear differential equations of second order with constant coefficients and constant delay in the form:

$$\ddot{x}(t) + \Omega^2 x(t-\tau) = 0, \tag{1}$$

where $t \geq 0$, $\tau > 0$ is a constant, $x \in \mathbb{R}^n$, $x(t)$ is twice continuously differentiable vector function and Ω is a nonsingular constant matrix.

For the investigation of the structure of solution of these systems is important the concept of delayed matrix cosine and delayed matrix sine which is defined in [9], [8]. Here are derived the theorems which specify the solution of the equation (1) which satisfies the initial conditions $x(t) = \varphi(t)$, $\dot{x}(t) = \dot{\varphi}(t)$. These results are obtained by step by step method and due to it the definition of delayed matrix cosine resp. sine is given with respect to intervals. Analogous results are derived for systems of linear differential equations with constant delay and a constant matrix and delayed exponential of matrix, for more details see [10], [2] and for difference systems [7], [5], [6] too. The purpose of this contribution is to describe asymptotic properties of delayed matrix cosine resp. sine by asymptotic properties of delayed exponential matrix which are studied in [12]. In a special case the matrix C is found such that the exponential of this matrix e^{Ct} has the similar asymptotic properties as delayed exponential of given matrix e_τ^{At} i.e., $\lim\limits_{n\to\infty}\left(e_\tau^{An\tau} - e^{Cn\tau}\right) = 0$ holds. The exponential matrix e^{Ct} is a matrix solution of the system

$$\dot{y}(t) = Ay(t-\tau) \tag{2}$$

and the matrix C is a solution of the equation $C = Ae^{-C\tau}$. The eigenvalues of the matrix C satisfying this equation depend on the values of LambertW well—known function of eigenvalues of the matrix A.

2 Step by step method for the linear system of differential equations of the first and second order.

The concept of delayed exponential of matrix is based on application of the step by step method for the linear system of differential equations of the first order with constant matrix and constant delay (2). The delayed exponential of a matrix is defined as follows:

$$e_\tau^{At} = \begin{cases} \Theta, & -\infty < t < -\tau; \\ I, & -\tau \le t < 0; \\ \vdots & \\ I + A\frac{t}{1!} + A^2\frac{(t-\tau)^2}{2!} \cdots + A^k\frac{(t-(k-1)\tau)^k}{k!}, & (k-1)\tau \le t < k\tau. \\ \vdots & \end{cases} \tag{3}$$

The matrix is together with an initial condition $e_r^{At} \equiv I$ for $-r \le t \le 0$ the solution of the equation (2). Moreover, for permutable matrices A, B, i.e. $AB = BA$, it is possible to describe the solution of initial value problem for equation $\dot{y}(t) = Ay(t-\tau) + By(t)$ $x(t) = \varphi(t)$ for $-r \le t \le 0$ as follows:

$$y(t) = e^{B(t-\tau)}e_\tau^{A_1(t-\tau)}\varphi(-\tau) + \int\limits_{-\tau}^{0} e^{A(t-\tau-s)}e_\tau^{A_1(t-\tau-s)}e^{A\tau}[\dot{\varphi}(s) - B\varphi(s)]\mathrm{d}s, \text{ where } A_1 = e^{-B\tau}A.$$

The structure of solutions is studied in [10] also for the equation

$$\dot{x}(t) = Ax(t) + Bx(t-r) + f(t),$$

and can be expressed by notion of delayed exponential of a matrix.

The same procedure yields for the system of second order (1) the definition of the notion of the delayed matrix sine and cosine:

$$\operatorname{Cos}_\tau \Omega t := \begin{cases} \Theta, & -\infty < t < -\tau; \\ I, & -\tau \le t < 0; \\ I - \frac{\Omega^2}{2}, & 0 \le t < \tau; \\ \vdots & \vdots \\ I - \Omega^2 \frac{t^2}{2!} + \dots (-1)^k \Omega^{2k} \frac{[t-(k-1)\tau]^{2k}}{(2k)!}, & (k-1)\tau \le t < k\tau. \\ \vdots & \vdots \end{cases}$$

$$\operatorname{Sin}_\tau \Omega t := \begin{cases} \Theta, & -\infty < t < -\tau; \\ \Omega(t+\tau), & -\tau \le t < 0; \\ \Omega(t+\tau) - \frac{\Omega^3}{3!} t, & 0 \le t < \tau; \\ \vdots & \vdots \\ \Omega(t+\tau) - \frac{\Omega^3}{3!} t + \dots (-1)^k \Omega^{2k+1} \frac{[t-(k-1)\tau]^{2k+1}}{(2k+1)!}, & (k-1)\tau \le t < k\tau. \\ \vdots & \vdots \end{cases}$$

These notions enable us to specify the solution of (1) satisfying the initial conditions $x(t) = \varphi(t)$, $\dot{x}(t) = \dot{\varphi}(t)$ for $-\tau \le t < 0$ in the form

$$x(t) = (\operatorname{Cos}_\tau \Omega t)\varphi(-\tau) + \Omega^{-1} \left[(\operatorname{Sin}_\tau \Omega t)\dot{\varphi}(-\tau) + \int_{-\tau}^{0} \operatorname{Sin}_\tau \Omega(t - \tau - \xi)\ddot{\varphi}(\xi) d\xi \right].$$

3 Asymptotic properties of delayed exponential of a matrix

The values of delayed exponential of matrix $e_\tau^{An\tau} = \sum_{k=0}^{n} \frac{(n+1-k)^k}{k!} A^k \tau^k = P_n(A\tau)$ are polynomial with respect to matrix A and delay τ. For constant matrix A such that at first its Jordan canonical form is a diagonal matrix,(i.e. there is an invertible matrix P such that $A = P^{-1} \operatorname{diag}(\dots \lambda_j \dots) P$) and at second for which modules of its eigenvalues of A satisfies the inequality $e|\lambda_j|\tau < 1$ there is the constant matrix C describing the limit

$$e^{-C\tau} = \lim_{n\to\infty} \frac{e_\tau^{A\tau n}}{e_\tau^{A\tau(n+1)}} = \lim_{n\to\infty} P^{-1} \operatorname{diag}\left(\dots \frac{P_n(\lambda_j \tau)}{P_{n+1}(\lambda_j \tau)} \dots \right) P =$$
$$\lim_{n\to\infty} P^{-1} \operatorname{diag}\left(\dots \sum_{j=1}^{n+2} \frac{(-j)^{j-1}}{j!} (\lambda_j \tau)^{j-1} + O((\lambda_j\tau)^{n+3}) \dots \right) P. \quad (4)$$

If λ_j satisfies the above inequality then the limit of eigenvalues of the sequence of matrices $e_\tau^{A\tau n} e_\tau^{-A\tau(n+1)}$ has the form of series $\sum_{j=1}^{\infty} \frac{(-j)^{j-1}}{j!} (\lambda_j \tau)^{j-1}$ and these series coincide with the fractions $\frac{W_0(\lambda_j \tau)}{\lambda_j \tau}$, where $W_0(x)$ is the principal branch of LambertW function.

Remark 3.1 *Lambert function, named after Johann Heinrich Lambert, see [11], is the inverse function of* $f(w) = we^w$*. The function satisfying* $z = W(z)e^{W(z)}$ *is multivalued (except at 0). For real arguments* x $(x > -1/e)$ *and real* w $(w > -1)$ *this equation defines a single-valued function* $W_0(x)$*. The Taylor series of* W_0 *around* 0 *is given by*

$$W_0(x) = \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} x^n = x - x^2 + \frac{3}{2}x^3 - \frac{8}{3}x^4 + \frac{125}{24}x^5 - \dots, \tag{5}$$

which has the radius convergence $1/e$*. The Lambert* W *function cannot be expressed in terms of elementary functions. For more details see [4].*

This fact is analogous with the routine for solving of system of linear differential equations for which it is possible to derived the characteristic equation in the form e^{Ct} of using by substitution. In this case we get the characteristic equation of (2) in the form:

$$C = Ae^{-C\tau} \Leftrightarrow Ce^{C\tau} = A. \tag{6}$$

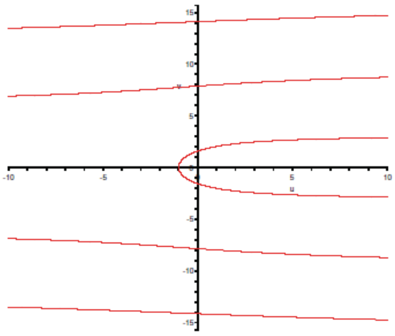
Moreover, is possible to express the matrix $C = Ae^{-C\tau} = P^{-1} \operatorname{diag}(\dots \frac{W_0(\lambda_j \tau)}{\tau} \dots)P$ and the eigenvalues of the matrix C equal to $\frac{W_0(\lambda_j \tau)}{\tau}$.

The branches of function $W(z)$ are numbered by integer k. The curve which separates the principal branch W_0 from the branches W_1 and W_{-1} is

$$\{-v \cot v + vi| - \pi < v < \pi\}$$

and the curves which separate the remaining branches are

$$\{-v \cot v + vi | 2k\pi < \pm v < (2k+1)\pi\}.$$



The fact $u + iv = W(z) \Rightarrow (u^2 + v^2)e^{2u} = |z|^2$ has the consequence that for any constant $|z|$ and any couple of values $u_k + iv_k = W_k(z)$, $u_l + iv_l = W_l(z)$ the implication $v_k^2 < v_l^2 \Rightarrow u_k > u_l$ holds and so, the inequality $\Re e W_k(z) > \Re e W_l(z)$ holds, too. Such $\Re e W_0(z)$ is the greatest real part of all values $W(z)$. Let the matrix A has the Jordan canonical form diagonal with eigenvalues λ_j satisfying the inequality $e\lambda_j \tau < 1$. Then the function $e^{\bar{C}t}$, which matrix $\bar{C}$ is defined by (4), bounds exponentials e^{Ct} of other matrices C which eigenvalues depend on values of other branches of LambertW function.

4 Main result

Let C be the constant matrix such that at first exponential e^{Ct} of this matrix is the matrix solution of the equation (1), and at second it has the similar asymptotic properties as $\operatorname{Cos}_\tau \Omega t$ resp. $\operatorname{Sin}_\tau \Omega t$, i.e.

$$\lim_{n\to\infty} \left(\operatorname{Cos}_\tau \Omega nt - e^{Cn\tau}\right) = 0 \;\text{ resp. }\; \lim_{n\to\infty} \left(\operatorname{Sin}_\tau \Omega nt - e^{Cn\tau}\right) = 0,$$

holds, then the matrix C satisfies the characteristic equation

$$C^2 + B^2 e^{-C\tau} = 0,$$

which can be obtained as the product of following couple of equivalent equations

$$C \pm iBe^{-C\frac{\tau}{2}} = 0.$$

This fact evokes the next Lemma describing the relation between delayed exponential of matrix and the delayed matrix cosine and delayed matrix sine.

Lemma 4.1 *For any square matrix Ω holds*

$$\mathrm{Cos}_\tau\, \Omega\left(t - \frac{\tau}{2}\right) = \frac{e_{\frac{\tau}{2}}^{i\Omega t} + e_{\frac{\tau}{2}}^{-i\Omega t}}{2}, \qquad \mathrm{Sin}_\tau\, \Omega(t-\tau) = \frac{e_{\frac{\tau}{2}}^{i\Omega t} - e_{\frac{\tau}{2}}^{-i\Omega t}}{2i}. \tag{7}$$

Proof: The proof immediately follows from the comparison of definition of delayed exponential of matrix with the definitions of the delayed matrix cosine and delayed matrix sine.

These relations is possible to read as a modification of well—known Euler's identity in the form:

$$e_{\frac{\tau}{2}}^{i\Omega t} = \mathrm{Cos}_\tau\, \Omega\left(t - \frac{\tau}{2}\right) + i\, \mathrm{Sin}_\tau\, \Omega(t-\tau).$$

Theorem 4.2 *Let the canonical form of a matrix Ω is a diagonal matrix, i.e. there is an invertible matrix Q such that $\Omega = Q^{-1}\,\mathrm{diag}(\ldots\omega_j\ldots)Q$ and the modules of eigenvalues of Ω satisfy the inequality $|\omega_j\tau| < \dfrac{2}{e}$. Then there exists the matrix C defined as follows*

$$C = Q^{-1}\,\mathrm{diag}\left(\ldots \frac{2W_0\left(\omega_j \frac{\tau}{2}\right)}{\tau} \ldots\right) Q$$

For delayed matrix sine and cosine the following equations hold:

$$\lim_{n\to\infty}\left(\mathrm{Sin}_\tau\, \Omega(n\tau) - \frac{e_{\frac{\tau}{2}}^{iC(n+1)\tau} - e_{\frac{\tau}{2}}^{-iC(n+1)\tau}}{2i}\right) = 0$$

and

$$\lim_{n\to\infty}\left(\mathrm{Cos}_\tau\, \Omega\,(n\tau) - \frac{e_{\frac{\tau}{2}}^{i\Omega\left(n+\frac{1}{2}\right)\tau} + e_{\frac{\tau}{2}}^{-i\Omega\left(n+\frac{1}{2}\right)\tau}}{2}\right) = 0$$

Proof: For the matrix Ω satisfying the assumptions of this Theorem there is a matric C such that $\lim\limits_{n\to\infty}\left(e_{\frac{\tau}{2}}^{iCn\tau/2} - e^{-Cn\tau/2}\right) = 0$ which is possible to defined by the matrix limit

$$e^{-C\frac{\tau}{2}} = \lim_{n\to\infty} \frac{e_\tau^{i\Omega n\frac{\tau}{2}}}{e_\tau^{i\Omega(n+1)\frac{\tau}{2}}} = \lim_{n\to\infty} Q^{-1}\,\mathrm{diag}\left(\ldots \frac{P_n(i\omega_j\frac{\tau}{2})}{P_{n+1}(i\omega_j\frac{\tau}{2})}\ldots\right) Q =$$

$$\lim_{n\to\infty} Q^{-1}\,\mathrm{diag}\left(\cdots \sum_{j=1}^{n+2} \frac{(-j)^{j-1}}{j!}(\omega_j\frac{\tau}{2})^{j-1} + O((\omega_j\frac{\tau}{2})^{n+3})\ldots\right) Q.$$

This matrix C is possible to describe as $C = Q^{-1}\,\mathrm{diag}\left(\dots \dfrac{2W_0(\omega_j\tau/2)}{\tau}\dots\right)Q$.

Moreover, $\mathrm{Sin}_\tau\,\Omega(n\tau) = \dfrac{e_{\frac{\tau}{2}}^{i\Omega(n+1)\tau} - e_{\frac{\tau}{2}}^{-i\Omega(n+1)\tau}}{2i}$ and $\mathrm{Cos}_\tau\,\Omega(n\tau) = \dfrac{e_{\frac{\tau}{2}}^{i\Omega(n+\frac{1}{2})\tau} - e_{\frac{\tau}{2}}^{-i\Omega(n+\frac{1}{2})\tau}}{2}$ hold, so the theorem is proved.

Acknowledgement

The paper was supported by grant Czech Grant Agency (Prague) no. P201-11-0768 and by the Czech Ministry of Education in the frame of projects MSM002160503

References

[1] R. BELLMAN, K. L. COOKE: Differntial-Difference equations, ACADEMIC PRES New York, 1963.

[2] A. BOICHUK, J. DIBLÍK, D. KHUSAINOV, M. RŮŽIČKOVÁ: *Fredholm's boundary-value problems for differential systems with a single delay.* Nonlinear Analysis 72, pp. 2251-2258, 2010.

[3] A. BOICHUK , J. DIBLÍK , D. KHUSAINOV , M. RŮŽIČKOVÁ: *Boundary Value Problems for Delay Differential Systems*Advances in Difference Equations,Volume 2010, Article ID 593834, 20 pages

[4] R. M. CORLESS, G. H. GONNET, D. E. G. HARE, D. E. KNUTH: *On the Lambert W Function* Advances in Computational Mathematics, Vol 5 , 329-359, 1996.

[5] J. DIBLÍK, D. KHUSAINOV: *Representation of solutions of discrete delayed system* $x(k+1) = Ax(k) + Bx(k-m) + f(k)$ *with commutative matrices,* Journal of Mathematical Analysis and Applications, **318**, No 1,(ISSN 0022-247X), pp. 63–76. , 2006

[6] J. DIBLÍK, D. KHUSAINOV: *Representation of solutions of linear discrete systems with constant coefficients and pure delay,* Advances in Difference Equations, Art. ID 80825, DOI 10.1155/ADE/2006/80825, pp. 1–13. (ISSN: 1687-1839, e-ISSN: 1687-1847, doi:10.1155/ADE), 2006

[7] J. DIBLÍK, D. YA. KHUSAINOV, M. RŮŽIČKOVÁ: *Controllability of linear discrete systems with constant coefficients and pure delay* SIAM J. CONTROL OPTIM. **Vol. 47**, No. 3, pp. 1140-1149,2008

[8] J. DIBLÍK, D. YA. KHUSAINOV, M. RŮŽIČKOVÁ, J. LUKÁČOVÁ: *Control of Oscillating Systems with a Single Delay,*Advances in Difference Equations Volume 2010, Article ID 108218, 15 pages, 2010.

[9] D. YA. KHUSAINOV, J. DIBLÍK, J. LUKÁČOVÁ M. RŮŽIČKOVÁ: *Representation of a solution of the Cauchy problem for an oscillating system with pure delay*, Nonlinear Oscillations, **Vol 11**, No 2, pp. 276-285, 2008.

[10] D. YA. KHUSAINOV, G. V. SHUKLIN: *Linear autonomous time-delay system with permutation matrices solving*, Studies of the University of Žilina, Mathematical Series **Vol 16**, 1-8, 2003.

[11] J. H. LAMBERT: *Observationes variae in mathesin puram.* Acta Helveticae physicomathematico- anatomico-botanico-medica, Band III, pp. 128-168, 1758.

[12] Z. SVOBODA: *Asymptotic properties of delayed exponential of matrix. Journal of Applied Mathematics.* Slovak University of Technology in Bratislava, p. 167 - 172, ISSN 1337- 6365, 2010.

Current address

Zdeněk Svoboda, RNDr. CSc.
UMAT FEKT VUT v Brně
Technická 8 61600 Brno Česká republika,
tel. 420541142830
svobodaz@feec.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

DECOMPOSITION METHOD FOR FRACTIONAL INTEGRO-DIFFERENTIAL EQUATIONS.

Šmarda Zdeněk, (CZ)

Abstract. This paper outlines a reliable strategy for solving fractional integro-differential equations. The modified Adomian decomposition method for the fractional calculus is applied. Numerical examples are presented to illustrate the Adomian decomposition approach.

Key words and phrases. Fractional integro-differential equations, Adomian decomposition method.

Mathematics Subject Classification. Primary 45J05; Secondary 34A08.

1 Introduction

Many problems in mathematical physics, theory of elasticity, viscodynamics of fluids, hydrology reduce to fractional differential and integro-differential equations [2].

Adomian decomposition method for solving integral has been presented by Adomian [1],[2] and then this has been extended by Wazwaz [9],[10] to higher-order integro-differential equations. In this method a solution is considered as an infinite series, rapidly converging to an accurate solution. The Adomian decomposition method provides solutions without any need for linearization or discrezations. Essentially the method provides a systematic computational procedure for equations of physical significance.

Consider the fractional integro-differential

$$\begin{aligned} D_*^{\alpha} y(t) &= a(t)y(t) + f(t) + \int_0^t K(t,s)F(y(s))ds, \qquad t \in [0,1], \\ y(0) &= t_0 \end{aligned} \tag{1}$$

where D_*^α is the Caputo derivative, $\alpha > 0$ is the order of the fractional derivative, $a(t), K(t,s), F(y(t))$ are continuous functions, t_0 is a constant.

Equation (1) of integer order was investigated by Adomian [10]. Momani [5] obtained local and global existence and uniqueness of a solution of integro-differential equation (1). Mital and Nigam [4] solved some classes of equation (1) using the Adomian decomposition method and obtained analytical solutions. Rawashdeh [7] solved the same fractional integro-differential equations as Mital and Nigam using the collocation method to approximate the solution of investigated equations.

In the paper we determine the Adomian decomposition method for solving of equation (1) and this method will be illustrated on a simple example.

At first, we recall some basic notions from fractional calculus:

The Riemann-Liouville fractional integral of any order $\alpha > 0$ for a function $\psi(t)$ for $t \in R^+$ is defined as

$$J^\alpha \psi(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \psi(s) ds, \quad t > 0, \quad \alpha > 0. \tag{2}$$

$J^0 = I$ is the identity operator. The Riemann-Liouville fractional derivative of order $\alpha > 0$ is defined as the left inverse of the corresponding Riemann-Liouville fractional integral, i.e.

$$D^\alpha J^\alpha = I.$$

For positive integer m such that $m - 1 < \alpha \le m$ we obtain

$$(D^m J^{m-\alpha}) J^\alpha = D^m (J^{m-\alpha} J^\alpha) = D^m J^m = I,$$

so that

$$D^\alpha = D^m J^{m-\alpha}.$$

From here and (2) we get

$$D^\alpha \psi(t) = \begin{cases} \frac{1}{\Gamma(m-\alpha)} \frac{d^m}{dt^m} \int_0^t \frac{\psi(s)}{(t-s)^{\alpha+1-m}} ds, & m-1 < \alpha < m \\ \frac{d^m}{dt^m} \psi(t), & \alpha = m. \end{cases}$$

Properties of the operators J^α and D^α were founded by Podlubny [6]. An alternative definition of fractional derivative was introduced by Caputo [3]. According to the Caputo definition

$$D_*^\alpha = J^{m-\alpha} D^m, \quad m-1 < \alpha < m.$$

Hence

$$D_*^\alpha \psi(t) = \begin{cases} \frac{1}{\Gamma(m-\alpha)} \int_0^t \frac{\psi(s)^{(m)}}{(t-s)^{\alpha+1-m}} ds, & m-1 < \alpha < m \\ \frac{d^m}{dt^m} \psi(t), & \alpha = m. \end{cases}$$

The basic property of the Caputo fractional derivative is

$$J^\alpha D_*^\alpha \psi(t) = \psi(t) - \sum_{k=0}^{m-1} \psi^{(k)}(0^+) \frac{t^k}{k!}.$$

2 Adomian decomposition method

Consider equation (1). Applying J^α on both sides of (1) we get

$$y(t) = \sum_{k=0}^{m-1} y^{(k)}(0^+)\frac{t^k}{k!} + J^\alpha\Big(a(t)y(t) + f(t) + \int_0^t K(t,s)F(y(s))ds\Big).$$

Adomian decomposition method defines a solution $y(t)$ by the series

$$y(t) = \sum_{n=0}^{\infty} y_n(t)$$

and the function F is decomposed as $F = \sum_{n=0}^{\infty} A_n$, where A_n are the Adomian polynomials given by

$$A_n = \frac{1}{n!}\left[\frac{d^n}{d\lambda^n} f\Big(\sum_{i=0}^{\infty} \lambda^i y_i\Big)\right]_{\lambda=0}, \quad n = 0, 1, 2, \dots .$$

The components $y_0, y_1, \dots$ are determined recursively by

$$\begin{aligned} &y_0(t) = \textstyle\sum_{k=0}^{m-1} y^{(k)}(0^+)\frac{t^k}{k!} + J^\alpha f(t),\\ &\vdots\\ &y_{k+1}(t) = J^\alpha\big(a(t)y_k\big) + J^\alpha\Big(\textstyle\int_0^t K(t,s)A_k ds\Big),\\ &\vdots \end{aligned}$$

Example. Consider the following fractional integro-differential equations

$$D_*^{1/2} y(t) = (\cos t - \sin t)y(t) + f(t) + \int_0^t t \sin s y(s) ds, \quad y(0) = 0 \tag{3}$$

where

$$f(t) = \frac{2}{\Gamma(5/2)} t^{3/2} + \frac{1}{\Gamma(3/2)} t^{1/2} + t(2 - 3\cos t + t\sin t + t^2 \cos t).$$

According to the decompsition method we get

$$\begin{aligned} &y_0(t) = y(0) + J^\alpha(f(t)) = t^2 + t + J^\alpha(2t - 3\cos t - t^2 \sin t + t^3 \cos t),\\ &\vdots\\ &y_{n+1}(t) = J^\alpha\big((\cos t - \sin t)y_n(t))\big) + J^\alpha\Big(\textstyle\int_0^t t \sin s y_n(s) ds\Big),\\ &\vdots \end{aligned}$$

We can simplify difficult fractional integrals by taking the truncated Taylor expansions for trigonometric terms but we get only approximate solution of (3). Rawashdeh [7] solved this equation by collocation method and obtained more accurate approximation of a solution of (3) than Mittal and Nigam [4].

Now we use the modified decomposition method suggested by Wazwaz [10]. Its modified algorithm for fractional integro-differential equation (3) is

$$\begin{aligned}
&y_0(t) = t^2 + t,\\
&y_1(t) = J^\alpha(2t - 3\cos t - t^2 \sin t + t^3 \cos t) + J^\alpha\big((t\cos t - \sin t)y_0(t)\big) + J^\alpha\Big(\int_0^t t \sin s y_0(s) ds\Big) = 0,\\
&y_2(t) = 0,\\
&\vdots\\
&y_{n+1}(t) = 0,\\
&\vdots
\end{aligned}$$

Hence $y(t) = t^2 + t$ is the exact solution of (3).

Acknowledgement

This research has been supported by the Czech Ministry of Education in the frames of MSM002160503 Research Intention MIKROSYN New Trends in Microelectronic Systems and Nanotechnologies and MSM0021630529 Research Intention Inteligent Systems in Automatization.

References

[1] ADOMIAN, G.: *A Review of the Decomposition Method in Applied Mathematics*, J. Math. Anal. Appl. **135**(2-3) (1988), 501-544.

[2] ADOMIAN, G.: *Solving Frontier Problems of Physics: The decomposition method,* Kluwer Academic Publishers, Boston, (1994).

[3] CAPUTO, M.: *Linear model of dissipation whose Q is almost frequency independent - II*, Geophys. J. R. Astr. Soc. **13** (1967), 529-539.

[4] MITTAL, S.M., NIGAM, R.: *Solution of fractional integro-differential equations by Adomian decomposition method,* Int. J. of Appl. Math. and Mech. **4**(2) (2008), 87-94.

[5] MOMANI, S.M.: *Local and global existence theorems of fractional integro-differential equations,* Journal of Fractional Calculus, **18** (2000), 81-86.

[6] PODLUBNY, I.: *Fractional Differential Equation,* Mathematics in Science and Engineering, **198**, Academic Press, (1999).

[7] RAWASHDEH, E.A.: *Numerical solution of fractional integro-differential equations by collocation method,* Applied Mathematics and Computation, **176**(1) (2005), 1-6.

[8] ŠMARDA, Z.: *Some properties of Fractional Integrals and Derivatives,* Proceedings of the Sixth International Conference on Soft Computing Applied in Computer and Economic Enviroments, ICSC 2008, Kunovice, 229-234, ISBN 978-80-7314-134-9 229-234.

[9] WAZWAZ, A.: *A reliable treatment for mixed Volterra-Fredholm integral equations,* Appl. Math. Comput. **127**(2-3) (2002), 405-414.

[10] WAZWAZ, A.: *A reliable algorithm for solving boundary value problems for higher-order integro-differential equations,* Appl. Math. Comput. **118**(2-3) (2001), 327-342.

Current address

Doc. RNDr. Zdeněk Šmarda, CSc.
Department of Mathematics
Faculty of Electrical Engineering and Communication
Brno University of Technology, Technická 8, 616 00 BRNO
e-mail: smarda@feec.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

SELF-ORGANIZED SYSTEMS

TREŠL Jiří, (CZ)

Abstract. Two models developed primarily in physics and leading to the origin of chaotic and self-organized behaviour are demonstrated. First, relatively simple nonlinear systems with discrete time are capable to break into chaotic state under small changes of control parameters. The important problem here is to distinguish between chaotic time behaviour and external random noise. In this connection, the use of correlation dimension and BDS test is discussed. Further, the origin of self-organized behaviour is manifested with the use of simple Ising spin model. It is shown, the resulting behaviour is critically dependent on the value of coupling strength.

Key words. deterministic chaos, correlation dimension, self-organized behaviour

Mathematics Subject Classification: Primary 62P20; Secondary 91B84.

1 Deterministic Chaos

Even some simple *non-linear deterministic systems* can under certain conditions pass to chaotic states [1], [2]. Chaotic behavior is bounded, not periodic and similar to random one. It is highly sensitive to small change of initial conditions and cannot be predicted for long time. It is called *deterministic chaos*.

In the case of autonomous systems with continuous time, chaotic behavior can appear in 3-rd order systems, in the case of non-autonomous already in 2-nd order systems. Some systems with discrete time (described by difference equations) are capable of producing of chaotic behaviour already in one-dimensional case.

As an simple example, let us consider discrete system described by *logistic difference equation*

$$x_{n+1} = Ax_n\left(1-x_n\right) = f\left(x_n\right) \qquad 0 \le A \le 4 \tag{0.1}$$

where A is control parameter. For $0 < A \le 4$, values from the interval < 0,1> will be mapped also into this interval. This system has been originally used for the modelling of the growth of a population at limited territory providing that individual generations do not overlap. State variable x_n

gives the number of objects in n-th generation. Quadratic term prevents the population from unlimited growth, parameter A describes the influence of surroundings.
The function $f(x_n)$ is so-called **iterative function**. The position of fixed points can be determined from the relation

$$\bar{x} = A\bar{x}\left(1-\bar{x}\right) \Rightarrow \bar{x}_1 = 0 \qquad \bar{x}_2 = 1-\frac{1}{A} \tag{0.2}$$

and their stability from the behavior of first derivatives

$$\frac{df}{dx}\left(x = A\right) = A \qquad \frac{df}{dx}\left(x = 1-1/A\right) = 2-A \tag{0.3}$$

For $A \le 1$, there is only one fixed point $\bar{x}_1 = 0$, which is stable (attractor). A sequence of values $x_0, x_1, x_2, \ldots$ tends to converge to zero (a population becomes extinct). In the range $1 < A < 3$, two fixed points exist. Now, the point $\bar{x}_1 = 0$ is unstable (repellor) and $\bar{x}_2 = 1-(1/A)$ is stable (attractor). Trajectories from arbitrary initial condition converge to one-point attractor (a population reach stationary state). For $A=3$, the first **bifurcation** occurs and the second fixed point will be unstable as well.
Within interval $3.57 < A \le 4$, the system behaviour becomes very complex. There is infinitely many intervals of the parameter A (periodic windows) with stable periodic trajectories. On the other hand, there are certain parameter values leading to chaotic behavior. In the limit case $A=4$, also analytical solution exists in the form

$$x_n = \sin^2\left(2^n \arcsin\sqrt{x_0}\right) \tag{0.4}$$

Clearly, from this solution, obvious extreme sensitivity to very small changes of initial value x_0 is seen. Thus, the solution of logistic difference equation leads in the case of increase of control parameter A from periodic solution with period two through bifurcation cascade of period doubling to chaotic behavior.
Now we mention briefly another way of the emergence of a chaotic state. It is the case of the class of *by part linear mapping*. For example, to this class belongs *symmetric roof mapping*

$$\begin{aligned} x_{n+1} &= 2Ax_n && \text{for} \quad 0 \le x_n \le 0.5 \\ x_{n+1} &= 2A\left(1-x_n\right) && \text{for} \quad 0.5 \le x_n \le 1 \end{aligned} \tag{0.5}$$

with control parameter $0 < A \le 1$. This function is continuous, but it has not the derivative at the point $x=0.5$. In the case $A < 0.5$, only one fixed point exists and namely $\bar{x} = 0$; this point is stable, because $2A < 1$. For $A > 0.5$, there are two fixed points

$$\bar{x}_1 = 0 \qquad \bar{x}_2 = \frac{2A}{2A+1} \tag{0.6}$$

and it holds

$$\left|\frac{df}{dx}\left(x = \overline{x_1}\right)\right| = \left|\frac{df}{dx}\left(x = \overline{x_2}\right)\right| = 2A > 1 \tag{0.7}$$

Both fixed points are unstable and trajectories are chaotic. In this case, chaotic behavior arises suddenly for $A > 0.5$ and no bifurcations occur.

2 Quantification of Chaotic Behaviour

The reasons for the construction of quantitative characteristics of chaotic behavior are the following:

- quantificators can help to distinguish deterministic chaos from „noisy" behavior, produced by the action of external random influences
- quantificators can help to determine minimum number of variables needed for the construction of a dynamical model of the system
- quantificators can help to classify systems according to universally valid regularities
- changes of quantificators may signalise changes in qualitative behavior of a system

Grassberger and Procaccia introduced the characteristic called *correlation dimension*, based on the behavior of so-called correlation sum [3]. For the computation of correlation dimension, we need data about the evolution of a trajectory (in sum n values). For each i-th point of the trajectory, we seek relative frequency $p_i(r)$ of trajectory points, lying at distance less than r from the point i (except i-th point)

$$p_i(r) = \frac{n_i}{n-1} \tag{2.1}$$

Correlation sum is then computed as average relative frequency

$$C_1(r) = \frac{1}{n}\sum_{i=1}^{n} p_i(r) \tag{2.2}$$

Obviously $C_1(r)=0$, if r is less than minimal distance among the points of a trajectory. On contrary $C_1(r)=1$ means, the distances among individual points do not exceed r . Minimal possible non-zero value $C_1(r)=2/(n(n-1))$ occurs in the case, only one distance is less than r.
Relative frequency can be formally expressed using Heaviside step function

$$\begin{aligned} H(x) &= 0 \qquad \mathit{for} \quad x < 0 \\ H(x) &= 1 \qquad \mathit{for} \quad x \geq 0 \end{aligned} \tag{2.3}$$

$$p_i(r) = \frac{1}{n-1}\sum_{j=1,\, j\neq i}^{n} H\left(r - \left|x_i - x_j\right|\right) \tag{2.4}$$

Similarly, correlation sum can be written as

$$C_1(r) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,\, j\neq i}^{n} H\left(r - \left|x_i - x_j\right|\right) \tag{2.5}$$

In limit case $n \to \infty$, correlation sum is melted into *correlation integral*. Correlation dimension is then given by formula

$$D_1 = \lim_{r\to 0} \frac{\log C_1(r)}{\log r} \tag{2.6}$$

A time series of single variable can be often sufficient for the determination of important characteristics of a multidimensional dynamical system. The groups of values

$$x_{t+1}, x_{t+2}, \ldots, x_{t+d} \qquad t = 0,1,2,\ldots,(n-d) \tag{2.7}$$

give the coordinates of a point in d-dimensional space. Then the sequence of these groups describes the time evolution of a system in d-dimensional *embedding space*. In this case, correlation sum can be written as

$$C_d(r) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} H\left(r - \left|\mathbf{x_i} - \mathbf{x_j}\right|\right) \tag{2.8}$$

because it depends on embedded dimesion d. Vector $\mathbf{x_i}$ of dimension d has components

$$\mathbf{x_i} = \left(x_i, x_{i+L}, x_{i+2L}, \ldots, x_{i+(d-1)L}\right) \tag{2.9}$$

where L is time lag between neighbouring values. The length of the difference of two vectors is mostly calculated as Euklid distance

$$\left|\mathbf{x_i} - \mathbf{x_j}\right| = \sqrt{\sum_{k=0}^{d-1} \left(x_{i+kL} - x_{j+kL}\right)^2} \tag{2.10}$$

Then it holds for correlation dimension

$$D_d = \lim_{r\to 0} \frac{\log C_d(r)}{\log r} \tag{2.11}$$

Grassberger and Procaccia have studied the behavior of these characteristic. In the case of i.i.d. (**i**ndependent **i**dentically **d**istributed) process with regular distribution is

$$D_1 = \lim_{r\to 0} \frac{\log C_1(r)}{\log r} = \lim_{r\to 0} \frac{\log 2 + \log r}{\log r} = 1 \tag{2.12}$$

$$D_2 = \lim_{r\to 0} \frac{\log C_2(r)}{\log r} = \lim_{r\to 0} \frac{\log 4 + 2\log r}{\log r} = 2 \tag{2.13}$$

and generally $D_d = d$. On the contrary, in the case of non-linear deterministic process is the behavior of correlation sum quite different. For example, in the case of roof mapping is $D_1 = D_2 = \ldots = D_d = 1$.

Brock, Dechert and Scheinkman showed, for a finite r and i.i.d. process the following relation is valid [4]

$$C_d(r) = \left[C_1(r)\right]^d \tag{2.14}$$

and suggested test statistic

$$T_d(r,n) = \frac{C_d(r,n) - \left[C_1(r,n)\right]^d}{s_d(r,n)} \qquad (2.15)$$

where $C_d(r,n), C_1(r,n)$ are sample correlation sums and $s_d(r,n)$ is the estimate of the standard deviation. This statistic has asymptotically standard normal distribution N(0,1) providing the validity of null hypothesis (i.i.d. process). It is seen from the Table 1, correlation dimensions for daily PX returns increase almost according to $D_d = d$, whereas they remains close to 1 for logistic deterministic chaos. However, for both processes, the null hypothesis (i.i.d.) is strongly rejected.

Table 1. Correlation dimensions computed for embedded dimensions 2 – 6.
Left: Daily PX returns Right: Logistic deterministic chaos
Last row: Test statistic for BDS test

	PX					Logistic				
Length	2	3	4	5	6	2	3	4	5	6
200	2.063	3.012	3.658	3.535	4.256	-	1.127	1.097	1.102	0.894
400	2.018	3.002	3.816	4.644	4.654	-	1.324	1.462	1.364	1.091
600	2.041	2.959	4.083	4.758	5.043	-	1.264	1.375	1.262	1.087
800	2.058	2.999	4.086	4.649	5.058	-	1.247	1.259	1.204	1.056
1000	2.131	2.893	3.846	4.492	4.852	-	1.172	1.234	1.146	0.967
1200	2.104	2.978	3.886	4.468	4.829	-	1.149	1.201	1.123	0.950
1400	2.125	2.986	3.878	4.496	4.890	-	1.194	1.254	1.173	0.978
1600	2.067	2.984	3.868	4.508	4.909	-	1.221	1.217	1.167	1.003
1800	2.067	3.004	3.894	4.546	4.938	-	1.241	1.232	1.178	1.012
2000	2.063	3.017	3.900	4.542	4.908	-	1.243	1.229	1.172	1.007
BDS	13.43	17.48	19.78	21.65	23.61	-11.30	-15.52	-18.29	-18.05	-17.24

3 Ising Model

Spin models belong to the best known and simplest models of interacting elements which are capable to exhibit self-organized behaviour [5]. In principle, these models express tle balance between ordering tendency following from particle interactions and disordering influence of external stochastic noise. A spin represents an entity, which can assume only finite discrete number of different states, in the simplest case two states 0 and 1.
A spin model is defined by the geometry of the lattice on which the spins are located and by the interactions between spins. We assume some network of interacting elements $i = 1, 2, \ldots, I$ and $N(i)$ denotes the set of elements directly connected to i–th element. In the simplest version called *Ising model,* we suppose only two possible states $s_i \in \{-1, +1\}$. Then the state of an element i is determined by

$$s_i = sign\left(K \sum_{j \in N(i)} s_j + u_i\right) \qquad u_i \approx N(0,1) \qquad (3.1)$$

where sign(x) function is equal to +1 (-1) for positive (negative) argument x and K is a positive parameter called coupling strength. It is important to realize, this equation describes the system state at given time moment. In the next time, new random disturbances change states of individual elements.

As an example, consider square lattice of elements

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$$

and suppose that the interactions can occur only either in horizontal or vertical direction. Thus, corner elements are influenced by two neighbours, border elements by three neighbours and inner elements by four neighbours according to following scheme (the numbers in parenthesis correspond to influencing elements)

$$\begin{bmatrix} 1(2,5) & 2(1,3,6) & 3(2,4,7) & 4(3,8) \\ 5(1,6,9) & 6(2,5,7,10) & 7(3,6,8,11) & 8(4,7,12) \\ 9(5,10,13) & 10(6,9,11,14) & 11(7,10,12,15) & 12(8,11,16) \\ 13(9,14) & 14(10,13,15) & 15(11,14,16) & 16(12,15) \end{bmatrix}$$

Taking this configuration, the time development of corresponding Ising model was computed for different values of coupling strength *K*. The results are depicted on Figures 1- 2, where the sum of states *S* over all 16 elements is charted. This variable can vary from the lowest value $S = -16$ (for all states equal to -1) to the highest value $S = +16$ (for all states equal to +1).

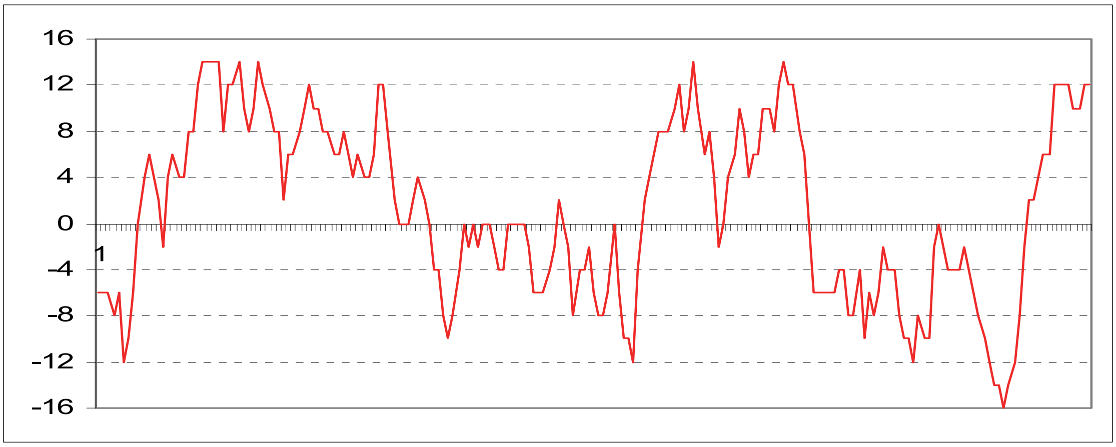


Figure 1. Time development of Ising model: K = 0.5, n = 200

Clearly, if there is no coupling strength, the sum S behaves purely randomly. The increase of coupling strength results in the onset of certain ordering. Finally, after exceeding of specific critical value ($K = 0.82$ in our case), the system starts to behave like large scale correlated one and bounded within certain limits.

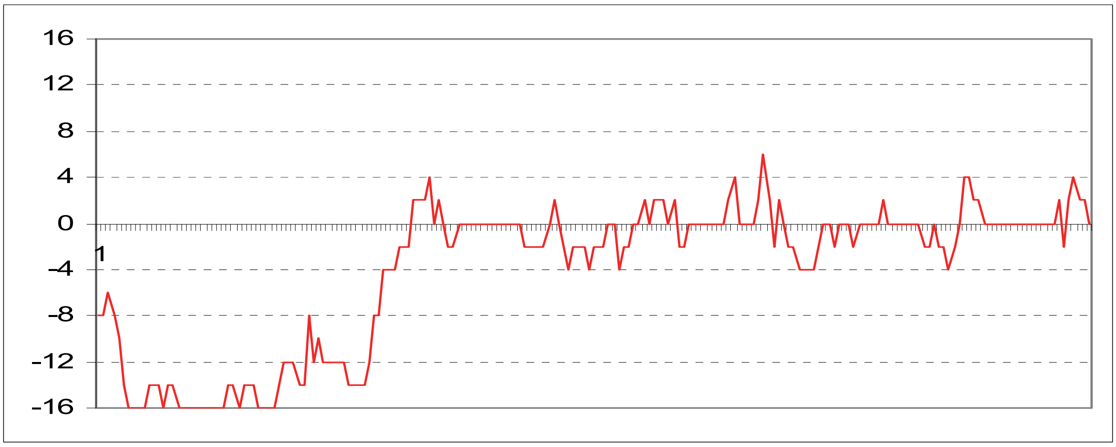


Figure 2. Time development of Ising model: K = 0.81, n = 200

Acknowledgement

Financial support of this research by grant from institutional support of longtime conceptual scientific and research development at FIS VŠE is gratefully acknowledged.

References

[1] HILBORN, R.C.: *Chaos and Nonlinear Dynamics.* Oxford University Press, Oxford, 2006.
[2] ALLIGOOD, K.T., et al.: *Chaos: an introduction to dynamical systems.* Springer Verlag, New York, 2001.
[3] GRASSBERGER, P., Procaccia, I.: *Characterization of strange attractors*. Physics Review Letters, Vol. 50 , pp.346-355, 1983.
[4] CAMPBELL, J.Y., LO, A.W., MACKINLAY, A.C.: *The econometrics of financial markets*. Princeton Univ.Press, Princeton, 1997.
[5] SORNETTE, D.: *Critical phenomena in natural sciences.* Springer Verlag, Berlin, 2000.

Current address

doc.Ing.Jiří Trešl,CSc.
University of Economics Prague,
W.Churchill Sq.4, 13067 Prague, CZ
e-mail: tresl@vse.cz

Journal of Applied Mathematics

volume IV (2011), number I

MODELLING OF THE 4 TH ORDER RESONANT FILTERS *LCLC* CONSIDERING NON-LINEARITIES

BENOVA Mariana, (SK), ABDAMULA Mahmud A.R., (LAR), DOBRUCKY Branislav, (SK)

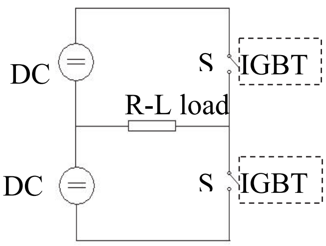
Abstract. The problem how to obtain sinusoidal voltage of load side at non-harmonic periodical supplying from the converters is very important in engineering practice. The paper shows that either LCLC resonant filter for frequency of fundamental harmonic component can be used, or LC filter tuned for switching frequency. Both filters have to remove higher harmonic components from the supplying voltage to reach the harmonic distortion roughly 5 %.
The paper deals mainly with analysis and modelling of the 4th order LCLC filter (of the first type) considering non-linearities, and with comparing to the other types of filtering. Simulation experiment results as verification confirm good quality of output quantities of the filter, voltage and current.

Keywords. State-space modelling, fictitious exiting functions, non-linear function, Fourier series, transient analysis

Mathematics Subject Classification: Primary 34A05, 34A09; Secondary 37A05

1 Basic connection of single-phase inverter with output resonant filter

The single-phase voltage inverter can be realised as full-bridge or half-bridge connection [1], Fig. 1, and completed by LCLC filter to provide sinusoidal output quantities.



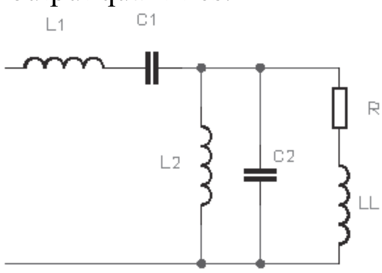


Fig. 1. Principle schematic connections of the single-phase half-bridge voltage inverter (left) and LCLC resonant filter (right)

2 Linearized mathematical model of both filters

Considering converter scheme and *LCLC* filter in Fig. 1 with series resistance r_1 and parallel resistance r_2 then the state-space equations can be

$$
\begin{aligned}
\frac{di_{L1}}{dt} &= \frac{1}{L_1}u(t) - \frac{r_{L1}}{L_1}i_{L1} - \frac{1}{L_1}u_{C1} - \frac{1}{L_1}u_{C2} \\
\frac{di_{L2}}{dt} &= \frac{1}{L_2}u_{C2} \\
\frac{du_{C1}}{dt} &= \frac{1}{C_1}i_{L1} \\
\frac{du_{C2}}{dt} &= \frac{1}{C_1}i_{L1} - \frac{1}{C_2}i_{L2} - \frac{1}{C_2 \cdot r_{C2}}u_{C2} - \frac{1}{C_2}i_L \\
\frac{di_{LL}}{dt} &= \frac{1}{L_L}u_{C2} - \frac{R}{L_L}i_L
\end{aligned}
\tag{1a,b,c,d,e}
$$

The system of equations (1) can be rearranged into matrix form

$$\frac{d}{dt}\overline{x} = \mathbf{A}\overline{x} + \mathbf{B}\overline{u} \tag{2}$$

where **A**, **B** are system matrices, and ***x, u*** are state-space and exciting vectors, respectively.

Let's use some of explicit methods for numerical integration of ordinary differential equations. Euler's method is a first-order numerical procedure for solving ordinary differential equations with a given initial value. We want to approximate the solution of the initial value problem

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0 \tag{3}$$

using the first two terms of the Taylor expansion of y, which represents the linear approximation around the point (t_0, $y(t_0)$). One step (if h is step size) of the Euler's method from t_n to $t_{\mathrm{n}+1} = t_{n+h}$ is

$$y_{n+1} = y_n + h\, f(t_n, y_n) \tag{4}$$

The direct Euler's method is explicit, i.e. the solution y_{n+1} is an function of y_i for $i \le n$ [7].

$$
\begin{aligned}
&\frac{d}{dt}\overline{x} = \frac{\overline{x}_{n+1} - \overline{x}_n}{h} \qquad \overline{x}_{n+1} - \overline{x}_n = h(\mathbf{A}\overline{x}_n + \mathbf{B}\overline{u}_n) \\
&\overline{x}_{n+1} = h(\mathbf{A}\overline{x}_n + \mathbf{B}\overline{u}_n) + \overline{x}_n = (1 + h\mathbf{A})\overline{x}_n + h\mathbf{B}\overline{u}_n
\end{aligned}
\tag{5a,b,c}
$$

The implicit Euler's method

$$y_{n+1} = y_n + h\, f(t_n, y_{n+1}) \tag{6}$$

yields

$$
\begin{aligned}
\overline{x}_{n+1} - \overline{x}_n &= h(\mathbf{A}\overline{x}_{n+1} + \mathbf{B}\overline{u}_n) \\
\overline{x}_{n+1} &= h(\mathbf{A}\overline{x}_{n+1} + \mathbf{B}\overline{u}_n) + \overline{x}_n \\
(1 - h\mathbf{A})\overline{x}_{n+1} &= \overline{x}_n + h\mathbf{B}\overline{u}_n
\end{aligned}
$$

$$\overline{x}_{n+1} = (1 - h\mathbf{A})^{-1}\overline{x}_n + (1 - h\mathbf{A})^{-1}h\mathbf{B}\overline{u}_n = (1 - h\mathbf{A})^{-1}(\overline{x}_n + h\mathbf{B}\overline{u}_n) \tag{6a,b,c,d}$$

Important is, that this method is absolutely stable, independently on integration step size.

After time discretization of system equations using implicit Euler's methods one obtains

$$\begin{bmatrix} i_{L1}(i+1) \\ i_{L2}(i+1) \\ u_{C1}(i+1) \\ u_{C2}(i+1) \\ i_{LL}(i+1) \end{bmatrix} = [\mathbf{J} - h\mathbf{A}]^{-1} \begin{bmatrix} i_{L1}(i) \\ i_{L2}(i) \\ u_{C1}(i) \\ u_{C2}(i) \\ i_{LL}(i) \end{bmatrix} + [\mathbf{J} - h\mathbf{A}]^{-1} h \begin{bmatrix} \frac{1}{L} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} u(t) \tag{7}$$

where

$$\mathbf{A} = \begin{pmatrix} \frac{-r_1}{L_1} & 0 & \frac{-1}{L_1} & \frac{-1}{L_1} & 0 \\ 0 & 0 & 0 & \frac{1}{L_2} & 0 \\ \frac{1}{C_1} & 0 & 0 & 0 & 0 \\ \frac{1}{C_1} & \frac{-1}{C_2} & 0 & \frac{-1}{C_2 . r_2} & \frac{-1}{C_2} \\ 0 & 0 & 0 & \frac{1}{L} & \frac{-R}{L} \end{pmatrix} \tag{7a}$$

i_{L1}, i_{L2} - currents through the inductors L_1 and L_2, respectively
i_L - current through the load R, L
u_{C1}, u_{C2} - voltages of the capacitors C_1 a C_2, respectively
J - unit matrix
A - system matrix
h - step size
$u(t)$ - output voltage of the inverter (filter input voltage)

2.1 Considering non-linearity owing to switching mode of source voltage

Positive half-period of source voltage (switches S+)

Considering scheme of converter with *LCLC* filter in Fig. 2 the state-space equations are the same as (1) but with filter input voltage

$$u_1 = U - 2u_T = U - 2U_{T0} - 2r_T\, i_{L1} \tag{8}$$

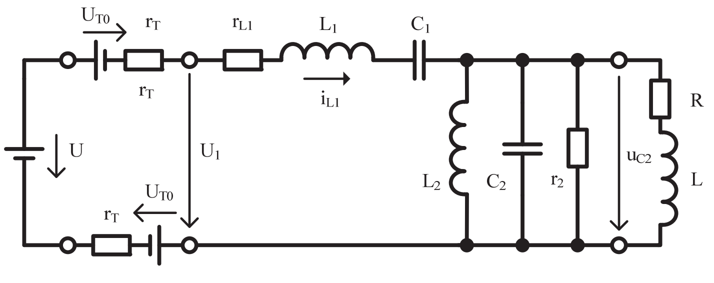


Fig. 2. Principle schematic connections of the single-phase voltage inverter and the output filter for positive half-period of source voltage

Negative half-period of source voltage (switches S-)

Considering scheme of converter with *LCLC* filter in Fig.3 the state-space equations are the same as (1) but with filter input voltage

$$u_1 = -(U - 2u_T) = -(U - 2U_{T0} - 2r_T\, i_{L1}) \qquad (9)$$

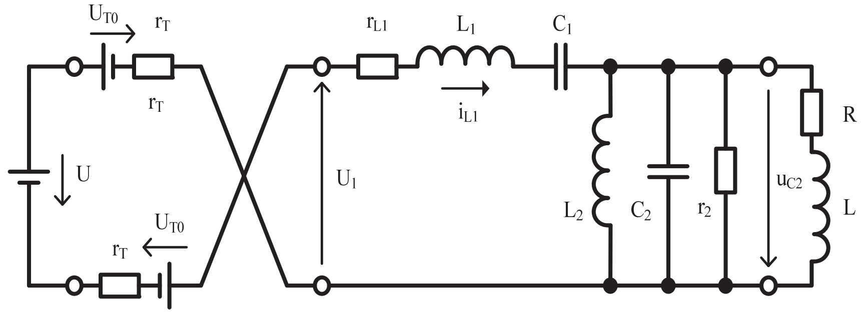


Fig. 3. Principle schematic connections of the single-phase voltage inverter and the output filter for negative half-period of source voltage

2.2 Considering non-linearity of magnetic circuit as a result of saturation of inductance L_1

It is well known that magnetizing characteristic B = f(H) is non-linear one, Fig. 4.

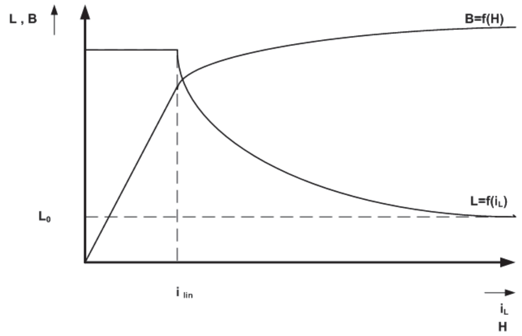


Fig. 4 Dependences of B = f(H) and L = f(i_L)

Since the inductance is derivative of magnetic flux

$$L = \mathrm{d}(\Psi)/\mathrm{d}i \qquad (9)$$

where $\Psi = N.S.B$ and i is the current of inductor,

and considering saturation of L_1 (L_1 = f(i_{L1})), then system matrices will be as

$$\mathbf{A}_{L_1}^{nonlin}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{L_2} & 0 \\ \frac{1}{C_1} & 0 & 0 & 0 & 0 \\ \frac{1}{C_2} & \frac{-1}{C_2} & 0 & \frac{-1}{C_2.r_2} & \frac{-1}{C_2} \\ 0 & 0 & 0 & \frac{1}{L} & \frac{-R}{L} \end{pmatrix} \qquad \mathbf{B}_{L_1}^{nonlin}=\begin{pmatrix} -r_1 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}\cdot\frac{1}{L_1(i_{L1})} \tag{10a,b}$$

and $\bar{u}=(i_{L1}\ u_1\ u_{C1}\ u_{C2}\ 0)^{\mathrm{T}}$. (10c)

2.3 Considering non-linearity owing to voltage U_{T0} and inductance L_1

Let's assume dependency of u_{T} on i_{L1} with the respect to Fig. 5

$$i_{L1}^{nonlin}=I_S\left(e^{k\,u_T^{nonlin}}-1\right) \tag{11}$$

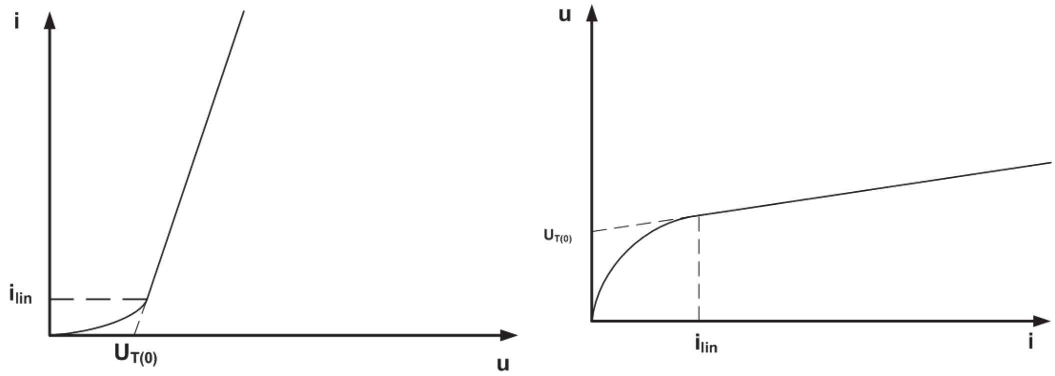


Fig. 5 Dependences of i_{L1} *on* u_{T} *and inverse function i.e.* $u_{\mathrm{T}}=\mathrm{f}(i_{\mathrm{L}})$

By separation of variables in (11) using natural logarithm function

$$e^{k\,u_T^{nonlin}}=\frac{1}{I_S}i_{L1}^{nonlin}+1$$

$$\ln e^{k\,u_T^{nonlin}}=\ln\left(\frac{1}{I_S}i_{L1}^{nonlin}+1\right)$$

one obtains $u_{\mathrm{T}}^{\text{non-lin}}$

$$u_T^{nonlin}=\frac{1}{k}\ln\left(\frac{1}{I_S}i_{L1}^{nonlin}+1\right) \tag{12a,b,c}$$

Then fictitious exciting vector is

$$\bar{u}=\left(i_{L1}\ u_1^{nonlin}\ u_{C1}\ u_{C2}\ 0\right)^{\mathrm{T}} \tag{13a}$$

where $u_1^{\text{non-lin}}$ for positive half period

$$u_1^{nonlin}=U-2u_T=U-2\left(u_T^{nonlin}+r_T\,i_{L1}^{nonlin}\right) \tag{13b}$$

and $u_1^{\text{non-lin}}$ for negative half period

$$u_1^{nonlin} = -(U - 2u_T) = -\left[U - 2\left(u_T^{nonlin} + r_T\, i_{L1}^{nonlin}\right)\right] \tag{13c}$$

3 Results of simulation experiments and verifications

Following figures show both the steady-state waveforms of input voltage u_1, input current i_{L1} and output voltage u_{C2} for linear system, and also for considered non-linear system. Only L_1 saturation non-linearity has been taking in account during simulation.

The parameters of the simulation and filter components:

$U = 250$ V; $R = 230\ \Omega$; $r_1 = 1\ \Omega$; $r_2 = 2e4\ \Omega$; $r_T = 1e\text{-}1\ \Omega$
$L_1 = L_2 = 7.3e\text{-}1$ H; $C_1 = C_2 = 1.38e\text{-}5$ F; $L = 1\ \mu$H;

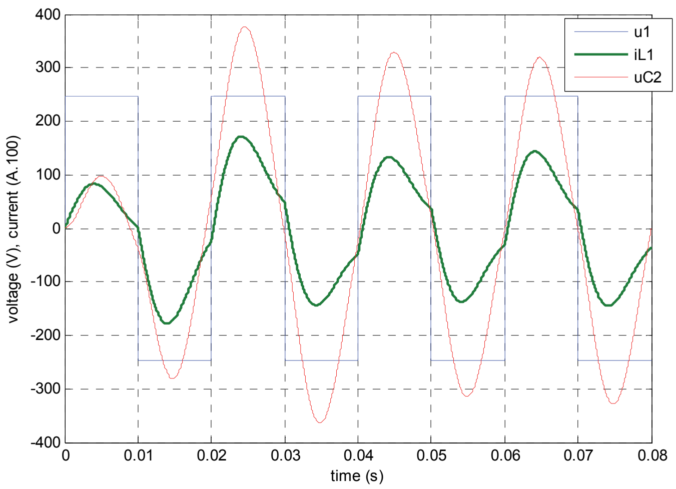


Fig. 6 Input- and output quantities of the LCLC filter for linear system

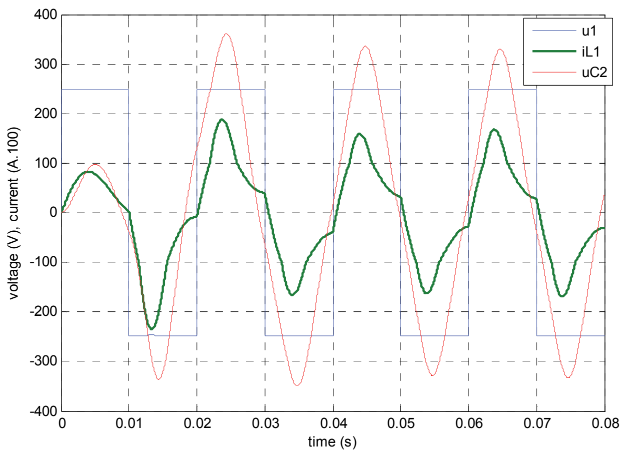


Fig. 7 Input- and output quantities of the LCLC filter considering non-linearities

4 Conclusion

A state-space transient analysis for both linear- and non-linear systems of LCLC resonant filter has been done. Euler's implicit method was used for obtaining of difference equation system of the filter. When filters are used, the output voltage has nearly harmonic waveform without substantial distortion. It has been shown that over current occurred during saturation of the series inductor could reach up one and half multiply of rated value. Simulation experiment results confirm good accordance of simulated waveforms and the theoretical assumptions.

Acknowledgement

Authors wish to thank for the financial support to Slovak grant agency VEGA through the projects No. 1/0470/09 and Education operational program Flexible and Attract Study at University of Zilina No. ITMS 26110230005 funded by European regional development fund (ERDF).

References

[1.] M. BEŇOVÁ, B. DOBRUCKÝ, E. SZYCHTA, M. PRAŽENICA: *Modelling and Simulation of HF Half-Bridge Matrix Converter System in Frequency Domain.* Logistyka, No. 6/2009, pp. 11 on CD.

[2.] P.P. BIRINGER, M.A. SLONIM: *Determination of Harmonics of Converter Current and/or Voltage Waveforms (New Method for Fourier Coefficient Calculations), Part II: Fourier Coefficients of Non-Homogeneous Functions*. IEEE Transactions on Industry Applications, *Vol. IA-16*, No. 2, March/April (1980), 248-253.

[3.] B. DOBRUCKÝ, R. ŠUL, M. BEŇOVÁ: *Mathematical Modelling of Two-Stage Converter using Complex-Conjugated Magnitudes- and Orthogonal Park/Clarke Transform Methods.* Aplimat – Journal of Applied Mathematics, Vol. 2 (2009), No. 3, 189-200.

[4.] B. DOBRUCKÝ, M. BEŇOVÁ, M. MARČOKOVÁ, R. ŠUL: *Analysis of Bipolar PWM Functions Using Discrete Complex Fourier Transform in Matlab.* Proc. of the 17th Technical Computing Prague Conf., Prague, Nov. 2009, pp. 22 on CD.

[5.] N. MOHAN, T.M. UNDELAND, W.P. ROBBINS: *Power Electronics: Converters, Applications, and Design*. John Wiley & Sons, Inc. 2003.

[6.] S.W. SMITH: *The Fourier Transform Properties and The z-Transform. Chapters 10&33 in: The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 1997-2007, pp. 185-210 and 605-630.

[7.] J. ŠKRÁŠEK, Z. TICHÝ: *Fundamentals of Applied Mathematics II* (in Czech). SNTL Publisher, Prague 1986.

[8.] M. ZÁSKALICKÁ et al.: *Analysis of Complex Time Function of Converter Output Quantities Using Complex Fourier Transform/Series.* Communications-Scientific Letters of the University of Zilina, No. 1 (2010).

Current addresses of the authors

Mariana Benova, Ing., PhD.

University of Zilina, Faculty of Electrical Engineering, Dept. Theoretical Electrotechnics, Univerzitná 1, 010 26 Žilina, Slovak Republic
Email: benova@fel.uniza.sk

Mahmud A. Rzig Abdalmula, Assoc. Prof.

University of Aljabel Algarbi, Faculty of Engineering, Sabrata, Lybia
E-mail: mahmudrzig@yahoo.com

Branislav Dobrucký, prof., Ing., PhD.

University of Zilina, Faculty of Electrical Engineering, Dept. of Mechatronics and Electronics, Univerzitná 1, 010 26 Žilina, Slovak Republic
Email: dobrucky@fel.uniza.sk

Journal of Applied Mathematics

volume IV (2011), number I

APPLICATIONS OF GAME THEORY TO NATURAL RESOURCE MANAGEMENT: THE CASE OF HIGH SEAS FISHERIES

COELHO, Manuel Pacheco, (P), FILIPE, José António, (P), FERREIRA, Manuel Alberto, M. (P)

Abstract: Property rights are in the center of fisheries management difficulties. The problem becomes more complex when fisheries are transboundary by nature. Extended Fisheries Jurisdiction gave the coastal states property rights and the potential of a sustainable management of fisheries resources. But, the High Seas remain with a statute where the principle of the "freedom of the seas" is in force. The imprecise definition of use rights in the areas of High Seas adjacent to the EEZs and the consequent difficulties in the management of the straddling stocks, made the origins of a lot of "fish wars", in the 90s. The U. N. Agreement (1995) on Transboundary Stocks and Highly Migratory Species pretended to be a formula of cooperation among interested states. The purpose of this paper is to survey the important contribution of Game Theory to study this problem of fisheries management and to highlight the possible routes for further research.

Keywords: Fisheries, Straddling Stocks, Game Theory, Optimal Control Theory

Introduction

Since the seminal papers of GORDON, SCHAEFER and SCOTT, in the 50s, the central idea in the Fisheries Economics is that, in conditions of free access and competition, the market leads to non - optimal solutions in the use of the resources. The "common property" nature of fisheries and the presence of externalities in the capture lead to market equilibrium solutions that implicate the overexploitation of the resources and overcapacity in the industry. That is, we're in the presence of Hardin's metaphor of "The Tragedy of the Commons" (HARDIN, 1968).

The property rights are in the core of fisheries management and the problem becomes more complex when fisheries are transboundary by nature. Extended Fisheries Jurisdiction gave the coastal states property-rights and the potential of a sustainable management of fisheries. However, the general evolution towards more exclusive rights didn't mean the exclusion of free access in international fisheries. The Law of the Sea (1982) doesn't exclude the principle of the "freedom of the seas" that remains in force in the High Sea.

One of the most penetrating subjects that emerge as a consequence of this statute is the management of straddling stocks. Given that the fish are endowed with mobility, it was inevitable that the coastal states, after the establishment of Economic Exclusive Zones (EEZs), verified that they were sharing some of those resources with neighbouring countries. Many coastal countries also verified that some of the acquired stocks passed the border of EEZ to the High Sea, where they were subject to the exploitation of distant waters fishing fleets from other countries. Some of those stocks moved at great distances, passing successively in EEZs of several countries and in areas of High Sea. There is no rigorous typology; we can designate the first ones as transboundary resources, the seconds as straddling stocks and the last ones as highly migratory species.

It is the second case that interests us, particularly. The development of a theory for these cases is still a work in progress, in spite of Economists and Mathematicians efforts to seek, from the end of the 70s, answers for these situations. The problem can be stated as the following:

The Law of the Sea attributes to the coastal states almost exclusive property rights on the fisheries to the 200 miles. The fundamental article (art. 56) reflects these sovereign rights to explore, to manage and to conserve the resources in EEZs.

One subject that was inconclusive, in 1982, concerned “transzonal” species. It rested for a clear debate the subject of who should be entitled management on these resources. During the Conference, the distant waters fishing nations argued that, given the mobility of those stocks, management should not be under jurisdiction of coastal states but under the competence of the Regional Fisheries Organisations. This position had the vigorous opposition of many coastal countries.

The debate took a commitment (established in the art. 64) that ended for being the focus of subsequent controversy. Art. 64 count two paragraphs seemingly contradictory. In the paragraph 1 it is said that, where International Organisation exist, coastal states should cooperate with the countries of distant fishing. For these countries it means, obviously, that, inside those Organisations, they can influence the regulation of the resources. The paragraph 2 says that the art. 64 should be applied “in addition to the other provisions of the part V of the Convention". Coastal states interpret this paragraph as implicating that the art. 56 should be applied integrally in (and out) their EEZs; that is, also to the migratory species.

An area of potential conflict grew up. The high negotiation costs implicated in the problem resolution were enough to maintain this vague stance situation. But the problem arose strongly, in the context of straddling stock fisheries. The consideration of the small importance of the highly migratory resources and the reasonable conjectures of certain coastal countries, who believed that the long distance fisheries fleets could only explore the resources of High Seas adjacent areas if it was guaranteed the access to EEZs, all showed to be wrong. Straddling stocks management was in the root-causes of serious “fish-wars” in the 90s.

In the essence, it is a problem of property rights. The conviction of coastal states, that they would be entitled “de facto" property rights on the transboundary resources, was wrong. These virtual rights ended for showing emptiness. Actually, these resources remain as “international common property" and the usual “tragedy of the commons” is well reflected in the overexploitation of these resources. The vague, imprecise form as they are defined in the Convention of 82 is in the origin of the problem. So, they can be called the" unfinished business" of the Law of the Sea (KAITALA e MUNRO (1993)).

The purpose of this paper is to survey the important contribution of Game Theory to study this problem of fisheries management and to highlight the possible routes for further research.

Shared Resources Management: Review of the Literature

The common analytical proposal has been the one that takes the basic model of the Fisheries Economics and combines it with Game Theory. In the core, the Theory grew for the transboundary resources. The importance of the straddling stocks is more recent. There is, however, a common trunk that we'll refer as Shared Resources Management.

The starting point is the Gordon–Schaefer model. We are confronted with two essential issues: the nature of free access of the resource and the consequent effect of total dissipation of rents, and the exercise of inter-temporal management of the resource, implicating a trade-off among present sacrifices and future gains.

The Game Theory, understood as an instrument of applicable analysis to situations in which a player is influenced not only for his decision and actions, as for those taken by the other players of the economic game, has an obvious value in this case.

2.1 Prisoner's Dilemma Or Cooperation?

The first subject to discuss is*: Is the cooperation worthwhile*? And one finds several alternative analysis: Colin CLARK (1980) and LEVHARI and MIRMAN (1980) classic approaches, and the developments of the Group of Helsinki (see, for example, KAITALA (1986)). The general conclusion is that the non-cooperation leads to inferior performances. The authors predict that the non-cooperation translates in results very similar to the non-regulated, free access fisheries case, that is, the dissipation of the rents.

CLARK (1980) combines the basic model of the fisheries with the Theory of Nash of non-cooperative Games, with two players (NASH, 1951).

1 THE GORDON-SCHAEFER MODEL

Consider a given country 1. The conditions of the basic model are assumed in:

$$\dot{x} = F(x) - h_{(t)}$$
$$h(t) = qE^{v}(t).x^{\phi}(t)$$

The first equation represents the dynamics of the resource as a function of the natural growth of the species and of the capture. The function of natural growth of the species, F(x), is given by a differential equation that relates the growth of the stock with the dimension of the biomass in every moment. In the model of Schaefer, a quadratic function is used that, integrated, drives to the popular logistic curve of Lotka/Volterra.

The second equation can be identified as the production function of the fishery, the capture depending on the stock dimension, on the level of applied effort and on a capture-ability coefficient specific for each species.

x represents the biomass*, h(t)* the capture, *q* the capturability coefficient and E(t) the measure of effort. The exponents *v* and *ø* are, for hypothesis, same to 1.

Supposing that the fishery is made by the country 1 alone:

The function of total cost is

$$C(t) = a_1 E(t),$$

a_1 is the unit cost of the effort, constant. This implies that the supply of effort is perfectly elastic. The same condition is put as for the demand, being *p* the fixed, constant, price of the fish.

The country 1 objective is the maximisation of the liquid benefits along the time:

$$\text{Max } PV_1 = \int_0^\infty e^{-\delta_1 t}\left(p - c_1(x)\right)h(t)\, d_t$$

δ_1 is the social discount rate in the country 1 and $c_1(x)$ it is the capture unit cost.

The Optimal Control Theory problem can be solved using the Maximum Principle of Pontryagin. The solution is the modified Golden Rule

$$F'(x*1) - \frac{c'\left(x*_1\right)F\left(x*_1\right)}{p - c_1\left(x*_1\right)} = \delta_1$$

This equation establishes the rule of resource use, the way as the society owes to invest/ disinvest in the resource along the time. The right side of the equation can be interpreted as the sustained marginal income of investing in an additional unit of the resource, divided by the cost of the investment, for that can be identified as the "rate of interest" of the resource. It is divided in two components. The first corresponds to the instantaneous marginal productivity of the resource. The 2nd term is the Marginal Stock Effect (EMS) that reflects the impact in the capture costs of the dimension of the biomass.

The approach to the optimal solution will be the fastest and we'll have a "bang-bang" result

$$h*_1(t) = \begin{cases} h_1 \text{ Max} & \text{se} \quad x_{(t)} > x*_1 \\ F(x) & \text{se} \quad x_{(t)} = x*_1 \\ 0 & \text{se} \quad x_{(t)} < x*_1 \end{cases}$$

h_1 Max is the maximum, arbitrary, capture rate.

The argument of the traditional model is that, in conditions of free access, the solution will be driven to the" bionomic equilibrium", where the rent is totally dissipated. The resource would be led to a level $x(t)=x_1$, in such a way that $c_1(x(t)) = p$.

MODEL WITH 2 PLAYERS

Supposing now the existence of a co-user 2 that shares the resource:

If the country 2 was the only user of the resource we could define the optimal biomass, according to the perspective of 2, x^*_2, in the same way we did for 1. Therefore, the "bionomic equilibrium" would be at the level x_2 in that $c_2(x_2) = p$

Supposing that there is no cooperation between the two countries and there is no communication among the managers, we are in presence of a non-cooperative game that leads us to the Prisoner's Dilemma. We fell back upon the Theory of Nash of no-cooperative games with two players.

The nature of the solution of Nash is that each player doesn't have incentive to alter his strategy given the other player's strategy. So, in the context of a fishery shared by two countries, the balance of Nash implicates, for both, capture rates (h1**(t), h2**(t)) stable.

These rates should satisfy the inequalities:

$$PV1(h1^{**}, h2^{**}) \geq (PV1(h1, h2^{**}) \qquad \text{for any } h1$$

$$PV2(h1^{**}, h2^{**}) \geq (PV2(h1^{**}, h2) \qquad \text{for any } h2$$

Supposing, for instance, that the costs of effort in the two countries are different and that barriers exist to the mobility of the labour and capital that perpetuate the inequality; and that $a_1 < a_2$ (the country 1 has low capture costs).

CLARK (1980) proves that, in these circumstances, supposing that Maxh1 and Maxh2 are sufficiently big, the solution for the no-cooperative game of Nash should satisfy

$$h1^{**}(t) = \begin{cases} h1\ Max & \text{if} \quad x > \min(x_1^*, x_2\infty) \\ F(x) & \text{if} \quad x = \min(x_1^*, x_2\infty) \\ 0 & \text{if} \quad x < \min(x_1^*, x_2\infty) \end{cases}$$

$$h2^{**}(t) = \begin{cases} h2\ max & \text{if} \quad x > x_2\infty \\ 0 & \text{if} \quad x < x_2\infty \end{cases}$$

This result means that the country with higher production costs will be pursued outside of the fisheries:

If $x_1^* < x_2\infty$ an optimal, lucky solution would be found. The result is not particularly good for the country 2 but the cooperation is punctual.

If, on the other hand, $x_1^* > x_2\infty$, then the resource will be driven to $x_2\infty$, a result clearly undesirable for both countries. We are confronted with the Prisoner's Dilemma - both players' rational decision leads to results that both consider undesirable, but that are inevitable without cooperation. Therefore, the incentive to the cooperation exists. The consequences of the non-cooperation approach the result that would be reached by a non-regulated fishery in the equivalent waters of only one country. Overexploitation and overcapacity will occur ((MUNRO, 1987, 1990); CLARK (1990)).

2.2 Cooperative Management: The Straddling Stocks Case

Recognising the advantage of the cooperation for some fisheries, we should continue an analysis of cooperative management. The process is the same, i.e., combination of the basic model of the fisheries with the Theory of Games, in this case, of the cooperative games among two people (NASH, 1953).

In the cooperative games it is assumed that the two players can communicate and are capable of establishing firm agreements.

The first subject is the one of knowing if the co-managers are willing to establish a formalised agreement, susceptible to enforcement, a coercive (binding) agreement, or simply more informal agreements, no-coercive (non-binding), without the establishment of a structure and rigorous enforcement rules.

The analysis of the cooperative fisheries is simpler in the cases of formalised, coercive agreements. There are several alternatives. The seminal analysis is the one of MUNRO (1979).

The functional objective of the two co-managers can be described in the following way:

$$\text{Max } PV_1 = \int_0^\infty e^{-\delta_1 t}\left(p - c1(x)\right)\alpha(t)\, h(t)\, dt$$

$$\text{Max } PV_2 = \int_0^\infty e^{-\delta_2 t}\left(p - c2(x)\right)\left(1-\alpha(t)\right)h(t)\, dt$$

$\alpha(t)$ it is the quota/share in total capture, for the country 1.

The co-managers have to consider two subjects:

- The division of the liquid benefits
- The possibility of different management objectives.

A potential agreement can be characterized in the following way:

$$\text{Max } PV = \beta PV1 + (1 - \beta)\, PV2$$

with $0 \le \beta \le 1$.

The objective is to maximize the global common profit. The coefficient β can be seen as the negotiation coefficient. If β=1 the preferences in terms of conservation policy of the country 1 are totally dominant, if β=0 the dominant preference is the one of the country 2. The value of this coefficient will be determined by the solution, if it exists, of the cooperative game of Nash.

Using the common procedures in treating these problems, we’ll find another modified Golden Rule. The equation is:

$$F'(x^*) - \frac{c'(x^*)F(x^*)}{p - c(x^*)} = \frac{\delta_1 \beta\alpha e^{-\delta_1 t} + \delta_2\left(1-\beta\right)\left(1-\alpha\right)e^{-\delta_2 t}}{\beta\alpha e^{-\delta_1 t} + \left(1-\beta\right)\left(1-\alpha\right)e^{-\delta_2 t}}$$

The fundamental result of the analysis is the following:

The differences in the social rates of discount produce different arrangements in the favourite strategies. Ceteris-paribus, the co-manager that uses a relatively lower discount rate prefers a conservationist policy and it is willing to invest in the resource. Therefore, the commitment favours in the immediate future the co-manager more short-sighted. Using a higher discount rate, he values more the close/near benefits. In the long term, the preferences of the more conservationist will be the more considered. According to MUNRO (1990): an "optimum-optimorum" will be found if the preferences of that, which attributed a higher value to the fishery, are predominant; he should establish the management program, and obviously, should compensate the other members, in any way. It is the "***The Compensation Principle***" [KAITALA and MUNRO, (1993)].

Also, the economic analysis indicates that the commitments in the fisheries policy through cooperative games with transfers are more efficient. The economic consequence of transfers (side payments) is that the partners are encouraged to focus on the allocation of the benefits instead of the division of capture shares.

STRADDLING STOCKS MANAGEMENT

When the resource is a straddling, the management analysis is similar to the applied to the shared resources. It is assumed that the relevant coastal state is confronted with one or more nations of distant fishing in the waters of High Seas, adjacent to EEZ.

However, an important difference in terms of the Theory of Games - it concerns the symmetry. In the relationship between two countries of contiguous EEZs there is a relationship of perfect symmetry, in the sense that each one has a power perfectly defined in his EEZ and none can use the resources of the other's EEZ without previous authorization. In the case of the straddling stocks, the relationship is asymmetrical. Nothing impedes the fleet of the coastal country in acceding to the waters of High Sea where the free access is maintained, but the fleets of distant fishing nations only enter the coastal countries EEZs if they are allowed.

In spite of this difference, the common trunk can be used with small alterations. The results don't also stand back significantly. If the non-cooperation prevails in the management of the resources the result will be the depletion of the resources.

Be noticed that, in the case of the straddling stocks, the number of participants can vary. Plus, that number can vary in time. When these issues are considered, the problem becomes significantly more complex.

Two additional issues are to be considered: the possibility of alliances between partners and the capacity of the "new entrants" to enter a given fishery. The existent analysis still constitutes an introduction to the problem (see KAITALA and MUNRO (1993). The theoretical analysis suggests some interesting conclusions:

With the number of players exceeding two, the possibilities of alliances among competitors must be considered. The analysis can get complicated considerably and, in practical terms, increases the difficulty of finding a stable cooperative arrangement.

The search of a cooperative agreement requests that each partner receives, at least, the payoff equivalent to the threaten point (the payoff of the situation of non-cooperation). But, also, that the partners of any sub-alliance obtain a result at least as good as that they would have if they chose any partner and they refused the cooperation with the third part of the organisation. That is: it is necessary an agreement whose payment is superior to the payment of the no-cooperative game and that's the largest of all the possible ones, in the possible alliances. In practice, it is the fundamental question of drawing the institutions (the International Organisations of Fisheries), and asking for their operational capacity: definition, constitution, game rules, powers, management purposes.

The problem suggested by the possibility of non-members of the Regional Organisations enter the fisheries in High Seas is very significant. Using the Game Theory we can conclude that the possibility of a member to transfer his property for the "**new entrant**" ends for increasing his bargain position, extracting a larger part from the liquid benefit resulting from the cooperation. The simple threat of transfer of his position increases his expected payoff immediately for the cooperative agreement.

This conclusion evidences the difficulty of reaching a stable agreement if in the Regional Organizations there aren't clear rules and restrictive regulation in relation with "new entrants". The blackmail strategies and bluff can happen. The negotiations become more difficult and the agreement, very unstable.

Anyway, the advantages of the cooperation are unquestionable. The process of establishing the agreements and his operational performance is a subject whose analysis stays unfinished. And it puts in relief the institutional subjects and the need of evaluating the transaction costs involved in the process of establishing the agreements.

3 The United Nations Agreement of 1995

In 1992, the United Nations accepted the accomplishment of a Conference on the Management of Transboundary Resources and Highly Migratory Species. The final Agreement came in 1995.

In the negotiations two thought schools emerged. For both it seems obvious that the management regime of the stocks in the adjacent areas of High Seas should be the same that guides the portions of that stock in EEZs.

The first school supports the "*consistency principle*". This simply states that the applied regime to the portion of the stock in the area of High Sea should be consistent with the established regime for the portion of the stock inside the EEZ. Innocuous (or maybe not), the principle seems to repeat the need of no divergence in the management regimes for the same stock. Be noticed, however, that the relationship, just as it is put, has not the two senses. By the article 56, the coastal country determines the management regime in his EEZ and, consequently, if it goes acceptance the consistence need, it owes the same regime to be in force for the remaining part of the stock. The preferences of the coastal State appear as dominant. MILES and BURKE (1989), defenders of this solution, maintain that the article 116 establishes that the coastal State has a superior right, responsibility and interest in the management of the straddling stocks, despite that the necessary distribution of competence is not prescribed.

For the maritime potencies that principle is just one more reflex of the "Creeping Jurisdiction" that shapes the recent evolution in the Maritime International Law. Some coastal countries, especially those with extensive Continental Platforms (for example, Canada), intend to maintain that principle to value his negotiation position. The distant waters fishing nations speak about co-management and justify their role in the determination of a management regime for those stocks. However, if such a rule was established, for consequence, the maritime potencies could influence the administration regime out of EEZs, and inside of them. For the coastal countries, this position, designated "*School of Article 64*", limits the sovereignty in their EEZs.

In this context, a commitment emerged:

- It maintains the free access over the 200 miles and guarantees to the Regional Fisheries Organisations the regulation power in the areas adjacent to EEZs. The largest innovation is the capacity of those Organisations to extend their rules to the non-members. Previously, a simple objection was enough for non-application of the rules, even for the members.
- It was not solved the problem of the “new entrant”. The Agreement just defined that any state with a “real interest" can be member and it should be encouraged to integrate the Organisation. However, it was not defined what means, in practice, “real interests".
- To the Organisations, the right is checked of establishing capture shares and controlling the number of boats for a given stock or area. But the Agreement doesn't say anything concerning as the decision it should be taken, if for consensus, if for majority. Once again, it will depend on the practice.
- The enforcement is another problem, because any state, by itself, can apply the law out of his territory. Each country member will have the inspection right on the ships of any other country. However, the legal action against eventual infraction only can be taken by the country of origin, of the ship found in fault. It seems that the potential effect of the enforcement is broadly bounded.

4 Routes for Further Research

The Agreement intended to promote a new cooperation formula among states interested in the resources management. Despite the cooperative atmosphere and some interesting results, this Agreement continues being the reason for discussion, especially in the context of NAFO. This situation stands for further research in, at least, three fundamental issues (Game Theory still developing a central role in every domain):

The “New Entrant” problem

Despite some interesting developments in this domain, many of them arising from the research of the so-called “Lisbon School”, this issue still requires further investigation.

The charter members of an RFMO (Regional Fisheries Management Organisation) are facing a dilemma. They can attempt to prevent non-members from becoming explicit free riders, that is, turn poachers into “game-keepers” by encouraging them to apply as new members. If the offer is too generous the existing RFMO will be undermined. If the prospective new members feel that the proposed shares are not enough they will return to explicit free riding. The solution of the problem involves the application of a coalition bargaining analysis in the form of a partition function. New developments are expected in this area.

Also the possibility of creating a market of “chart member” has to be investigated. The possibility that each member has the right of selling his chart member, creating a market for the rights to access the organization, is a matter of discussion and research because it involves a lot of problems in the division of the benefits from the cooperative use. Also, the problems of coalitions between partners.

The "Time Consistency" issue

As said before, the consideration of side payments in the models is a form of getting more stable commitments. The problem of time consistency refers to the question of knowing what are the conditions that make the commitments more stable for the future. Should the rules be more or less flexible? In a situation of uncertainty of stocks' recovery, what kind of agreement can be more trustable and less dependent of member states own motivations? What are the effects of introducing the climate change issues?

We return to the central question: coercive or non-coercive agreements? How can we design the organizations (their structures and rules) to make them more resistant to time passing and changes?

The "Interlopers" issue

This is a different form of looking at the "new entrant" issue. Suppose that a possible new entrant in the fisheries decides not to enter the RFMO and maintain a situation of free rider, exploiting the straddling stock (even with the better results that come from others' cooperative management in the RFMO). With the present rules of the game how can the "co-managers" enforce their rules to non-members?

Without a real capacity of intervention and enforcement (from detection to conviction), the efforts of cooperation will turn into disillusion and more incentives to get into free riding behavior, even for previous members of RFMO.

Most of the literature on fisheries management implicitly assumes law can be perfectly and cost-less enforced. Even when such costs and imperfections are recognised, they are not incorporated in the analysis to show how management and regulatory policies are affected by their presence. We could explore this issue with a formal model of fisheries law enforcement to show how fishing firms behave and fisheries policies are affected by costly, imperfect enforcement of fisheries law. This type of models should combine standard Economics of Fisheries analysis (Gordon/Schaefer model), Game Theory and the Theory of "Crime and Punishment" of Becker.

References

[1] AL-HUSAINI, M. (2003), "Fishery of Shared Stock of the Silver Pomfret, Pampus Argenteus, in the Northern Gulf; A Case Study", in FAO (Ed) *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[2] ARMSTRONG C. and FLAATEN, O. (1998), "The Optimal Management of a Transboundary Fish Resource: The Arcto-Norwegian Cod Stock", reprinted from *Essays on the Economics of Migratory Fish Stock,* Springer / Verlag; University of Troms⌀, pp 137-151.

[3] ARMSTRONG, C. (1994), "Cooperative Solutions in a Transboundary Fishery: The Russian-Norwegian Co-Management of the Arcto-Norwegian Cod Stock*", Marine Resource Economics,* Vol. 9, pp 329-351.

[4] BJORNDAL, T. (2003), "Management of a straddling fish stock: the case of the Norwegian Spring-spawning Herring fishery", in FAO (Ed) *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[5] BJORNDAL, T. and MUNRO, G. (2005), "The Management of High Seas Fisheries Resources and the Implementation of the UN Fish Stocks Agreement of 1995", in *Yearbook of Environmental and Resource Economics*, Edgar Elgar Pub. , pp.1-37.

[6] BJORNDAL, T., KAITALA, V. and MUNRO, G. (2000), "The Management of High Seas Fisheries", *Annals of Operations Research*, Vol. 94, pp.183-196.

[7] BRASÃO, A., DUARTE, C. and CUNHA-E-SA, M. (2000), "Managing the Northern Atlantic Bluefin Tuna Fisheries. The Stability of the UN fish Stock Agreement Solution", *Marine Resource Economics*, Vol. 15, pp.341-360.

[8] CHALULEU. J. (2003). "Shared fishery Argentina-Uruguayan Common Fishery Zone", in FAO (Ed) *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[9] CHAN, K. (1978), "The economic consequences of the 200-mile seabed zone: The replenishable resources case*", Canadian Journal of Economics*, Vol. 11, Nº 2, pp 314-318.

[10] CLARK, C. (1980), "Restricted Access to Common-Property Fishery Resources: A Game-Theoretic Analysis", in P. LIU (Ed), *Dynamic Optimization and Mathematical Economics*, New York: Plenum Press, pp 117-132.

[11] CLARK, C. (1985), *Bioeconomic Modelling and Fisheries Management,* John Wiley Sons.

[12] CLARK, C. (1990), *Mathematical Bioeconomics, The Optimal Management of Renewable Resources",* 2nd edition, Wiley-Interscience Publication, John Wiley & Sons, Inc.

[13] CLARK, C. and MUNRO, G. (1975), "The Economics of Fishing and Modern Capital Theory: A Simplified Approach*", Journal of Environmental Economics and Management*, Vol. 2, Nº 2, pp 92-106.

[14] COELHO, M. (1999) ; *A Tragédia dos Comuns Revisitada. A Pesca do Bacalhau na Terra Nova: Consequências do Regime das 200 Milhas*, ISEG/UTL, Lisboa.

[15] COELHO, M., FILIPE, J., and FERREIRA, M. (2007), "Fisheries Management and "Creeping Jurisdiction": The Statute of the Continental Platform Revisited", *Proceedings of the APFA 6 - Applications of Physics in Financial Analysis Conference*, ISCTE, Lisboa.

[16] COELHO, M., FILIPE, J. and FERREIRA, M., (2009) "Creeping Jurisdiction: The Enlargement of Economic Exclusive Zones"; *Proceedings do 15º Congresso da APDR (Associação Portuguesa de Desenvolvimento Regional),* , Cidade da Praia, Cabo Verde.

[17] COELHO, M., FILIPE, J., FERREIRA, M. and PEDRO, M. (2008), "Illegal Fishing: An Economic Analysis"**,** *APLIMAT, Journal of Applied Mathematics*, Volume 1, Number 2, pp. 167-173.

[18] COELHO, M. and LOPES, R. (2002), "Straddling Stocks and the Management of High Sea Fisheries", *Annual Conference of the European Association of Fisheries Economists*, EAFE / Universidade do Algarve - Faculdade de Economia, http://www.ualg.pt/feua/uk/eafe/

[19] CROTHERS, G. and NELSON, L. (2007), “High Seas Fisheries Governance: A Framework for the Future”, *Marine Resource Economics*, Vol. 21, pp.341-353.

[20] CROWLEY, R. and PALSSON, H. (1992), "Rights Based Fisheries Management in Canada", *Marine Resource Economics,* Vol. 7, pp 1-21.

[21] DOCKNER, E., FEICHTINGER, G. and MEHLMANN, A. (1989), “Noncooperative Solutions for a Difference Game Model of Fishery*”, Journal of Economic Dynamics and Control*, Vol. 13, Nº 1, pp 1-20.

[22] DUARTE, C., BRASÃO, A. and PINTASSILGO, P. (2000), “Management of the Northern Atlantic Bluefin Tuna: An Application of C-Games”, *Marine Resource Economics,* Vol.15, pp.21-36.

[23] EHTAMO, H. and HÄMÄLÄINEN, R. (1993), “A Cooperative Incentive Equilibrium for a Resource Management Problem*”, Journal of Economic Dynamics and Control*, Vol. 17, pp 659-678.

[24] EKERHOVD, N. (2008), *Essays on the Economics of Shared Fishery Resources*, PhD Thesis, Norwegian Scholl of Economics and Business Administration, Department of Economics, Bergen.

[25] FAO (2003), *Report of the Norway - FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, FAO, Committee on Fisheries, Rome.

[26] FILIPE, J. , FERREIRA, M. and COELHO, M. (2009), “A Teoria dos Jogos e os Comuns da Pesca”, in *Temas em Métodos Quantitativos-6*, Sílabo Editora.

[27] FISCHER, R. and MIRMAN, L. (1992), "Strategic Dynamic Interaction: Fish Wars*", Journal of Economic Dynamics and Control*, Vol. 16, Nº 2, pp 267-287.

[28] FISHER, R. and MIRMAN, L. (1996), "The Complete Fish Wars: Biological and Dynamic Interactions"*, Journal of Environmental Economics and Management*, Vol. 30, Nº 1, pp 34-42.

[29] GARDNER, M. (1989), "The Enterprise Allocation System in the Offshore Groundfish Sector in Atlantic Canada", in NEHER, ARNASON e MOLLETT (eds.), *Rights Based Fishing,* Kluwer Academic Publishers.

[30] GORDON, H. S. (1954), "The Economic Theory of a Common Property Resource: The Fishery*", Journal of Political Economy*, Vol. 62, pp 124-142.

[31] HÄMÄLÄINEN, R.; and KAITALA, V. (1990), "Cartels and Dynamic Contracts in Sharefishing*", Journal of Environmental Economics and Management*, Vol. 19, Nº 2, pp 175-192.

[32] HANNESSON, R. (1997), "Fishing as a Supergame*", Journal of Environmental Economics and Management*, Vol. 32, Nº 3, pp 309-322.

[33] HARDIN, G. (1968), "The Tragedy of the Commons*", Science*, Vol. 162, pp 1243-1247.

[34] HOUTTE, A. (2003), “Legal Aspects in the Management of Shared Fish Stocks - A Review”, in FAO (Ed) *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[35] INTERNATIONAL INSTITUTE FOR SUSTAINABLE DEVELOPMENT, (1995) " A summary of the final session of the Conference on straddling fish stocks and highly migratory stocks", *Earth Negotiations Bulletin,* Vol. 7, N.54, pp.1-12.

[36] KAITALA, V. (1986), "Game Theory Models of Fisheries Management - A Survey", *in BASAR, T. (ed.), Dynamic Games and Applications in Economics,* Berlin: Springer-Verlag, pp 252-266.

[37] KAITALA, V. and MUNRO, G. (1995), "The Management of Transboundary Resources and Property Rights Systems: The Case of Fisheries", in HANNA, S. e MUNASINGHE, M. (eds.), *Property Rights and the Environment*, World Bank.

[38] KAITALA, V. and MUNRO, R. (1993), "The Management of High Sea Fisheries", *Marine Resource Economics,* Vol. 8, pp 313-329.

[39] KAITALA, V. and POHJOLA, M. (1988), "Optimal Recovery of a Shared Resource Stock: A Differential Game Model With Efficient Memory Equilibria", *Natural Resource Modelling,* Vol. 3, Nº 1, pp 91-119.

[40] KRONBAK, L. and LINDROSS, M. (2007), "Sharing Rules and Stability in Coalition Games with Externalities", *Marine Resource Economics,* Vol.22, pp.137-154.

[41] LEVHARI, D. and MIRMAN, L. (1980), "The great fish war: an example using a dynamic Cournot-Nash Solution", *The Bell Journal of Economics*, vol. 11, Nº 1, pp 322-334

[42] LI, E. (1999), "Cooperative High-Seas Straddling Stock Agreement as a Characteristic Function Game", *Marine Resource Economics*, Vol.13, pp.247-258.

[43] LINDROSS, M., KRONBAK, L. and KAITALA, V. (2005), "Coalition Games in Fisheries Economics", Working Paper, mimeo.

[44] McRAE, D. and MUNRO, G. (1989), "Coastal State "Rights" within the 200 - Mile Exclusive Economic Zone" in NEHER, ARNASON e MOLLETT (eds.) *Rights Based Fishing*,, Kluwer Academic Publishers, Dordrecht.

[45] MILES, E. and BURKE, W. (1989), "Pressures on the United Convention on the Law of the Sea of 1982 arising from new fisheries conflicts", *Ocean Development and International Law*, 20, pp 343-357.

[46] MILLER, K. and MUNRO, G, (2004), "Climate and Cooperation: A new perspective on the management of shared fish stocks", *Marine Resource Economics*, Vol.19, pp.367-393.

[47] MUNRO, G. (1979), "The optimal management of transboundary renewable resources", *Canadian Journal of Economics*, Vol. 12, Nº 3, pp 355-376.

[48] MUNRO, G. (1982), *"Fisheries, extended jurisdiction and the economics of common property resources", Canadian Journal of Economics*, Vol. 15, Nº 3, pp 405-425.

[49] MUNRO, G. (1987), "The Management of Shared Fisheries Resources under Extended Jurisdiction", *Marine Resource Economics*, Vol. 3, Nº 4, pp 271-296.

[50] MUNRO, G. (1990), "The Optimal Management of Transboundary Fisheries: Game Theoretic Considerations", *Natural Resource Modelling*, Vol. 4, Nº 4, pp 403-426.

[51] MUNRO, G. (2001), "The United Nations Fish Stocks Agreement of 1995: History and Problems of Implementation", *Marine Resource Economics*, Vol.15, pp.265-280.

[52] MUNRO, G (2002), "Economics, The 1995 UN Fish Stocks Agreement and the Future of Transboundary Fishery Resources", Paper presented at the XIVth Annual Conference of European Association of Fisheries Economists, EAFE, Universidade do Algarve, Faro.

[53] MUNRO, G. (2003), On the management of shared fish stocks", in FAO (Ed*) Report of the Norway- FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[54] MUNRO, G. (2006), "Game Theory and the Development of Resource Management Policy: the Case of International Fisheries", *Proceedings of the 6th Meeting on Game Theory and Practice*, Zaragosa, Spain.

[55] MUNRO, G. (2007), "Internationally Shared Fish Stocks, The High Seas, and Property Rights In Fisheries", *Marine Resource Economics*, Vol. 22, pp. 425-443.

[56] MUNRO, G., Van HOUTTE, A. and WILLMANN, R. (2004), "The Conservation and Management of Shared Fish Stocks: Legal and Economic Aspects", *FAO Fisheries Technical Paper 465*, FAO, Rome.

[57] NAITO, T. and POLASKY, S. (1997), "Analysis of a Highly Migratory Fish Stocks Fishery: A Game Theoretic Approach", *Marine Resource Economics*, Vol.12, pp.179-201.

[58] NASH, J. (1951), "Noncooperative Games*", Annals of Mathematics,* 54, pp 289-295.

[59] NASH, J. (1953), *"*Two-Person Cooperative Games*", Económetrica*, Vol. 21, pp 128-140.

[60] PINTASSILGO, P. and DUARTE, C. (2001), "The New-Member Problem in the Cooperative Management of High Seas Fisheries", *Marine Resource Economics*, Vol. 15, pp.361-378.

[61] RANKE, W. (2003), "Cooperative Fisheries Management Issues in the Baltic Sea", in FAO (Ed) *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[62] SCHAEFER, M. (1957), *"*Some Considerations of Population Dynamics and Economics in Relation to the Management of the Commercial Marine Fisheries*", Journal of the Fisheries Research Board of Canada*, Vol. 14, pp 669-681.

[63] SCHRANK, W. (1995), "Extended Fisheries Jurisdiction: Origins of the current crisis in Atlantic Canada's fisheries*", Marine Policy*, Vol. 19, Nº 4, pp 285-299.

[64] SCOTT, A. (1955), "The Fishery: The Objectives of Sole Ownership", *Journal of Political Economy*, Vol. 63, pp 116-124.

[65] SCOTT, A. and MUNRO, G. (1985), *"*The Economics of Fishery Management*", in Handbook of Natural Resource and Energy Economics*, Vol. II, North-Holland, Amsterdam, pp 623-676.

[66] STAPLES, D. (2003), "Management of Shared Fish Stocks - Australian Case Studies", in FAO (Ed) *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[67] STOKKE, O. (2003), "Management of Shared Fish Stocks in the Barents Sea", in FAO (Ed) *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[68] STOKES, R. (1981), *"*The New Approach to Foreign Fisheries Allocation: An Economic Appraisal*", Land Economics*, Vol. 57, Nº 4, pp 568-582.

[69] SUMAILA, R., NINNES, C, and OLOFSEN, B. (2003), "Management of Shared Hake Stocks in the Bengal Marine Ecosystem", in FAO (Ed) *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002*, Rome.

[70] TRANSFORM AQORAU (2003), "Cooperative Management of Shared Fish Stocks in the South Pacific", in FAO (Ed), *Report of the Norway-FAO Expert Consultation on the Management of Shared Fish Stocks, Bergen, Norway, October 2002,* Rome.

[71] TORRES, J., SERRA, R. and BASCH, M. (2000), "Who Can Fish What and Where: Chile's Tradeoffs in High Seas Fishing of Straddling Stocks", *Marine Resource Economics*, Vol. 14, pp.245-262.

[72] TRONDSEN, T., MATTHIASSON, T. and YOUNG, J. (2005), " Towards a Market Oriented Management Model for Straddling Fish Stocks", *Marine Policy,* Vol. 30, pp.199-205.

[73] UNITED NATIONS (1995), *Agreement for the Implementation of the Provisions of the United Nations Convention on the Law of the Sea of 10 December 1982, Relating to the Conservation and Management of Straddling Fish Stocks and Highly Migratory Fish Stocks*, General Assembly, 6th session, New York, A/ CONF., 164/37, September.

Current addresses

Manuel Coelho, Professor Auxiliar
ISEG- UTL
SOCIUS
Rua do Quelhas, 6, 1200-781
LISBOA, Portugal
Phone: +(351) 213925800.
Fax: +(351) 213922808.
Email: coelho@iseg.utl.pt

José António Filipe, Professor Auxiliar
ISCTE – IUL
UNIDE – IUL
Av. Forças Armadas 1649-026 Lisboa, Portugal
Tel.+351 217 903 000
E-mail: jose.filipe@iscte.pt

Manuel Alberto M. Ferreira, Professor Catedrático
ISCTE – IUL
UNIDE – IUL
Av. Forças Armadas 1649-026 Lisboa, Portugal
Tel. +351 217 903 000
E-mail: manuel.ferreira@iscte.pt

Journal of Applied Mathematics

volume IV (2011), number I

MODELLING AND SIMULATION OF FINITE MARKOV MULTI-SERVER QUEUEING SYSTEM SUBJECT TO BREAKDOWNS

DORDA Michal, (CZ), TRESLER Filip, (CZ),
TUREK Michal, (CZ), TUREK Richard, (CZ)

Abstract. This paper is devoted to modelling and simulation of a finite Markov multi-server queueing system subject to breakdowns and with an ample repair capacity. The paper introduces a mathematical model of the studied system and a simulation model created by using software CPN Tools, which is intended for modelling and a simulation of coloured Petri nets. At the end of the paper the outcomes which were reached by both approaches are presented and statistically evaluated.

Key words. M/M/n/m, Queueing system, Breakdown, Petri net

Mathematics Subject Classification: Primary 60K25, 90B22; Secondary 68U20.

1 Introduction

The server that is working without failures is usually assumed in the queueing theory. But in many practical cases this assumption is not correct; servers are often technical devices and every technical device can be broken. Many authors studied a behavior of single server queueing systems subject to breakdowns under diverse assumptions. Modeling of multi-server queueing systems is not so often done due to mathematical complexity of their analysis.

Mitrani and Avi-Itzak [1] investigated a multi-server Markov queueing model with an infinite queue capacity, server breakdowns and an ample repair capacity. Neuts and Lucantoni [2] and Wartenhorst [3] considered a multi-server Markov queueing model with an infinite queue capacity and server breakdowns as well, but with a limited repair capacity. Some experimental outcomes related to a behavior of an unreliable multi-server Markov queueing system were presented by Mitrani and King in paper [4]. Queueing systems with two unreliable servers were studied by Madan, Abu-Dayyeh and Gharaibeh [5] and by Yue D., Yue W., Yu and Tian [6]. In paper [5] the authors assumed homogeneous servers, in paper [6] heterogeneous servers were considered. Wang and Chang [7] studied a finite Markov queueing system with balking, reneging and server breakdowns.

2 Assumptions of the model

Let us consider a finite Markov multi-server queueing system consisting of n homogenous parallel placed servers subject to breakdowns. Incoming customers can wait for the service in a finite queue with a capacity equal to m-n, customers are served one by one according to FIFO (First In – First Out) discipline. Thus there are in total m places in the queueing system.
Customers come to the system according to the Poisson process with the parameter λ, hence the customers inter-arrival times are exponentially distributed with the mean value equal to $\frac{1}{\lambda}$. The customers service times follow the exponential distribution with parameter μ, thus the mean service time is equal to $\frac{1}{\mu}$.
Each of the servers is successively failure-free and broken; let us assume that failures of the servers are mutually independent. Furthermore let us consider that a server breakdown can occur at any time. That means the server can break if it is busy or idle. Time of failure-free state is an exponential random variable with the parameter η, thus the mean value is equal to $\frac{1}{\eta}$. The server repairing time is an exponential random variable as well, but with the parameter ξ; the mean repairing time is therefore equal to $\frac{1}{\xi}$. Let us assume an ample repair capacity – repairing of each server starts immediately after the occurrence of the breakdown.
And finally let us assume the so called preemptive-repeat discipline that means the customer servicing is interrupted after the occurrence of the server breakdown (if the server is servicing a customer at this moment) and its service starts from the beginning. Generally one of the three different events can occur:

- If there is an idle server in the system, the customer will go over to the idle server.
- If there is not an idle server in the system, but there is a free queue place, the customer will come back to the queue.
- If there is neither an idle server nor a free queue place, the customer will be rejected.

3 Mathematical model

Let us consider a random variable $X(t)$ being the number of broken servers and a random variable $Y(t)$ being the number of customers finding in the system at the time t. On the basis of the assumptions established in Section 2 it is clear that $\{X(t), Y(t)\}$ is a two-dimensional Markov process with the state space

$$\Omega = \{(i, j), i = 0,1,...,n; j = 0,1,...,m - i\}.$$

The system is found in the state (i,j) at the time t if $X(t) = i$ and $Y(t) = j$, let us denote a corresponding probability $P_{(i,j)}(t)$.
Let us illustrate the queueing model graphically as a state transition diagram (see in fig. 1). The vertices represent the particular states of the system and oriented edges indicate the possible transitions with the corresponding rate.

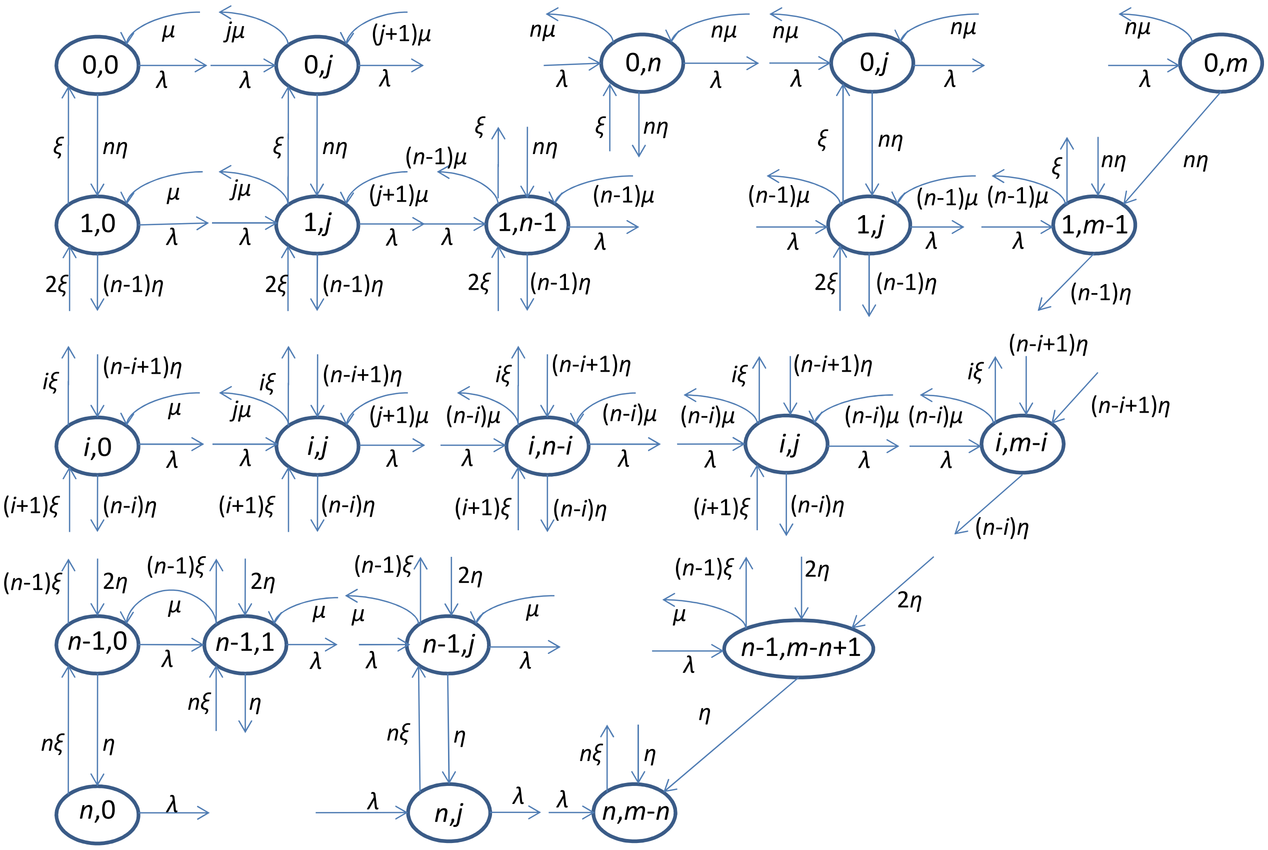


Figure 1: The state transition diagram.

On the basis of the state transition diagram we can obtain a finite system of the differential equations for the probabilities $P_{(i,j)}(t)$ depending on the time *t*. For $t \to \infty$ we get the system of the linear equations for steady state probabilities $P_{(i,j)}$ that are not dependent on the time *t*. The steady state balance equations are:

$$(\lambda + n\eta)P_{(0,0)} = \mu P_{(0,1)} + \xi P_{(1,0)}, \tag{1}$$

$$[\lambda + (n-i)\eta + i\xi]P_{(i,0)} = \mu P_{(i,1)} + (n-i+1)\eta P_{(i-1,0)} + (i+1)\xi P_{(i+1,0)} \text{ for } i = 1,\dots,n-1, \tag{2}$$

$$(\lambda + n\xi)P_{(n,0)} = \eta P_{(n-1,0)}, \tag{3}$$

$$(\lambda + j\mu + n\eta)P_{(0,j)} = \lambda P_{(0,j-1)} + (j+1)\mu P_{(0,j+1)} + \xi P_{(1,j)} \text{ for } j = 1,\dots,n-1, \tag{4}$$

$$\begin{aligned}[\lambda + j\mu + (n-i)\eta + i\xi]P_{(i,j)} &= \lambda P_{(i,j-1)} + (j+1)\mu P_{(i,j+1)} + (n-i+1)\eta P_{(i-1,j)} + \\ &+ (i+1)\xi P_{(i+1,j)} \text{ for } i = 1,\dots,n-2;\ j = 1,\dots,n-i-1,\end{aligned} \tag{5}$$

$$(\lambda + n\mu + n\eta)P_{(0,j)} = \lambda P_{(0,j-1)} + n\mu P_{(0,j+1)} + \xi P_{(1,j)} \text{ for } j = n,\dots,m-1, \tag{6}$$

$$[\lambda+(n-i)\mu+(n-i)\eta+i\xi]P_{(i,j)}=\lambda P_{(i,j-1)}+(n-i)\mu P_{(i,j+1)}+(n-i+1)\eta P_{(i-1,j)}+$$
$$+(i+1)\xi P_{(i+1,j)} \text{ for } i=1,\dots,n-1;\ j=n-i,\dots,m-i-1\,, \tag{7}$$

$$(n\mu+n\eta)P_{(0,m)}=\lambda P_{(0,m-1)}\,, \tag{8}$$

$$[(n-i)\mu+(n-i)\eta+i\xi]P_{(i,m-i)}=\lambda P_{(i,m-i-1)}+(n-i+1)\eta P_{(i-1,m-i)}+(n-i+1)\eta P_{(i-1,m-i+1)}$$
$$\text{for } i=1,\dots,n-1\,, \tag{9}$$

$$(\lambda+n\xi)P_{(n,j)}=\lambda P_{(n,j-1)}+\eta P_{(n-1,j)} \text{ for } j=1,\dots,m-n-1\,, \tag{10}$$

$$n\xi P_{(n,m-n)}=\lambda P_{(n,m-n-1)}+\eta P_{(n-1,m-n)}+\eta P_{(n-1,m-n+1)}\,. \tag{11}$$

Clearly, the probabilities $P_{(i,j)}$ must satisfy the normalization equation:

$$\sum_{i=0}^{n}\sum_{j=0}^{m-i}P_{(i,j)}=1\,. \tag{12}$$

By solving of the linear equations system, obtained from equations (1) – (10) and (12), we get steady state probabilities of the particular states of the system that are needed for performance measures computing of the studied queueing system. Please notice that the equation (11) is a linear combination of the equations (1) – (10), therefore it is dropped and replaced by the normalization equation (12).

On the basis of the steady state probabilities the following performance measures can be computed. The mean number of the costumers in the service, denoted as *ES,* can be computed according to the formula:

$$ES=\sum_{i=0}^{n-1}\sum_{j=1}^{n-i}jP_{(i,j)}+\sum_{i=0}^{n-1}(n-i)\sum_{j=n-i+1}^{m-i}P_{(i,j)}\,. \tag{13}$$

The utilization of the servers denoted as χ_s is expressed by the formula:

$$\chi_s=\frac{ES}{n}=\frac{\sum_{i=0}^{n-1}\sum_{j=1}^{n-i}jP_{(i,j)}+\sum_{i=0}^{n-1}(n-i)\sum_{j=n-i+1}^{m-i}P_{(i,j)}}{n}\,. \tag{14}$$

The mean number of the waiting customers *EL* can be expressed by the formula:

$$EL=\sum_{i=0}^{n}\sum_{j=n-i+1}^{m-i}(j-n+i)P_{(i,j)}\,. \tag{15}$$

Clearly for the mean number of the customers finding in the system *EK* it must be satisfied:

$$EK=ES+EL=\sum_{i=0}^{n}\sum_{j=1}^{m-i}jP_{(i,j)}\,. \tag{16}$$

For the mean number of broken servers *EP* it can be written:

$$EP=\sum_{i=1}^{n}i\sum_{j=0}^{m-i}P_{(i,j)}\,. \tag{17}$$

And finally the utilization of the repair capacity χ_r can be expressed by the formula:

$$\chi_r = \frac{EP}{n} = \frac{\sum_{i=1}^{n} i \sum_{j=0}^{m-i} P_{(i,j)}}{n}. \tag{18}$$

4 Simulation model

In order to validate the outcomes, which were reached by solution of the above-mentioned mathematical model, Petri net model of the studied queueing system was created by using CPN Tools – Version 2.2.0. The software CPN Tools is designed for editing, simulating and analyzing coloured Petri nets. The created simulation model is shown in fig. 2. The model is compound of 11 places and 9 transitions. The Petri net presented in fig. 2 models the unreliable M/M/3/6 (thus n=3, m=6) queueing system fulfilling the conditions mentioned in Section 2, where λ=6 h^{-1}, μ=2 h^{-1}, η=150^{-1} h^{-1} and ξ=0,2 h^{-1}.

The concrete values of the random variables are generated during the simulation through the defined function fun ET(EX) = round(exponential(1.0/EX)). We apply a minute as the unit of time and the parameter of the defined function is equal to the mean value of the exponential distribution. So the mean value of each random variable must be expressed in [min]. We substitute $\frac{1}{\lambda} = 10$ min, $\frac{1}{\mu} = 30$ min, $\frac{1}{\eta} = 9000$ min and $\frac{1}{\xi} = 300$ min in this case.

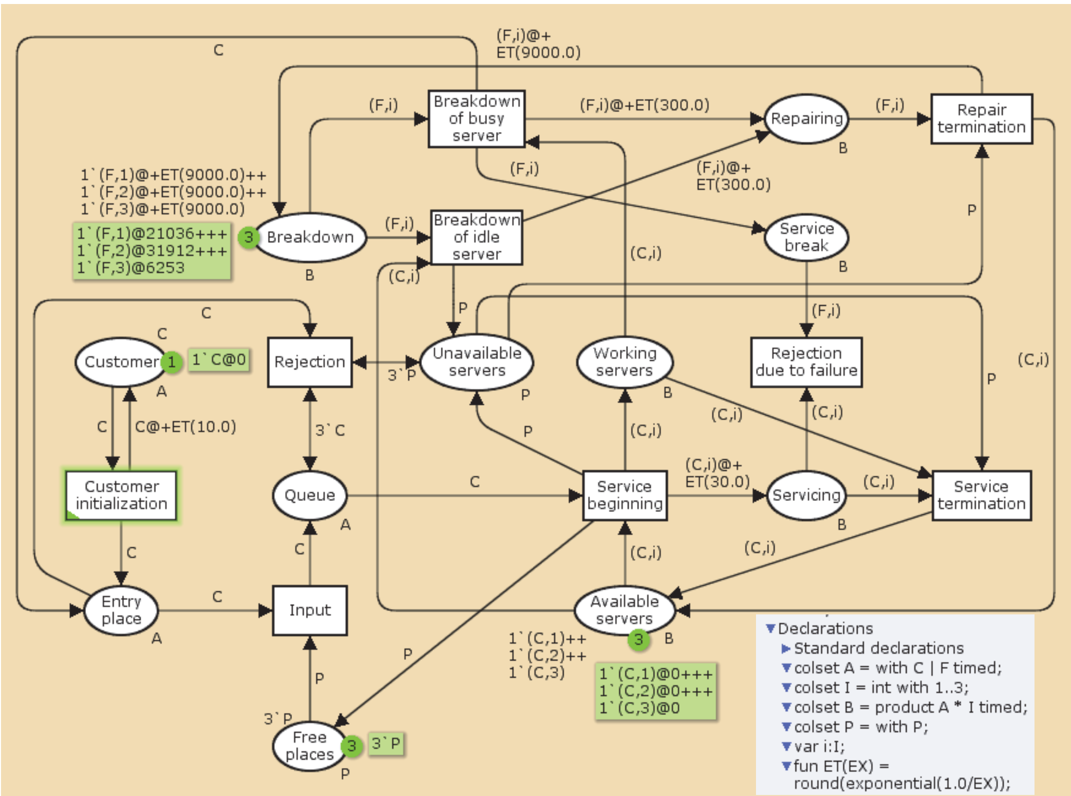


Figure 2: The simulation model created by using CPN Tools – Version 2.2.0.

The created Petri net works with following tokens:

- Tokens C represent incoming and waiting customers.
- Tokens (C,*i*), where $i \in \{1,2,3\}$, represent customers as well, but in the phase of the service process (each customer in the service is labeled by a number of the server where is serviced); these tokens are used for modelling of available and working servers as well.
- Tokens (F,*i*), where $i \in \{1,2,3\}$, model failures of the servers.
- Auxiliary tokens P serve for modelling of the free queue places and unavailable servers.

The place "Customer" with the initial marking C and the transition "Customer initialization" model the Poisson arrival process of customers; first customer comes to the system at the time 0. A customer, approaching to the queueing system finding in the place "Entry place", will come in the queue modelled by the place "Queue" if there is a token P in the place "Free places" (the initial marking of the place is equal to the queue capacity – 3`P), otherwise the customer will be rejected through the transition "Rejection". The transition "Rejection" will be enabled only if there are exactly 3 customers in the queue and all of the servers are unavailable (busy or broken) – the transition is connected with the place "Queue" by the testing arc with the arc expression 3`C and with the place "Unavailable servers" by the testing arc with the arc expression 3`P. The transition "Service beginning" will be enabled if there is a customer in the queue and a server is available (idle) – the place "Available servers" with the initial marking 1`(C,1)++1`(C,2)++1`(C,3) models available servers. In the place "Servicing" there are placed tokens modelling servicing of customers, the appropriate service time is ensured by the time stamp update through the transition "Service beginning" firing. The transition "Service termination" will be enabled if there is a customer in the service and corresponding server is working (or busy).

In the place "Breakdowns" with the initial marking 1`(F,1)@+ET(9000.0)++ 1`(F,2)@+ET(9000.0)++1`(F,3)@+ET(9000.0) there are placed tokens modelling failures of servers. When the actual simular time equals to the value of the time stamp of the token modelling breakdown of *i*-th server, its failure happens. If *i*-th server is idle at this moment (there is the token (C,*i*) in the place "Available servers"), the breakdown of the idle server will occur. If *i*-th server is busy at this moment (there is the corresponding token in the place "Working servers"), the breakdown of the busy server will occur; this breakdown will cause the rejection of corresponding customer through the transition "Rejection due to failure" firing and the input of a new costumer to the place "Entry place". The tokens (F,*i*) finding in the place "Repairing" model repairing of servers, the exponential repair time is ensured by the time stamp update. The repairing of *i*-th broken server is terminated by the transition "Repair termination" firing and corresponding token (F,*i*) is sent with the updated time stamp back to the place "Breakdowns".

There are three marking size monitoring functions that were applied for the computing of selected performance measures during simulation:

- The monitoring function which is bonded with the place "Working servers" enables the estimation of the mean number of the customers in the service *ES*.
- The monitoring function which is bonded with the place "Queue" serves to the estimation of the mean number of the waiting customers *EL*.
- The monitoring function which is associated with the place "Repairing" enables the estimation of the mean number of the servers in failure *EP*.

5 Evaluation of executed experiments

Let us consider the studied queuing system with 3 parallel servers (n=3) and in total 6 places in the system (m=6), therefore the queue capacity is equal to 3. Let us consider applied values of the random variables parameters being λ=6 h^{-1}, μ=2 h^{-1}, $\eta \in \{150^{-1}, 125^{-1}, 100^{-1}, 75^{-1}, 50^{-1}\}$ h^{-1} and ξ=0,2 h^{-1}, consequently we consider 5 model configurations differing in the parameter η.
The steady state probabilities were obtained by the solution of the linear equations system introduced in Section 3, the equations system was solved numerically by using the software Matlab. On the basis of steady state probabilities knowledge we can compute the performance measures according to the formulas (13) up to (18). Let us focus on the performance measures *ES*, *EL* and *EP*.
The experimental estimation of the performance measures we got by simulation of the coloured Petri net presented in Section 4. Thirty independent experiments were executed for each model configuration; each experiment was terminated after a million steps (a step corresponds to a transition firing).
With the view of the further statistical evaluation the normality of the simulation outcomes was tested by using the χ^2 goodness-of-fit test. As for all cases p-value is greater than 0,05 we do not reject individual hypotheses about the normal distribution. We can compute 95% confidence intervals for selected performance measures, because the normality was not rejected. The reached outcomes are summarized in the table 1 below. Please notice that all statistical computations were executed by using software Statgraphics plus 5.0.

Table 1: The reached outcomes.

η [h^{-1}]	Performance measure	Normality testing			Analytic result	95% confidence interval	Difference [%]
		Mean value μ	Dispersion σ^2	P-value			
150^{-1}	*ES*	2,43403	2,49E-05	0,61596	2,43097	(2,43216; 2,43589)	0,13
	EL	1,09976	3,11E-05	0,24144	1,09160	(1,09768; 1,10184)	0,75
	EP	0,09565	8,40E-06	0,11569	0,09677	(0,09457; 0,09673)	-1,16
125^{-1}	*ES*	2,42187	2,25E-05	0,24144	2,41931	(2,42009; 2,42364)	0,11
	EL	1,11362	3,54E-05	0,85761	1,10555	(1,11140; 1,11584)	0,73
	EP	0,11491	3,03E-05	0,91608	0,11538	(0,11286; 0,11697)	-0,41
100^{-1}	*ES*	2,40414	2,25E-05	0,11569	2,40199	(2,40236; 2,40591)	0,09
	EL	1,13288	3,86E-05	0,70293	1,12614	(1,13056; 1,13520)	0,60
	EP	0,14210	2,73E-05	0,70293	0,14286	(0,14015; 0,14405)	-0,53
75^{-1}	*ES*	2,37679	2,31E-05	0,70293	2,37355	(2,37499; 2,37858)	0,14
	EL	1,16837	3,91E-05	0,36904	1,15954	(1,16603; 1,17070)	0,76
	EP	0,18589	3,33E-05	0,52892	0,18750	(0,18373; 0,18804)	-0,86
50^{-1}	*ES*	2,32002	4,06E-05	0,61596	2,31825	(2,31764; 2,32240)	0,08
	EL	1,23248	3,05E-05	0,95798	1,22311	(1,23041; 1,23454)	0,77
	EP	0,27296	5,51E-05	0,61596	0,27273	(0,27019; 0,27573)	0,09

On the basis of the reached outcomes it is possible to say that there are statistical significant differences for all the cases, because the analytic results do not fall into the corresponding confidence intervals. However these differences are not considerably in practice as we can see in the last column of table 1. In this column the percentage differences are shown between the analytic results and the estimations of the performance measures mean values gained by the simulation experiments Let us illustrate the gained relations graphically – see figures 3, 4 and 5. In all graphs two lines are drawn, first line denoted as "Analytic result" corresponds to the value gained by the mathematical model solution and the second line "Mean value" to the estimation of the mean value (sample average) gained by the simulation experiments. The figure 3 shows the relationship between the mean number of the customers in the service *ES* and the mean time of failure-free state $\frac{1}{\eta}$. We can see that the increasing value $\frac{1}{\eta}$ causes the increasing of *ES*, this relationship could be logically expected.

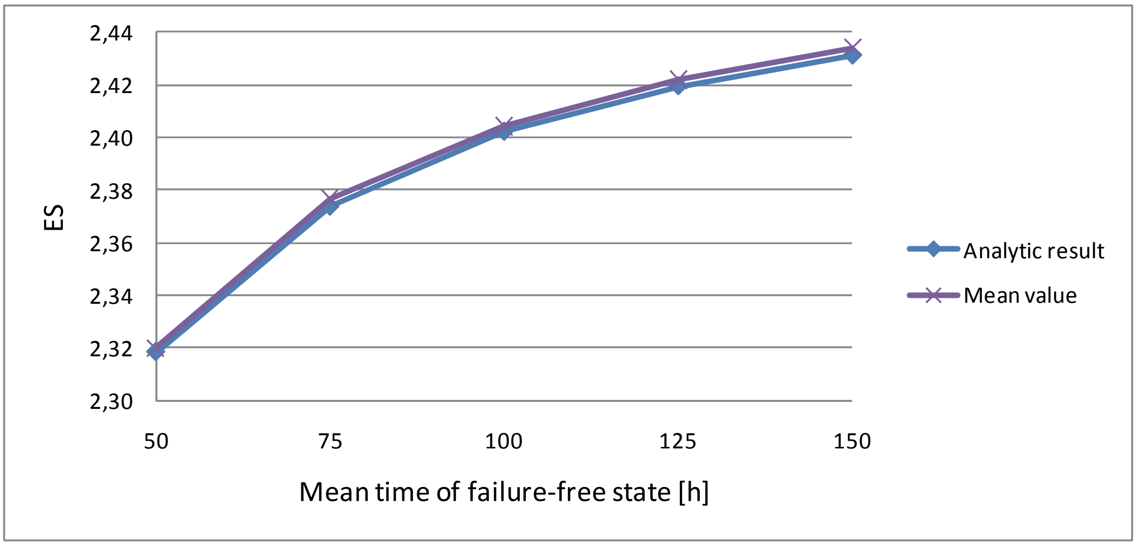


Figure 3: The relationship between *ES* and $\frac{1}{\eta}$.

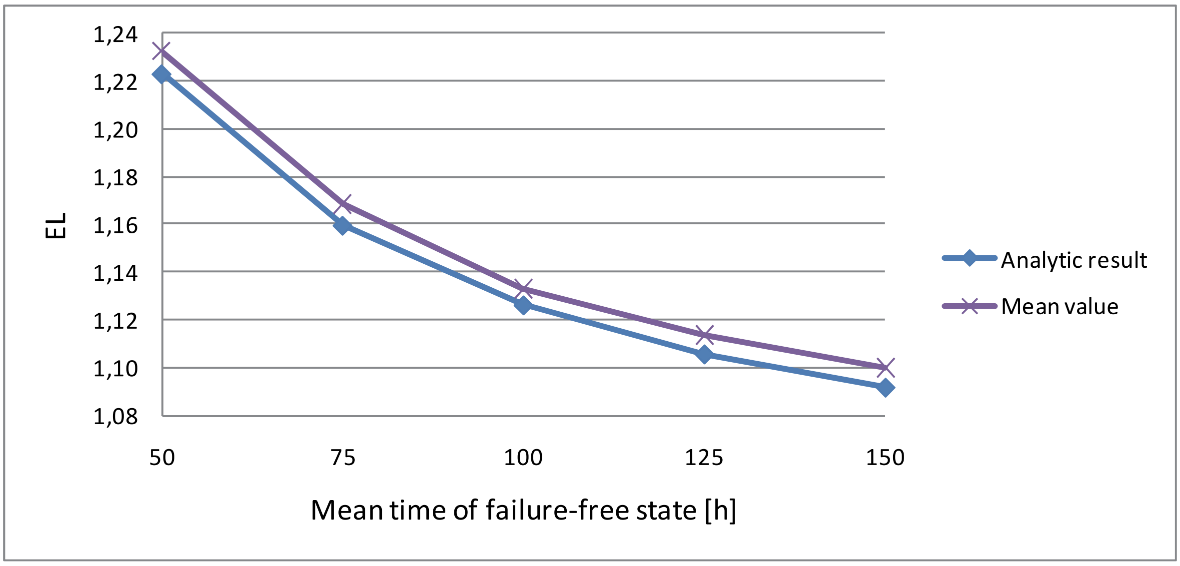


Figure 4: The relationship between *EL* and $\frac{1}{\eta}$.

The relation between the mean number of the waiting customers *EL* and the mean time of failure-free state $\frac{1}{\eta}$ is shown in figure 4. And finally the dependence of the mean number of broken servers *EP* on the mean time of failure-free state $\frac{1}{\eta}$ is drawn in figure 5. Both relationships are logically decreasing.

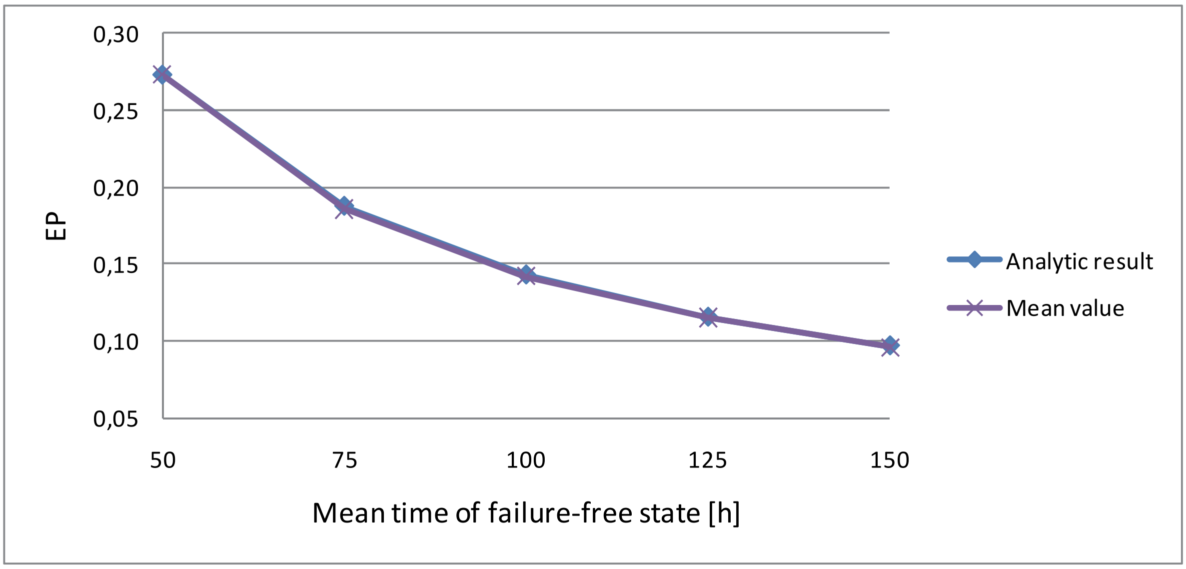


Figure 5: The relationship between *EP* and $\frac{1}{\eta}$.

5 Conclusion

There were presented two models of the M/M/*n*/*m* queueing system with the servers subject to breakdowns and an ample repair capacity in this paper – the mathematical model and the simulation model created as coloured Petri net by using CPN Tools. The major part of the paper was focused on the mathematical model of the studied system and on the description of the created simulation model. At the end of the paper reached outcomes were evaluated for the unreliable M/M/3/6 queueing system with the concrete parameters of the random variables.

Please notice that more accurate simulation outcomes we would probably obtain by using lower unit of time, for example a second. As regards to following research on the studied queueing system we would like to find the formula for the customer rejection probability above all. And finally, the model will be generalized by the assumption of an insufficient repair capacity $r<n$ that means only *r* servers can be repaired simultaneously.

References

[1.] MITRANI, I. L., AVI-ITZHAK, B.: *A many-server queue with service interruptions*. Operations Research, 16 (1968), 628 – 638.

[2.] NEUTS, M. F., LUCANTONI, D. M.: *A Markovian queue with N servers subject to breakdowns and repairs*. Management Science, 25 (1979), 849 – 861.
[3.] WARTENHORST, P.: *N parallel queueing systems with server breakdown and repair*. European Journal of Operational Research, 82 (1995), 302 – 322.
[4.] MITRANI, I., KING, J. B.: *Multiserver systems subject to breakdowns: an empirical study*. IEEE Transactions on Computers, C-32:1 (1983), 96 – 98.
[5.] MADAN, K. C., ABU-DEYYEH, W., GHARAIBEH, M.: *On two parallel servers with random breakdowns*. Soochow Journal of Mathematics, 29:4 (2003), 413 – 423.
[6.] YUE, D., YUE, W., YU, J., TIAN, R.: *A heterogeneous two-server queueing system with balking and server breakdowns*. The Eighth International Symposium on Operations Research and Its Applications, Zhangjiajie, China, 2009, pp. 230 – 244.
[7.] WANG, K., CHANG, Y.: *Cost analysis of a finite M/M/R queueing system with balking, reneging, and server breakdowns*. Mathematical Methods of Operations Research, 56 (2002), 169 – 180.

Current addresses

Michal Dorda, Ing., Ph.D.
Institute of Transport, Faculty of Mechanical Engineering,
VŠB – Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava – Poruba,
tel. number: +420 597 325 754,
e-mail: :michal.dorda@vsb.cz.

Filip Tresler, Ing.
Institute of Transport, Faculty of Mechanical Engineering,
VŠB – Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava – Poruba,
tel. number: +420 597 325 754,
e-mail: filip.tresler@email.cz.

Michal Turek, Ing.
Institute of Transport, Faculty of Mechanical Engineering,
VŠB – Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava – Poruba,
tel. number: +420 597 325 755,
e-mail: michal.turek.st1@vsb.cz.

Richard Turek, Ing.
Institute of Transport, Faculty of Mechanical Engineering,
VŠB – Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava – Poruba,
tel. number: +420 597 325 755,
e-mail: richard.turek.st@vsb.cz.

Journal of Applied Mathematics

volume IV (2011), number I

ANALYSIS OF THE AVAILABILITY FOR IMMEDIATE USE OF THE EMERGENCY VEHICLES AT AN EMERGENCY AID CENTRE

EVTIMOVA Stoyanova Veselina, (BG)

Abstract: The paper examines the availability for immediate use of the emergency vehicles at an emergency aid centre to respond to patients' calls. Determining the number of specialized emergency vehicles which are to be used over a certain time interval is of utmost importance. On the one hand, they have to guarantee that medical staff and patients can be transported in the shortest possible time (when this is necessary), on the other hand the emergency vehicles are expensive assets, which means that any extra vehicles will be an unjustified burden on the budget. The transport vehicles at an emergency aid centre from a system of discrete states which operates continuously in time. On each of the vehicles operate flows of effects with intensities $\lambda(t)$ and $\mu(t)$, respectively, which transform it from state S_1 to state S_2 and vice versa.
A system of three vehicles is analysed and its state is described using Kolmogorov's equations. The probabilities of these states are calculated and the resulting conclusions are pointed out.
The emergency aid centre should have at least three emergency vehicles at the ready at any point in time, to respond to incoming calls as it could be fatal if patients are made to queue.

Key words. Random processes, Modelling, Emergency medical care, Transport

Mathematics Subject Classification: Primary 60K35, 93A30; Secondary 90B06

1. Introduction

The national emergency aid system has been established to save the lives of suffering patients when their condition is a threat to their existence.
The emergency aid centres are of essential importance especially in the years of ongoing reforms in the public health care system and with the anticipated close down of many hospitals. The emergency aid centres have to be well- equipped and organized in order to give adequate health care [5].
Responding to emergencies is only possible if the emergency aid centres have highly qualified medical staff and an absolutely reliable, modern transport service.

Determining the number of specialized emergency vehicles which are to be used over a certain time interval is of utmost importance. On the one hand, they have to guarantee that medical staff and patients can be transported in the shortest possible time (when this is necessary), on the other hand the emergency vehicles are expensive assets, which means that any extra vehicles will be an unjustified burden on the budget [4].
A major problem which has to be solved is the organization and coordination of the medical units since the time intervals between the incoming calls and the time needed to provide the medical attendance have probabilistic characteristics.
The emergency aid centres work round-the-clock to give medical care to patients in need.

2. Description of the Model and the Investigation

In a previous paper [2] it is said that the transport vehicles at an emergency aid centre from a system of discrete states which operates continuously in time.
Let's examine one vehicle about which we can roughly assume that it is in one two possible states:

S_1 – the vehicle is in perfect working order (in operation),

S_2 – the vehicle is out of order (it is under repair).

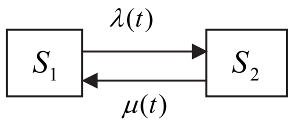


Fig.1. Graph of the states of one vehicle

The vehicle which is in state S_1 is under the impact of a flow of actions resulting in failures with intensity $\lambda(t)$, which leads it to state S_2. The same vehicle, now in state S_2, is under the effect of a flow of actions leading to restoration with intensity $\mu(t)$ (fig.1). We assume that these are two Poisson's flows and they are independent of each other [7]. We examine the processes within the system working out Kolmogorov's equations [9] for its probabilistic states and solve them with the condition that at the initial moment with $t=0$ the vehicle is in order.
For the case we examine the Kolmogorov's equations are:

$$\begin{cases} \dfrac{dp_1(t)}{dt} = \mu(t).p_2(t) - \lambda(t).p_1(t) \\ \dfrac{dp_2(t)}{dt} = \lambda(t).p_1(t) - \mu(t).p_2(t) \end{cases} \tag{1}$$

The normalizing condition on this case is as follows:

$$p_1(t) + p_2(t) = 1, \text{ so}$$
$$p_2(t) = 1 - p_1(t). \tag{2}$$

By substituting from Eq.(2) in the first equation of (1) we get:

$$\frac{dp_1(t)}{dt} + [\lambda(t) + \mu(t)].p_1(t) = \mu(t) \tag{3}$$

The linear differential Eq. (3) is a first order differential equation with variable coefficients. Since we have assumed that at initial moment the vehicle is in working order, then $p_1(0) = 1$. The solution is the following:

$$p_1(t) = e^{-\int_0^t [\lambda(\tau)+\mu(\tau)]d\tau} \left[\int_0^t \mu(\tau).e^{\int_0^\tau [\lambda(x)+\mu(x)]dx} d\tau + 1 \right] \tag{4}$$

With failure intensity $\lambda(t)$ and restoration intensity $\mu(t)$, dependent on time, it will be necessary to use numerical methods to find the solution Eq. (4).
We will confine to studying the partitive case when the two intensities $\lambda(t)$ and $\mu(t)$ do not depend on time, i.e.

$$\lambda(t) = \lambda = const, \quad \mu(t) = \mu = const\,. \tag{5}$$

Then the differential Eq. (3) assumes the following Eq. (6):

$$\frac{dp_1(t)}{dt} + [\lambda + \mu].p_1(t) = \mu \tag{6}$$

Following the solution of Eq. (6) we get

$$p_1(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu}.e^{-(\lambda+\mu).t} \tag{7}$$

After a substation from Eq. (7) in Eq. (2) the result is:

$$p_2(t) = 1 - p_1(t) = \frac{\lambda}{\lambda + \mu}\left[1 - e^{-(\lambda+\mu).t}\right]. \tag{8}$$

As it is expected that most of the time the vehicle will be in working order, then $\mu > \lambda$. When $t \to \infty$ the system switches into a stationary regime, where the probabilities p_1 and p_2 do not depend on time any more. From Eq. (7) and Eq. (8) we get the following Eq.(9) and Eq.(10):

$$p_1 = \lim_{t\to\infty} p_1(t) = \frac{\mu}{\lambda + \mu}$$

$$p_2 = \lim_{t\to\infty} p_2(t) = \frac{\lambda}{\lambda + \mu}.$$

In a stationary regime the vehicle will change its state from S_1 to S_2 and vice versa, but the probabilities of these states will not depend on time any more [3]. These probabilities can be interpreted as the vehicle's average relative time of being in state S_1 and S_2 respectively.
The graph in fig.2 is constructed with $\lambda = 0.01$ and $\mu = 0.15$.
Conclusion: The graph in fig.2 shows that in a stationary regime, the probability of the vehicle to be in working order is 0.94. The time in twenty-four-hour intervals is marked on the abscissa.
In a previous paper [6] it has been estimated that with three and more teams available for duty at the Emergency aid centre in Ruse, the average queuing time for the patient before he is taken care of, is

practically nought. This is the reason why we will examine a system of three vehicles and we will investigate with what probabilities it will be in each of its possible states.
Each vehicle breaks down independently of the other two. We will assume that the failure flows of vehicles are Poisson's flows [8] with variable intensity $\lambda_1(t), \lambda_2(t), \lambda_3(t)$. After the breakdown each vehicle is repaired and the restoration flows are considered to be Poisson's flows, too with intensities $\mu_1(t), \mu_2(t), \mu_3(t)$.

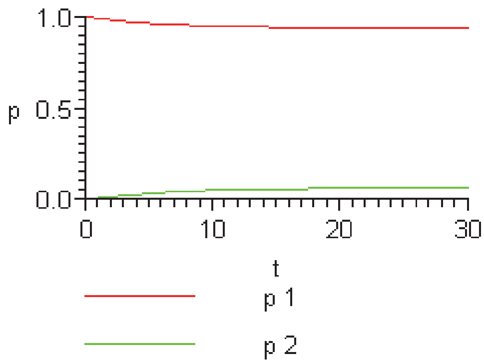


Fig.2. Graphics of the probabilistic states of one vehicle

We examine the following states of the system:

S_1 -all vehicles are in working order,

S_2 -vehicle 1 is out of order, vehicles 2 and 3 are in working order,

S_3 - vehicle 2 is out of order, vehicles 1 and 3 are in working order,

S_4 - vehicle 3 is out of order, vehicles 1 and 2 are in working order,

S_5 -vehicles 1 and 2 are out of order, vehicle 3 is in working order,

S_6 - vehicles 1 and 3 are out of order, vehicle 2 is in working order,

S_7 - vehicles 2 and 3 are out of order, vehicle 1 is in working order,

S_8 - all vehicles are out of order.

The graph of the system described is shown in fig.3.

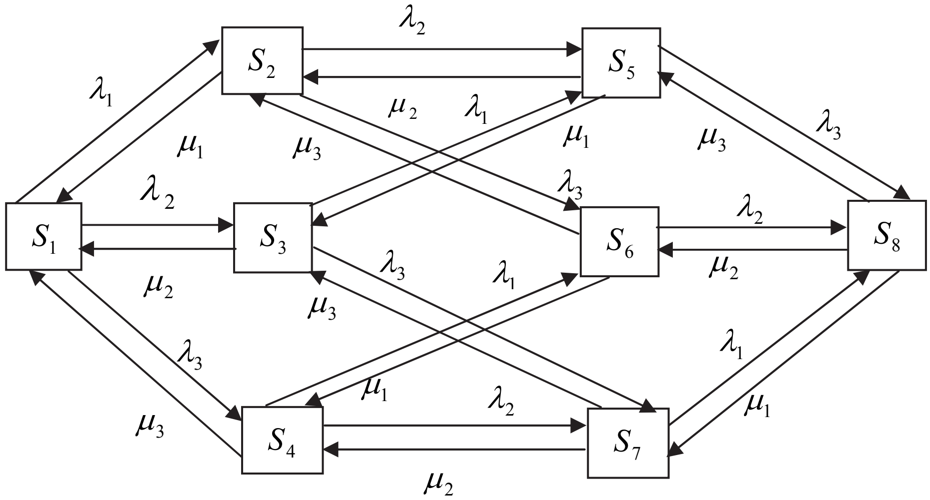


Fig.3. Graph of the states of the system of three vehicles

We work out a system of differential equations (11) of Kolmogorov for the probabilities of states S_1, S_2, ... , S_8 respectively $p_1(t)$, $p_2(t)$,..., $p_8(t)$. We will assume that at the initial moment of time $t = 0$ all vehicles were in working order. For short we write: $p_i(t) = p_i$, $\lambda_j(t) = \lambda_j$, $\mu_j(t) = \mu_j$.

The normalizing condition is $\sum_{i=1}^{8} p_i(t) = 1$.

Assuming that each vehicle is of a different make and has been in operation for a different period of time. A simplified case is examined when the failure flows and the restoration flows do not depend on time. We assume: $\lambda_1 = 0.01$; $\lambda_2 = 0.02$; $\lambda_3 = 0.01$; $\mu_1 = 0.12$; $\mu_2 = 0.15$; $\mu_3 = 0.18$;

As we assumed that at the initial moment $t = 0$ all vehicles are in working order, then we solve the system (11) at the following initial conditions: $p_1(0) = 1$, $p_2(0) = p_3(0) = ... = p_8(0) = 0$.

$$\begin{cases} \dfrac{dp_1(t)}{dt} = \mu_1.p_2 + \mu_2.p_3 + \mu_3.p_4 - (\lambda_1 + \lambda_2 + \lambda_3).p_1 \\ \dfrac{dp_2(t)}{dt} = \lambda_1.p_1 + \mu_2.p_5 + \mu_3.p_6 - (\mu_1 + \lambda_2 + \lambda_3).p_2 \\ \dfrac{dp_3(t)}{dt} = \lambda_2.p_1 + \mu_1.p_5 + \mu_3.p_7 - (\mu_2 + \lambda_1 + \lambda_3).p_3 \\ \dfrac{dp_4(t)}{dt} = \lambda_3.p_1 + \mu_1.p_6 + \mu_2.p_7 - (\mu_3 + \lambda_1 + \lambda_2).p_4 \\ \dfrac{dp_5(t)}{dt} = \lambda_2.p_2 + \lambda_1.p_3 + \mu_3.p_8 - (\mu_1 + \mu_2 + \lambda_3).p_5 \\ \dfrac{dp_6(t)}{dt} = \lambda_3.p_2 + \lambda_1.p_4 + \mu_2.p_8 - (\mu_1 + \mu_2 + \lambda_2).p_6 \\ \dfrac{dp_7(t)}{dt} = \lambda_3.p_3 + \lambda_2.p_4 + \mu_1.p_8 - (\lambda_1 + \mu_2 + \mu_3).p_7 \\ \dfrac{dp_8(t)}{dt} = \lambda_3.p_5 + \lambda_2.p_6 + \lambda_1.p_7 - (\mu_1 + \mu_2 + \mu_3).p_8 \end{cases} \tag{11}$$

The results of the calculations are as follows:

$$p1(t) = \frac{240}{2431} e^{\left(-\frac{17}{100} t\right)} + \frac{180}{2431} e^{\left(-\frac{11}{100} t\right)} + \frac{2}{2431} e^{\left(-\frac{41}{100} t\right)} + \frac{24}{2431} e^{(-7/25\, t)} + \frac{20}{2431} e^{(-3/10\, t)} + \frac{15}{2431} e^{(-6/25\, t)} + \frac{1800}{2431} + \frac{150}{2431} e^{\left(-\frac{13}{100} t\right)}$$

$$p8(t) = -\frac{2}{2431} e^{\left(-\frac{17}{100} t\right)} - \frac{2}{2431} e^{\left(-\frac{11}{100} t\right)} - \frac{2}{2431} e^{\left(-\frac{41}{100} t\right)} + \frac{2}{2431} e^{(-7/25\, t)} + \frac{2}{2431} + \frac{2}{2431} e^{(-3/10\, t)} + \frac{2}{2431} e^{(-6/25\, t)} - \frac{2}{2431} e^{\left(-\frac{13}{100} t\right)}$$

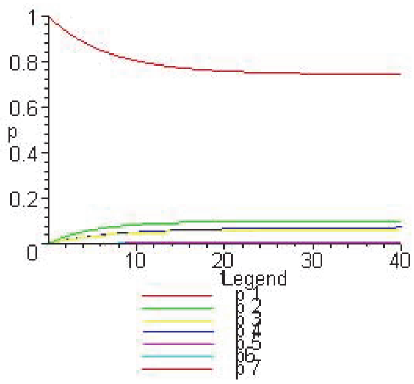


Fig.4. Graphics of the probabilistic states of a system of three vehicles

The MAPLE [1] program product has been used both for the calculations and for the graphs.
At a stationary working regime of the system the result is $p_1 = 0.74$ and $p_8 = 0.0008$. In fig.4 the time in twenty-four-hour intervals is marked on the abscissa.

3. Conclusions

At the established intensities of the failure and restoration flows, the conclusion is, that the examined system of three vehicles cannot be relied on to provide adequate attendance to the patients' needs. The emergency aid centre should have at least 3 emergency vehicles at the ready at any point in time to respond to incoming calls as it could be fatal if patients have in queue.

References

[1] AMERICH, E.: *A Guide to Maple,* Springer Verlag, New York, 1994.

[2] EVTIMOVA, V.: *Assessment of the Operational Characteristics of the Ambulance Cars at an Emergency Aid Centre*, Interdisciplinary Forum Bulgaria – Russia - Directions of Reciprocity, Ruse, pp. 135-146, 2008.

[3] HENDRICKSON, C.: *Research in transportation system characteristics and operations,* Transportation Research Part A, Vol. 19, Issues 5-6, pp. 367-369, 1985.

[4] LUBICZ, M., MIELCZAREK, B.: *Simulation modelling of medical services*, European Journal of Operational Research, Vol. 29, Issue 2, pp. 178-185, 1987.

[5] ROESSLER, M., ZUZAN, O.: *EMS systems in Germany*, Resuscitation, Vol. 68, Issue 1, pp. 45-49, 2006.

[6] SIMEONOV, D., EVTIMOVA, V., PENCHEVA, V.: *Investigation of the Influence of the Number of Teams on the Efficiency of an Emergency Aid Centre Service,* Fifth International Science and Technique Conference MOTAUTO'98, Sofia, pp. 61-65, 1998.

[7] SU, S., SHIH, C.L.: *Modelling an emergency medical service system using computer simulation,* International Journal of Medical Informatics, Vol. 72, Issues 1-3, pp. 57-72, 2003.

[8] TARAKANOV, K.V., OVCHAROV, L.A., TIRISHKIN, A.N.: *Analytical Methods of Systems Research,* Soviet Radio, Moskow, 1974.

[9] VOLKOV, I.K., ZUEV, S.M., TSVETKOVA, M.G.: *Random Processes,* The Bauman MGTU Publishing House, Moskow, 2003.

Current address

Veselina Stoyanova EVTIMOVA , PhD,
Department of Algebra and Geometry,
University of Ruse "Angel Kanchev",
8 Studentska str., Ruse 7017, Bulgaria,
tel. +359 886 267 996, +359 82 844 509, fax +359 82 845 708
e-mails: vevtimova@uni-ruse.bg, v.evtimova@gmail.com

Journal of Applied Mathematics

volume IV (2011), number I

ANALYSIS OF SOME RESULTS OF AN EXAMINATION OF THE FLOW PROCESSES AND THE TRANSPORT SERVICE OF AN EMERGENCY AID CENTRE

EVTIMOVA Stoyanova Veselina, (BG)

Abstract. The paper presents a study of the flow processes and the functioning of an emergency aid centre as a system. It examines the transport service to establish the relationship and conditions which are necessary for its most effective performance.
The amount of transport expenses (about 30% of the budget of the centre) requires an improvement of the transport service which will result in a reduction of these expenses.
The paper shows that the emergency aid system has the specific characteristics of a queuing system; therefore it uses the theory of the queuing system to describe the processes which take place in it. It defines the type of specialization of the teams by means of the apparatus of the theory of information. The possibilities of registering high priority incoming calls are investigated. The seasonal fluctuation in the type of incoming calls is studied. Conclusions made on the basis of this comprehensive research lead to suggestions for improving the organization of the transport service.

Key words. Queueing theory, Information theory, Factorial analysis, Variance analysis

Mathematics Subject Classification: Primary 60K25, 94A17; Secondary 62J10

1. Introduction

The emergency aid is a national priority and this means that the state takes the responsibility both to provide for it and to administer it. The restructuring of the emergency aid in Bulgaria aims at creating an efficient national system, organized and equipped to the best European standards, and which is capable of giving high - quality emergency aid in the shortest possible time to any Bulgarian citizen or any visitor to the country [16].
The emergency aid centre is the basic structural unit of the national emergency aid system.
Highly – qualified medical staff and adequate transport service are equally important for the successful performance of an emergency aid centre. Ambulance cars with special equipment are used as means of transport. It is of essential importance to determine the number of specialized ambulance cars that are to be used in a certain time interval as they have to guarantee the

transportation of the medical teams and the patients in the shortest possible time (when it is necessary) and, on the other hand, as ambulance cars are an expensive asset, any extra vehicles would impose an additional burden on the budget.
It is a really difficult task to organize the operation of the medical units of the probabilistic characteristics of the time intervals between the incoming calls and the time needed to provide the medical service [3].
The theoretical and practical investigations employ the apparatus and methods of: mathematical statistics, queuing theory, information theory, variance analysis, factorial analysis.
The solutions which the paper presents find application both to this particular task and to the general theory on the subject.
The approach used in this investigation can serve as a methodological basis for solving analogous problems for other emergency aid centres.
The investigation of the performance of the emergency aid centre in Ruse and of its branches, shows that the transport service can be improved so that the patient is reached and given the necessary help in the due time.

2. Some Basic Principles of the Organization of the Emergency Medical Aid

The principle of social and economic equality is pointed out as the aim of public health service systems. Evans and Wolfson [5] come up with the following algebric form of the function of usefulness for the individual:

$$U = f[x_1, x_2, ..., x_n, HS(h_1, h_2, ..., h_m, E)], \tag{1}$$

where U is a function of the usefulness as a combined notion, a sum of the usefulnesses of the listed goods; $x_1, x_2, ..., x_n$ - the consumer goods which bring about usefulness for the individual; *HS*-health (health status); $h_1, h_2, ..., h_m$ - the different kinds of health service, including medical help; *E* – environment related factors which affect human health.
People in North America have arrived at the conclusion that when the necessary specialist is at the right place at the right moment, in many cases the patient's life can be saved [13].
After an analysis of the health service systems in the world and of the emergency aid organization [1, 2, 8, 12, 18] the following conclusions can be drawn:

-All patients have equal access to emergency aid [6];

-Giving emergency aid is a complex process. It starts at the place where the suffering patient is and continues during the transportion and getting the patient into the hospital;

-The use of modern means of communication the potential of the means of transport results in the fastes possible getting to the patient [7,17].

The budget analysis of some emergency aid centres in different towns of the country shows that transport expenses come up to 30% of the budget. The high amount of these expenses requires an improvement in the transport service which will lead to their reduction.
The nature, the purpose and the conditions for effectiveness of the investigated system are the base for optimization of the transport service organization. Finding a rational solution to the problem is only possible with mathematical models.
The system has to remain flexible, allowing continual adjustment to the changes in health service and considering the existing limitations.

3. Analysis of the Emergency Medical Aid System

The emergency medical aid system is formalized. The formalization includes the following units (fig.1):

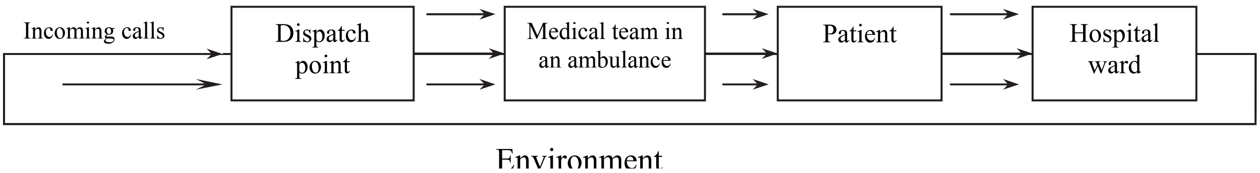


Fig.1. Diagram of the attendance to the incoming calls flow by the emergency medical aid centre

The incoming calls are received at the dispatch point of the emergency aid centre where they are classified and directed to a medical team; it has to give the patient prompt emergency treatment within a strictly limited time interval. This is how the incoming calls flow is formed. Depending on the patient's condition, the team either manages or doesn't manage to give appropriate and expeditions help. When the patient has received adequate care, the team is free to move on to attending to the next call. If no other calls have been received, the team returns to the coordinating centre. If the patient's condition requires further attention and urgent hospitalization, he/she is taken to an emergency consultative ward, and only after this task is completed the team is free and able to respond to another call.
The suggested model of service time is shown in fig.2.

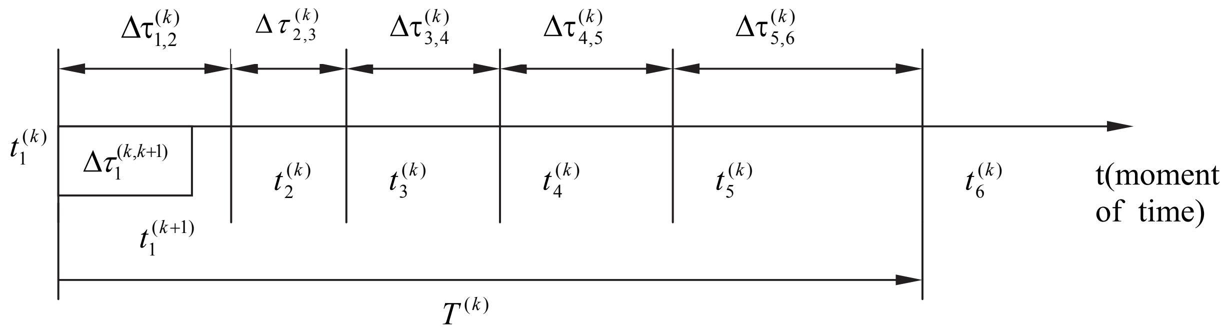


Fig.2. Graphic representation of the time moments of giving the service at *k*-number of calls

When describing the service process, the κ-number of calls is examined in a twenty-four-hour interval with n number of calls ($\kappa=1,2,\ldots,n$). If the total service time is expressed by $T^{(k)}$, $k=1,2,\ldots,n$, then the service time model for this call will have the form shown in fig.2, where

- $t_1^{(k)}$ - the moment of time for the *k*-number of calls ($\kappa=1,2,\ldots,n$);
- $t_2^{(k)}$ - the moment of time when the ambulance with the medical team sets off;
- $t_3^{(k)}$ - the moment of time they get to the patient;
- $t_4^{(k)}$ - the moment of time they leave the patient's home;
- $t_5^{(k)}$ - the moment of time the patient is driven to the hospital;
- $t_6^{(k)}$ - the moment of time the team is free to deal with the κ-number of calls.

The emergency medical aid system displays the characteristics of a queuing system. So it is appropriate to use the queuing theory to describe the processes in this system.

3.1. Analysis of the Emergency Aid System with the Help of the Queuing Theory Apparatus

The incoming calls to some emergency aid centres (in Sofia, Ruse, Silistra, Belene) have been studied. The laws of distribution of the incoming calls flow, of the time intervals between two calls and of the service time have been analysed. It is ascertained that between 7 a.m. and 11 p.m. the medical teams at the emergency aid centre in Ruse spend 23% of the time awaiting emergency calls; during the rest of the twenty-four-hour period this time is 37%.
The emergency aid centre is viewed as a multichannel queuing system with unlimited queue. The states of the system are shown with the graph in fig.3.

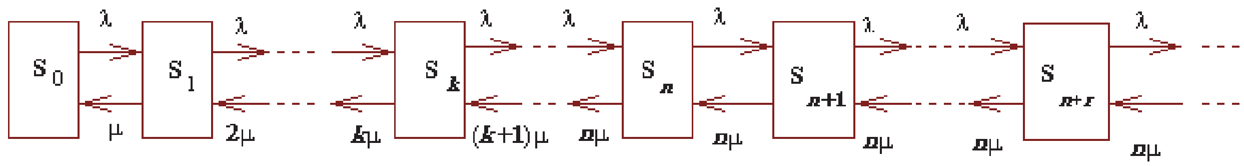


Fig.3. Graph of the system's states

The parameters which characterize the system's efficiency can be calculated with the formulae [15]:

$$L_0 = \frac{\rho^{n+1}}{n.n!\left(1-\dfrac{\rho}{n}\right)^2}; \quad \bar{k} = \rho; \quad L_c = L_o + \rho; \quad W_0 = \frac{L_0}{\lambda}; \quad W_c = \frac{L_c}{\lambda}, \tag{2}$$

where L_o - the average length of the queue; L_c-the average number of emergency calls in the system; W_o - the average waiting interval; W_c -the average time the incoming call remains in the system; $\bar{\kappa}$ -the average number of medical teams engaged in attending to calls; n-the number of teams; λ -the intensity of the incoming flow; μ - the intensity of calls which are being attended to; ρ - the redused intensity or the system's load coefficient.
The investigation of the emergency aid centre system as a queuing system has some special features. The incoming calls cannot remain the gueue, waiting to be attended to, for a long time. In some cases minutes are of vital importance to prevent death when there are casualties of major industrial accidents, natural disasters, or when people have sudden heart attacks, strokes, etc.Therfore, the time the emergency call stays in the queue, has to be reduced to a bare minimum – the time the call spends in the queue is an important factor which determines the system's efficiency.
Sometimes the system may be overloaded, i.e. the productivity of the teams attending to emergency calls, can be very high. However, this leads to a considerable waste of time for the queuing patients. This is why the problems studied in the present article, should be approached as a complex and individually, at the same time the patients' interests and the possibilities for financing the emergency units by the budget have to be considered as well. These factors serve as a starting point for finding the rational solution.
When simulating the performance of an emergency aid centre as a queuing system with high-priority calls, it is assumed that these priority calls come from citizens with higher income or higher rank, who pay for the medical service [4].

It is found out that the time spent by the patient in the queue does not depend on the call priority and is equal to $W_o = \sum_{k=1}^{n} \frac{\lambda_k}{\lambda} W_o^{(k)}$, where $\lambda = \sum_{i=1}^{n} \lambda_i$ and $W_0^{(k)}$ is the average period of time the emergency call of a particular priority, remains in the system's queue. The average waiting time of the emergency calls in the system is

$$W_c = \sum_{k=1}^{n} \frac{\lambda_k}{\lambda} W_c^{(k)} \quad . \tag{3}$$

The collected information is used to calculate the average waiting time in each of the two queues (tabl.1). The result is $W_o^{(1)} = 0.3$ hours; $W_o^{(2)} = 4{,}92$ hours. The average waiting time of a randomly chosen emergency request call is $W_o = 3{,}74$ hours. The average length of the queue for each priority level is calculated, too $L_o^{(1)} = \frac{1}{2} . \lambda_1 . W_o^{(1)} = 0{,}15$ patients; $L_o^{(2)} = \lambda_2 . W_o^{(2)} = 7{,}13$ patients.

When two emergency medical teams are engaged giving emergency aid, it turns out that the waiting time for the first and for the second team is respectively: $W_o^{(1)} = 0.085$ hours; $W_o^{(2)} = 0.16$ hours. The waiting time for a randomly chosen emergency call in the system is

$$W_o = \frac{\lambda_1}{\lambda} W_o^{(1)} + \frac{\lambda_2}{\lambda} W_0^{(2)} = 0.14 \text{ hours.}$$

(4)

The average number of emergency calls in the system is $L_o = \lambda W_o = 0.27$.

Table1.Average waiting time for one emergency call

Average waiting time for one emergency call with one medical team on duty		**Average waiting time for one emergen-cy call with two medical teams on duty**	
1^{st} priority [h]	2^{nd} priority [h]	1^{st} priority [h]	2^{nd} priority [h]
0.3	4.92	0.085	0.16

So, if two medical teams are on duty attending to emergency calls, and if priority is assigned to the calls, the average waiting time of the calls in the system is redused from 3.74 h to 0.14 h.
The analysis of the results in the table 1 makes it clear that emergency calls prioritising with one medical team on duty must not be allowed because the waiting time for patients with 2^{nd} priority is almost 5 hours and this does not comply with the requirements and regulations for emergency medical aid.

3.2. Defining the Type of Specialisation of Medical Teams with the Information Theory Apparatus

When defining the type of specialisation of medical teams, it is assumed that each incoming call brings along some information about a disease S_i, which is proportional to a fraction of the type $\frac{1}{p_i}$, whose entropy is [19]

$$e_i = -\log_2 p_i \,. \tag{5}$$

The average or the expected value of the partitive content of information of the symbols, called entropy of symbol – H, is

$$H = \sum_{i=1}^{m} p_i e_i = -\sum_{i=1}^{m} p_i \log_2 p_i \,, \text{ where } \sum_{i=1}^{m} p_i = 1; \quad H_{\max} = \log_2 m \,. \tag{6}$$

The types of incoming calls for different kinds of disease have appeared in Ruse, and the other emergency aid centres in the municipalities of Byala, Vetovo, Dve mogili and Slivo pole, with the incidence shown in table 2.

Data from the investigations carried out at the emergency aid centre in Ruse, are used to build the dependency of the entropy on the probability of a certain disease to appear (fig.4). From the continuousness of the function $-p.\log_2 p$ (fig.4) it follows that the entropy is a continuous function of the probability for the sepatate results of the experiment (when a call is received to attend to a patient with a particular kind of disease). Therefore, when very little change of these probabilities occurs, the entropy changes very little, too.

Table 2. Diseases at Ruse Emergency Aid Centre

Types of disease	**Ruse**	**Byala**	**Vetovo**	**Dve mogili**	**Slivo pole**
Hypertensions	2509	317	163	207	167
Ischemic Heart Disease	1433	76	105	56	86
Surgical Diseases	149	15	20	24	28
Ulcer/Gastritis	242	21	51	17	82
Renal/Biliary Crisis	1351	155	108	83	74
Cerebrovascular accident (CVA) / Transient disorders of cerebral cortex	795	60	74	38	104
Obstetrics and gynecological diseases	29	5	28	9	11
Mental disorders/Substance dependence	414	7	10	1	0
Other diseases	7358	903	651	465	480

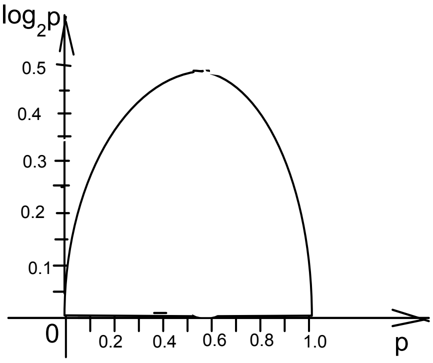


Fig.4. Dependency of the entropy on the probability of a certain type of emergency call

As a measure for the surplus of symbols in the sequence of information transfer, the value surplus of information R (7) is used.

$$R = (H_{\max} - H)/H_{\max} = 1 - H/H_{\max} \tag{7}$$

Table 3. R=$(H_{\max} - H)/H_{\max}$

Settlement System	Entropy (H)	Surplus of Information (R)
Ruse	2.096	0.339
Byala	1.851	0.416
Vetovo	1.993	0.371
Dve mogili	2.062	0.350
Slivo pole	2.348	0.259

The results analysis shows that the type of incoming calls flow is close to even distribution (R=0) and this is very strongly expressed at the Slivo pole emergency aid centre and very weakly expressed at the Byala emergency aid centre (table 3).
The results can lead to some conclusions about the type of specialisation of the medical teams. The relative part of obstetrical and gynecological disorders is the smallest one. The specialisation of the doctors on the staff at the emergency aid centre is based on this, and only one staff member is a gynecologist. The investigation shows that medical teams with broad specialisation in all areas of emergency medicine should form the staff of the emergency aid centres.

3.3. Investigation of the Seasonal Variations in the Type of Incoming Calls

The results from this investigation are necessary for the planning and management of the studied queuing system.
It is appropriate to use Spearman's rank correlation coefficient [9]:

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)}, \tag{8}$$

for the examined case, where N is the number of observations or realisations; $\sum d^2$ - the sum of the squares of the differences between the ranks of the two indications.

In order to draw any conclusions about the ascertained relationship it is necessary to check the rank correlation coefficient in connection with its statistical significance. The hypothesis H_0, which claims that there is no correlation, is used to estimate the statistical significance of Spearman's rank correlation coefficient.
The influence of the seasonal factor on the type of incoming calls is checked with the analysis of variance apparatus [14]. The deviations between the values of the result indications with the separate units and the group mean are found. These deviations characterise the random dispersion in the studied population. The intragroup variance is calculated [11]

$$\sigma_{в}^{2}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_i})^2}{n-m}, \qquad (9)$$

where i is the group number, j is the number of units in the i- group, $\overline{y_i}$ is the average value of the observed variable in the i- group, m is the number of groups, n-the number of the observed units.
The level of differences for the population as a whole is found with the calculation of the inter–group variance [10]

$$\sigma_{\text{int}\,er}^{2}=\frac{\sum_{i=1}^{m}(\overline{y_i}-\overline{y_0})^2 n_i}{m-1}, \qquad (10)$$

where n_i is the number of units in the respective group, $\overline{y_0}$ is the total average value of the observed variable number of emergency calls for a certain type of disease.
Figure 5 shows the distribution of the number of calls for a particular type of disease over the months in a year.

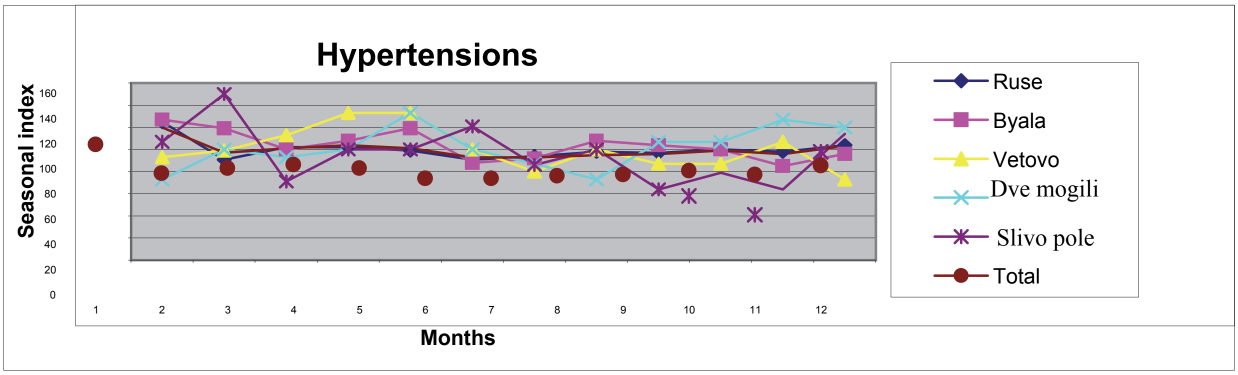


Fig.5 Distribution of emergency calls for hypertensions over a twelve-month period

The calculated values for Spearman's rank correlation coefficient, for the investigated emergency aid centre in Ruse and the district are shown in table 4.

Table 4. Spearman's rank correlation coefficients

	Types of disease	For Ruse	For Ruse District
1	Hypertensions	-0.063	-0.18
2	Ischemic heart disease	-0.014	-0.16
3	Renal and biliary crises	-0.05	-0.12
4	Cerebrovascular accident	-0.73	-0.84
5	Other diseases	-0.61	-0.52

The results analysis shows that for most diseases there is no correlation between the type of incoming emergency calls for different diseases and seasons. Therefore, the specialist on the staff should remain unchanged throughout the year.
The investigation results are generalised in table 5.

Table 5. Empirical value of the variance ratio $F_{emp} = \frac{\sigma^2_{\text{int}er}}{\sigma^2_{в}}$

	Types of disease	$F_{emp} = \sigma^2_{\text{int}er} / \sigma^2_{в}$	
		For Ruse	For Ruse District
1	Hypertensions	1.23	1.075
2	Ischemic heart disease	2.18	3.98
3	Renal and biliary crises	0.81	0.74
4	Cerebrovascular accident	2.02	2.44
5	Other diseases	0.84	1.29

The results analysis shows that the seasonal factor has considerable influence on the ischemic heart disease cases only. The result can be used by the medical teams so that they can plan for enough specialists on the staff and for the relevant medication; on the other hand the patients suffering from this disease can take the necessary precautions.
If random variations are eliminated by monthly averaging for a three-year period and when there are no seasonal variations $k_c = 0$. The stronger the seasonal variations factor, the closer the coefficient is to 1 (table 6)

Table 6. Seasonal coefficient for some types of disease

	Types of disease	For Ruse	For Ruse District
1	Hypertensions	0.063	0.05
2	Ischemic heart disease	0.079	0.075
3	Renal and biliary crises	0.074	0.1
4	Cerebrovascular accident	0.12	0.091
5	Other diseases	0.05	0.072

The results analysis from the investigation of the emergency request calls related to the types of disease, shows, in table 6, that the seasonal variations are of little consequence, and throughout the whole year the medical teams have to be ready to attend to patients from all areas of emergency medicine.

4. General Conclusions

1. The emergency aid system investigation with the queuing theory apparatus is based on sound knowledge of the distribution laws of incoming and output flows in different functioning conditions of the system.
2. The time which patients spend waiting in the system's queue is taken as a criterion for optimal management of the system. It is proved that the average time the patient remains queuing in the system, does not does not depend on the patient's priority and is $W_o = \sum_{k=1}^{n} \frac{\lambda_k}{\lambda} W_o^{(k)}$.
3. The study of the teams' specialisation has shown that the entropy is a continuous function of the probability of the experiment. The emergency aid centres have to be staffed with teams with teams with broad specialisation in all areas of emergency medicine.
4. It is established that the seasonal variations of the incoming calls, based on the types of disease, are insignificant, which is why the teams should be ready to attend to patients from all areas of the emergency medicine and at all times.

References

[1] ALI, M., MIYOSHI, C., USHIJIMA, H.: *Emergency medical services in Islamabad, Pakistan: a public-private partnership*, Public Health, Vol. 120, Issue 1, pp. 50-57, 2006.
[2] BÁEZ, A.A., LANE P.: *Emergency medical services in the city of Santo Domingo'*, Prehospital Emergency Care, Vol.6, Issue 3, pp. 336-339, 2002.
[3] BERALDI, P., BRUNI, M.E.: *A probabilistic model applied to emergency service vehicle location,* European Journal of Operational Research, Vol.196, Issue 1, pp. 323-331, 2009.
[4] DAVID, G., CHIANG, A.J.: *The determinants of public versus private provision of emergency medical services,* International Journal of Industrial Organization, Vol.27, Issue 2, pp.312-319, 2009.
[5] EVANS, R., WOLFSON, A.: *Fait hope an charity: health care in the utility function*, Department of Economics, Discussion paper №80-46, Vancouver: University of British Columbia, 1980.
[6] FELDER, S., BRINKMANN, H.: *Spatial allocation of emergency medical services: minimising the death rate or providing equal access?*, Regional Science and Urban Economics, Vol. 32, Issue 1, pp. 27-45, 2002.
[7] FRAZER, E.: *The national transportation safety board's helicopter emergency medical services public hearings*, Air medical journal, Vol.28, Issue 3, pp.128-129, 2009.
[8] HESS, A., THOMAS, T., CONTRERAS, R., GREEN, G.: *Development of emergency medical services in Guatemala,* Prehospital Emergency Care, Vol. 8, Issue 3, pp. 308-312, 2004.
[9] IAN, V., VALLE, H.: *Factor analysis and its application*, Sofia, Technica, 1974.
[10] KENDALL, M.: *Rank correlations*, Moskow, Finance and Statistics, 1975.
[11] KENDALL, M.: *Time serties*, Moskow, Finance and Statistics, 1981.
[12] LATEEF, F.: *The emergency medical services in Singapore*, Resuscitation, Vol. 68, Issue 3, pp.323-328, 2006.
[13] MEISEL, Z.F., Branas, C., Pines, J.M.: *Type of Insurance Is Associated with Ambulance Use for Transport to Emergency Departments in the United States*, Annals of Emergency Medicine, Vol. 54, Issue 3, Supplement 1, S78, 2009.

[14] MISHEV, G., TSVETKOV, St.: *Statistics for Economists*, Sofia, University Publishing House "Stopanstvo", 1998.
[15] MITKOV, At., MINKOV, D.: *Mathematical methods of engineering research*, Ruse: University Publishing House, 1985.
[16] *Ordinance №7 of the Ministry of Health* of 25.05.1994, publ. State Journal, Issue 47,1994.
[17] REDDY, M.C., PAUL, S.A., ABRAHAM, J., MCNEESE, M., DEFLITCH, C., Yen, J.: *Challenges to effective crisis management: Using information and communication technologies to coordinate emergency medical services and emergency department teams,* International Journal of Medical Informatics, Vol. 78, Issue 4, pp. 259-269, 2009.
[18] TANIGAWA, K., TANAKA K.: *Emergency medical service systems in Japan: Past, present, and future,* Resuscitation, Vol.69, Issue 3, pp. 365-370, 2006.
[19] [19] YAGLOM, A.M., YAGLOM, I.M.: *Probability and information*, Moskow, Science, 1983.

Current address

Veselina Stoyanova EVTIMOVA , PhD,
Department of Algebra and Geometry,
University of Ruse "Angel Kanchev",
8 Studentska str., Ruse 7017, Bulgaria,
tel. +359 886 267 996, +359 82 844 509, fax +359 82 845 708,
e-mails: vevtimova@uni-ruse.bg, v.evtimova@gmail.com

Journal of Applied Mathematics

volume IV (2011), number I

COGNITIVE TECHNOLOGIES FOR AGENT-BASED MODELING

FRIEDRICH Václav, (CZ)

Abstract. Exact methods are used in economics for more than three centuries. In economics as in other disciplines there were initially dominated the dynamic models based on ordinary differential equations (ODEs), so-called equation-based models. In recent years a new approach to computer modeling of economic systems have been appeared – there are so called agent-based models, where the behavior of individual entities is simulated, using the methods of the parallel computation. The paper indicated the parallels between these two types of models. Some well-known models are presented, such as the model of disease spread or a predator – prey model. In connection with this, the paper offers an interesting look at the parallels between biological and economic (or social) models.

Key words. agent-based models, cognitive technologies, NetLogo, equation-based models, LOGO program

Mathematics Subject Classification: Primary 37M05, 68U20; Secondary 37N40.

1 Equation-based Models and Agent-based Models

In economics the exact methods and models are utilized for more than last three centuries. Already the classical economic theory developed in 1st mid of 18th century by Adam Smith used mathematical formalization, but the biggest development of exact tools in economics has occurred during the 20th century, when a separate discipline on the border between economics and mathematics called mathematical economics was developed by Roy G. D. Allen [2].

1.1 Equation-based Models

Initially, in economics as in other disciplines (physics, biology, etc.) there were dominated the dynamic models based on ordinary differential equations (ODEs), so-called equation-based models. These models represent the economic system as a ”black box“ which reacts to external stimuli and changes, while the internal structure of the system is not explored.

One of the famous dynamic biological models is "predator – prey" model, which describes the population density of the prey and predators using Lotka – Volterra set of differential equations:

$$\begin{aligned} \frac{dn_1}{dt} &= n_1\left(b - k_1 n_2\right) \\ \frac{dn_2}{dt} &= n_2\left(k_2 n_1 - d\right) \end{aligned} \tag{1}$$

where n_1 is the population density of the prey, n_2 is the population density of the predators, b is the birth rate of the prey and d is the death rate of the predators [4].

Another interesting model is "S-I-R" model (Susceptible – Infected – Recovered) of the disease spread in population. This model can be described using Kermack – McKendrick set of equations:

$$\begin{aligned} \frac{dS}{dt} &= -rSI \\ \frac{dI}{dt} &= rSI - aI \\ \frac{dR}{dt} &= aI \end{aligned} \tag{2}$$

where S is a number of susceptible people, I is a number of people infected, R is the number of people who have recovered from the infection, r is the infection rate, and a is the recovery rate [3].

Although these models originally came from the area of biology, they can be also used to describe behavior of some economic systems. For example, the original biological model of disease spread can be also used as a model of marketing promotion, or the predator – prey model can be used to simulate the behavior of competitors in the market with limited resources. These parallels between the biological and economic systems are very interesting and would deserve a separate study.

The development of computer technologies in the last three decades of 20th century has been accompanied by the emergence of new numerical methods. These algorithms were used to solve the dynamic equations and their systems, which have been virtually insoluble using classical mathematical procedures and methods. The specialized computer programs called Computer Algebra Systems (CAS), such as MatLab Mathematica, or MathCAD, allow to implement these models relatively easy and with vivid graphics support. Therefore, these software tools have become popular in laboratories as well as in classrooms. The mathematical economics has developed into computational economics.

1.2 Agent-based models

In recent years, however, the classical equation-based models has been often discussed and criticized. The opponents complain that the models describe the economic system as an aggregated monolithic structure at a high level of abstraction, although it is actually composed of many separate economic subjects with individual rearing. That is why, among other things, these models have not been able to respond to turbulent changes that come with the current economic situation brings. Also the process of aggregation at the equation-based models is too abstract and not just simply understandable for non-mathematicians.

Some economists (for example so called Austrian School of Economics or famous Czech economist Tomas Sedlacek [6]) even argue that the present mathematics is unable to bring any suitable tools for new modern economic theories. But that is not true. Many times in the history mathematics was able to respond to new demands that come with the development of various scientific disciplines with suitable mathematical theories and tools (eg. fuzzy sets, theory of games, fractal theory, and others).

Such a new approach to computer modeling of economic systems there are so called agent-based models. In these models the behavior of individual entities is simulated, using the methods of the parallel computation. For many non-mathematicians this decentralized approach to economic modeling process is even easier to understand and solve than the analysis of differential equations with aggregated variables. The agent-based models are also much easier to implement some specific features and effects of present economic situation than the complex equation-based models.

The multi-agent simulation models can be created using various computer programs. In this paper one of them is focused – it is NetLogo developed at the Center for Connected Learning at Northwestern University (Chicago, USA). This program represents a new generation of LOGO computer software which has been developing for more than 40 years.

2 LOGO – a Cognitive Tool for Modeling and Simulation

The origin LOGO program was developed by Wally Feurzeig and Seymour Papert in 1967 by simplifying LISP algorithmic language into a programming tool to control cyber turtles. During four last decades several generations of young people have used different versions of LOGO program to develop their logical and algorithmic thinking. Thanks to the LOGO program also some new mathematical and computational disciplines were developed - such as fractal geometry, complex base systems, or artificial intelligence. After creating a version of LOGO with more turtles, which can interact with each other and influence their behavior and properties, it was only a short step to multi-agent modeling. With these capabilities, the current LOGO system is more than child's computer game. It has changed into a serious tool for modeling and simulation of processes and phenomena from different fields, including economics.

The author of this article has been interested in cognitive technologies including LOGO program for many years. He is a co-author of the Czech textbook of LOGO program called “LOGO for both children and adults” (1989, with Vojtěch Sedláček). At presence he is interested in the use of NetLogo program to create and simulate multi-agent models and to apply the experience from this research in education.

2.1 NetLogo for Agent-based Modeling

The first multi-agent version of LOGO language named StarLogo was developed in 1991 – 1994 at MIT Media Laboratories in Cambridge, Massachussets by Mitchel Resnick and his research group, as a programmable modeling environment for modeling and simulating natural and social phenomena. The program was created as a Java virtual machine application, so it has worked on all major platforms (Mac, Windows, Linux, et al). Many models has been developed for StarLogo as an Application Models Library.

The next generation of multi-agent LOGO environment is NetLogo as a direct successor of StarLogo. It was authored by Uri Wilensky in 1999 at the Center for Connected Learning and Computer-Based Modeling at Nortwestern University, USA, as a part of "Connected Mathematics" research project. In 2010 the 4th generation of NetLogo has been published. The NetLogo software, the documentation, and the Models Library are distributed free of charge for use by the public to explore and construct models. The program can therefore be used free of charge for educational and research purposes.

NetLogo is particularly well suited for modeling complex systems with hundreds or even thousands of agents which can be controlled independently. There are three levels of objects in NetLogo – turtles (or breeds) as micro-level individuals, patterns as macro-level objects and the observer which control all the individuals together.

NetLogo was originally developed for both the research and educational purposes in many fields of interest, including biology and medicine, physics, chemistry, mathematics and computer science, and also economics. It means that NetLogo is not only the powerful tool for researchers but it is also a very good authoring environment for education which enables students and teachers to create their own models and run simulations. Models created by users can be run as Java applets in a web browser so they can be implemented into web courses, for example in Moodle e-learning environment.

2.2 Disease spread model in NetLogo

In this article the model of disease spread created in NetLogo will be presented. As we have mentioned above, this originally biological model can be interpreted also as an economical model of marketing promotion (or general, as a model of information spread).

The model described in this section is a simplified version of S-I-R model described in 1st chapter. This model is created by students of educational subject called Logical processes and methods which the author guarantees and teaches. The creation of this model is so simple that it can be developed and tested by most the students for less than an hour.

Figure 1 shows the finished model. On the right side of the screen this is a desktop with agents in different colors (healthy or sick). On the left side there are the buttons to control model, the sliders to set parameters, and the chart, which shows the development of model variables – number of healthy and sick agents.

Figure 2 is an example of program code in NetLogo syntax. The NetLogo language is very simple, it is actually still the same language used by small children for control of the turtles in the simpler versions of LOGO program.

Multi-agent modeling allows reseachers to perform various model changes, which would be reflected into equation-models only with difficulties, if at all possible. So students can try to make different variations of the model, such as:

- disease can be spread not only by contact, but also at a distance
- disease has an incubation period, after a certain time it lapses itself

and so on.

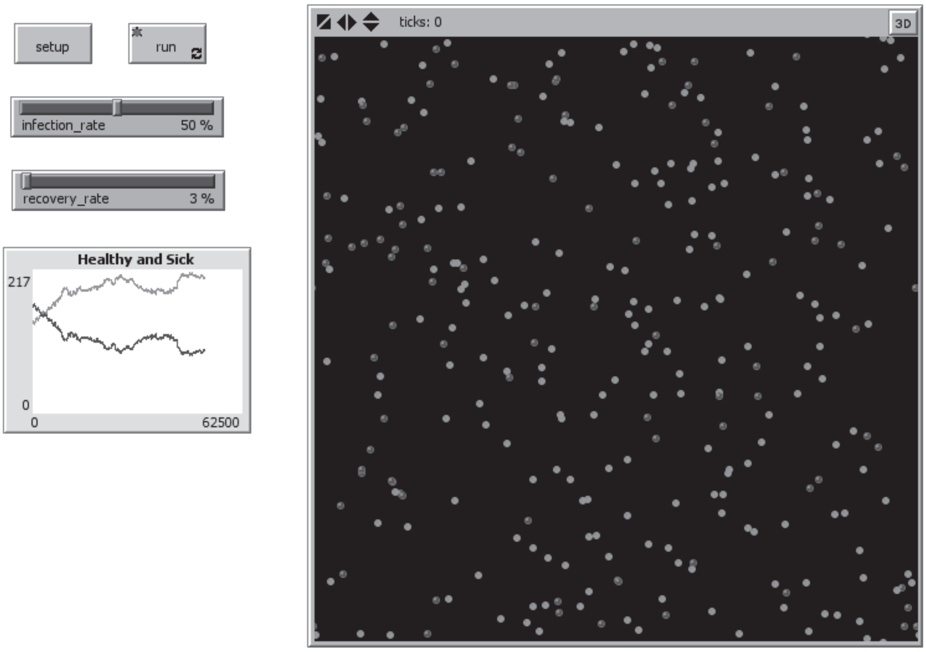


Fig. 1 Model S-I-R in NetLogo

```
to move
  fd 1
  rt random 10
  lt random 10
  collision-test
  healing
  set-current-plot-pen "infect"
    plot count turtles with [color = red]
    set-current-plot-pen "default"
    plot count turtles with [color = green]
end

to collision-test
  if count turtles-here >= 2
    [if color = red
      [ask turtles-here
         [set color red]
      ]
    ]
end
```

Fig. 2 Agent-based model in NetLogo (part)

3 Conclusion

Multi-agent modeling is a relatively new method of scientific modeling in various fields of interest including economy, which could be exploited thanks to the development of computer technology. These models deliver many advantages over classical analytical or equation-based models. They are often easier to create and simulate, they are better to describe the dynamics of the system and in some cases they even have no alternative analytical models.

Our students meet the problems of economic processes modeling and simulation in several educational subjects. Mostly they learn about analytical models based on linear relationships between variables, which can be solved as a system of linear differential or difference equations. The other approaches such as multi-agent modeling, they usually have not learnt too much. At our faculty we try to familiarize students with the NetLogo program, and multi-agent modeling, at least in some of the elective courses.

References

[1.] ABRAHAMSON, D., Wilensky, U., Levin, J. *Agent-Based Modeling as a Bridge Between Cognitive and Social Perspectives on Learning.* Annual meeting of the American Educational Research Association, Chicago, IL, April 9-13, 2007.

[2.] ALLEN, R. G. D. *Mathematical Economics.* 1st Edition. Macmillan; St. Martin's Press, 1956.

[3.] BULAVA, P. *Discrete models of infectious disease spread.* Proceedings of the 12th WSEAS International Conference On Computers. Heraklion, Greece, 2008.

[4.] COLELLA, V. S. – KLOPFER, E. – RESNICK, M. *Adventures in Modeling*. Teachers College Press. Columbia University, New York, 2001. ISBN 978-0-8077-4082-8.

[5.] KOMOSINSKI, M., ADAMATZKY, A.: *Artifical Life Models in Software*. Second Edition. Springer Verlag, Heidelberg London New York, 2009. ISBN 978-1848822849.

[6.] SEDLÁČEK, T. *Ekonomie dobra a zla : Po stopách lidského tázání od Gilgameše po finanční krizi.* Nakladatelství 65. pole, Praha 2002. ISBN 978-80-903944-3-8.

[7.] VATTAM, S., GOEL, A. K., RUGABER, S., HMELO-SILVER, C., JORDAN, R. *From Conceptual Models to Agent-based Simulations: Why and How.* Proceedings of 14th International Conference on Artificial Intelligence in Education (AIED), 2009. [online].

Current address

Václav Friedrich, Ing., Ph.D., ING-PAED IGIP
VSB – Technical University of Ostrava
Faculty of Economics
Sokolská 33
708 00 Ostrava
Czech Republic
e-mail: vaclav.friedrich@vsb.cz

Journal of Applied Mathematics

volume IV (2011), number I

LASER ENGRAVING MODELLING – COMPARISON OF METHODS FOR THE HEAT-AFFECTED AREA DETECTION

HÁJKOVÁ Jana, (CZ)

Abstract. The paper deals with the topic of laser engraving modelling. More concretely it describes methods for the automatic detection of the heat-affected area on the real samples and aims at their comparison. Real samples are used as the main input for the modelling process and the detection should be provided several times during data pre-processing. The paper outlines the format of processed data and the way of its acquisition. The main part of the paper is dedicated to three methods designed for the automatic detection of the heat-affected area and their comparison.

Key words. laser engraving, comparison, modelling, data processing, high map, detection

Mathematics Subject Classification: 68U20, 68Q68.

Introduction

The work described in this paper is a part of a larger project that deals with laser engraving control and modelling of laser engraving results. We are interested mainly in the real data exploration, comparison and modelling. We are working on finding the way how to get the optimal laser setting description to be able to model the engraving process result without providing the practical engraving. The reason is quite simple, the real experiments are expensive, time consuming and the result depends a lot on the engraving parameters and used material.

During the data pre-processing we need to detect the area modified during the engraving process by the laser beam. This detection should be used several times, e.g. during the sample parameterization or in the phase of the modelling verification and data comparison. That is why the designed method should give optimal results for different processed samples, should be quick and reliable enough. During our research we have designed several methods. Some of them were not suitable, but three of them seem to be possible to use. To choose the best one, we have decided to compare them from different points of view. This comparison was made also as a part of [5.]. Compared methods used for the automatic detection are shortly described in Section 0 and their results are compared in Section 0.

But before the detection design itself, we have to analyze the real engraved samples to learn about the structure of samples and we also have to understand the physical background of the laser engraving process. The format of data is described in Section 0 whereas Section 0 focuses on the physical fundamentals of the engraving process.

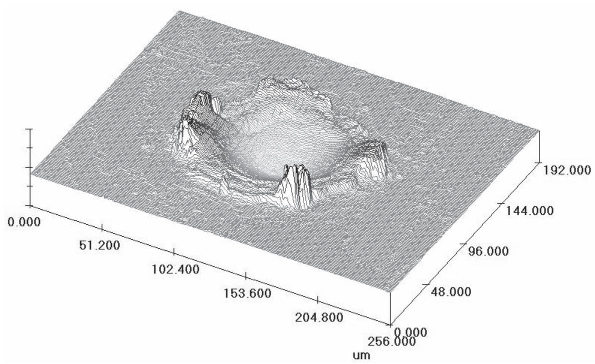
Data Acquisition and Description

For the modelling process, we use real data as an input. Data originates from the real samples engraved by a laser into the material surface. All samples are prepared under the same conditions, only the description of the experiment changes to get a larger data set. All engraved samples are measured by a confocal microscope. During the measurement, for each engraved experiment the close surroundings of the engraved area is chosen, scanned and saved in the form of a high map. This height map is formed by a matrix of real numbers, which express the heights in a uniform rectangular grid (as can be seen in Fig. 1a). The basics of data acquisition process are described in detail in [2.]. The dimension of samples reaches approximately several hundreds of micrometers.

We use the special testing data set consisting of samples engraved by the laser into a single point in the material. Such testing data should prevent potential faults caused by the external influences in a maximal possible way. The number of pulses goes in sequence, e.g. from 1 to 100. Each sample differs from the others even if the same experiment description has been engraved under the same conditions and with the same laser settings. This is the reason, why each experiment is repeated several times and so the input set should be representative enough.

To get a better imagination about the appearance of a real sample, see Fig. 1b, where the sample with 50 laser pulses engraved into one point of steel is shown in a 3D view.

a)



b)

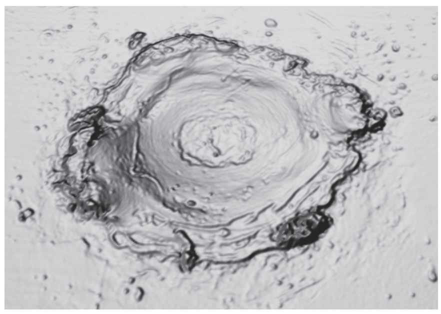


Fig. 1: High map of the sample with 50 pulses engraved into steel: a) 3D view from the confocal microscope; b) 3D view from different angle generated by our SW.

Engraving Process

Engraving is the process during which the laser beam, which is an electromagnetic radiation, affects the surface of the material. When the laser radiation strikes the material, some radiation is reflected, some absorbed and some transmitted. In our case, the most important is the absorption of the radiation, because it causes changes directly on the surface of the material. Although the physical reason depends on the type of the used material (metal, insulator or semiconductor), the effect is the same – it generates heat on the material surface. The heat generated at the surface directly affected by the laser beam is further conducted into the material.

If the laser intensity is high enough, the incident material heats, melts and if it reaches the boiling point, it starts to vaporize. These three described phases can be seen from the cross-section view in Fig. 2a-c. The solid arrows signal the laser beam direction, dashed arrows show the heat conduction in the material and the melted and vaporized material is highlighted with the gray colour. The vapour interacts with the laser, it ionizes and creates plasma. Vaporized particles, which are not affected by the laser beam, move away from the surface of the material, loose their energy and approximately 18% of them condense back to the surface. This process causes the increase of the sample surface roughness (especially in the heat-affected area closest surroundings). Moreover, the evolving vapor from the surface exerts a recoil pressure on the surface, which causes a melt expulsion, which is schematically shown in Fig. 2d. The whole process and the details related to the plasma phase are described in [1.] or [2.].

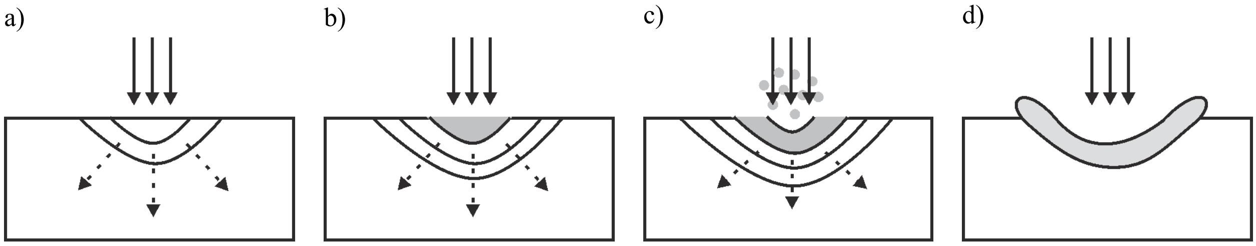


Fig. 2: Phases of interaction of laser and material: a) heating; b) melting; c) vaporization.
d) Schematic representation of melt expulsion process.

After the engraving process a pit with a transition ring around it is left behind at the exposure site. An idealized described cross-section relief is shown in Fig. 3a, the cross-section of real samples can be seen for two examples in Fig. 3b.

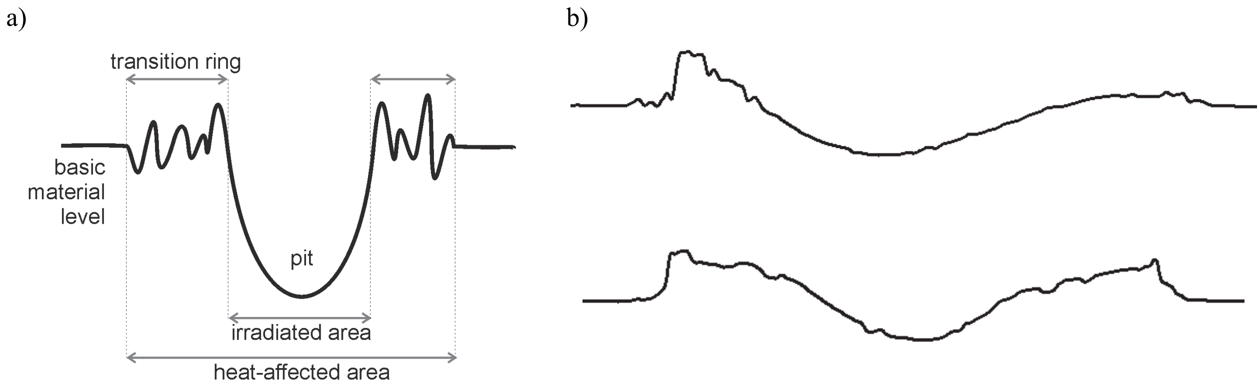


Fig. 3: a) Idealized cross-section relief of a sample and its description; b) cross-section reliefs of two different real samples.

Methods for the Automatic Heat-Affected Area Detection

The main task of the heat-affected area detection is to define the area of the material surface, which was affected by the laser beam during the engraving process, as exactly as possible. The required algorithm should detect the area automatically and accurately so that it can be used as a part of the data pre-processing.

In the idealized case, the detection of the heat-affected area would be simple, but for the real samples many problems appears. These problems (such as local defects, material surface roughness and other inaccuracies) can be seen Fig. 4. In Fig. 4a, the sample burned into the smooth surface of steel is shown in the 3D view. The detection of the heat-affected area in this case is even in despite of several local defects less problematical than in the case of the sample burned into cermet (Fig. 4b), where the roughness of the basic material complicates the detection a lot.

The most exact and reliable way of the heat-affected area detection is the manual way. But if we want to use it as a part of the whole data pre-processing, the manual usage slows the computation down and prevents its automation. The detection should be used several times during the process of the laser engraving modelling, e.g. during the parameterization of the sample or if we want to compare two samples. Hence, in order to speed up and simplify the whole pre-processing, the system has to be maximal self-contained. But the precision and accuracy should be preserved in a maximal possible way. While the user is able to distinguish roughness or defects on the material during the manual pulse detection well, for an automatic method it is difficult to differentiate inaccuracies on the sample surface from the outer border of the pulse.

a) b)

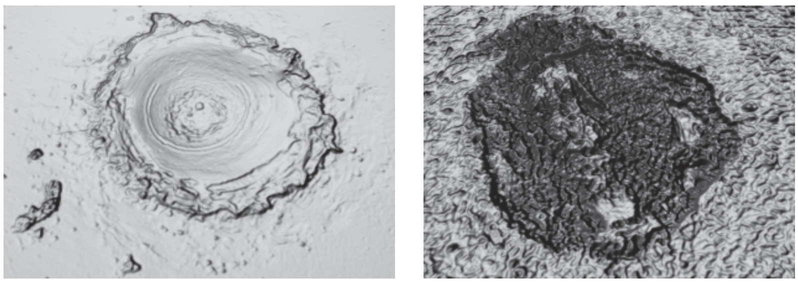


Fig. 4: a) Surface of steel with relatively smooth surface, but several local defects on the surface; b) cermet sample with globally higher roughness of the material surface.

We have designed several methods, which are described in detail especially in [2.] and [3.]. In the following explanation each algorithm will be shortly outlined and the main focus will be given on their comparison (in Section 0).

Clipping Method

This detection algorithm goes from unmodified sample surface. At the beginning we need a starting point, which should be determined as the point placed preferably in the middle of the heat-affected area. The starting point for automatic pulse detection is computed as the position of centre of mass in the sample. The standard procedure of the centre of mass computation has to be adapted for the sample representation and so equation (1) for the sample centre of mass computation has been derived. Value $f_i = f[x, y]$ represents the single point of the sample and *materialLevel* represents the height of the basic material. This method returns the starting position which is placed in the area of engraved pulse and is affected by local material defects in a minimal way.

$$x_c = \frac{\sum m_i . x_i}{\sum m_i}, \text{ where } m_i = (f_i - materialLevel)^4 \qquad (1)$$

From the starting position, columns of the height map to the left and to the right side are inspected and the height difference of points in each single column (it means the difference between the maximal and minimal value in the column) is computed. If the value does not exceed given height limit, an inspection in the direction is finished and columns behind these left and right borders are clipped. After clipping the columns on the left and right side of the heat affected area, the same process of border searching is started for rows. Horizontal borders are appointed, the other rows are clipped. The procedure is in a simplified way shown in Fig. 5a.

There is a question, how to gain the value of a difference height limit. If we do not want to set the constant manually, because we explore the sample automatically, we can compute the value as the minimal height difference of several border columns or rows. This approach works for the majority of samples, because the borders of the sample are not modified during the laser engraving and so, they represent the original material surface.

Spiral Method

All particular steps of spiral method are shown in Fig. 5b. Also this method goes also from unmodified sample surface and uses the starting point that is computed as the centre of mass of the sample. If we put through the starting point two lines parallel with axis *x* and *y* (in Fig. 5b see two solid lines), we can get two cross-sections of the sample along these lines similar to those shown in Fig. 3b – the vertical and the horizontal one. On each curve we can determine two points, where the heat-affected area finishes and to define borders there. By this approach we get left and right border of the pulse from the horizontal cross-section curve and the top and bottom border from the vertical cross-section curve.

From this starting bordering rectangle (in Fig. 5b highlighted with dotted rectangle) we have to continue the final border searching. The final borders are searched in a spiral way. All borders are periodically tested if it is possible to move them for one row or one column further from the starting point. In each step for each single border (left, bottom, right and top), the height difference between the minimal and the maximal value in the shifted position is computed and compared with the difference limit (computed in the same way as in the case of the clipping method) for the processed sample. The sequence of borders is preserved through the whole computation (it means borders are rotating during the algorithm). If any border can not be moved further, it is skipped in the next rotation. The final bordering rectangle is in Fig. 5b marked with the dashed line.

a) b)

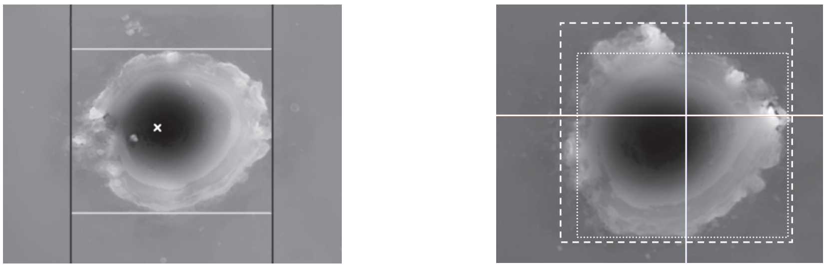


Fig. 5: a) Borders determined during the clipping method (white cross represents the starting point). b) Spiral method detection process – cross-section positions (solid lines), starting rectangle (dotted rectangle), final borders (dashed rectangle).

Statistical Method

Statistical method works on a completely different principle than methods described in Sections 0 and 0. The whole sample height map is divided into the regular rectangular grid and for each cell of the grid, a representing value (as a difference of the minimal and maximal height in the cell) is computed. Part of such sample with computed values can be seen in Fig. 6a. In this way we reach simplification of the sample and for its further processing we use the statistical approach. To get optimal results, it is very important to determine size of the statistical grid cell, because we need to simplify it enough, but not too much. Our experiments show that the grid size depends especially on the used material and so, the optimal value is determined experimentally for each tested data set.

To distinguish the values of the heat-affected area from the rest of the sample, we can use thresholding. After doing this, cells representing the basic material are labelled with zero value; the others are indicated by the value of 1 and further processed. In Fig. 6b such cells are highlighted (the dashed rectangle borders the area zoomed in Fig. 6a). To remove the small areas of the local defect, we need to find the largest labelled area and so we use the binary image segmentation, more concretely the connected component labelling, described in [4.] or [5.]. Finally, we get the largest area which represents the heat-affected area (see Fig. 6c).

a) b) c)

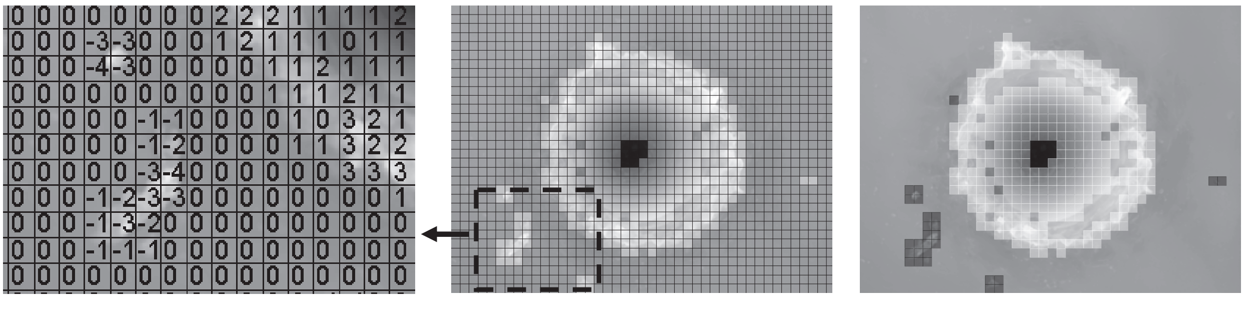


Fig. 6: a) Part of the sample with computed values for the statistical grid (bottom left corner of the sample); b) the grid after the thresholding; c) the largest thresholded area determined.

Methods Results and Comparison

All three described methods were tested on the set of real samples. To test all methods properly, we have used samples engraved in two different materials – steel and cermet. For each material several samples with various defects and other problematic parts were chosen. Methods are compared from two points of view – their universality and accuracy. Speed of all methods is comparable, because it depends especially on the sample dimension (on an average 300ms is needed for one detection; Intel Core 2Duo CPU 3GHz, 3.25GB RAM, Windows 7, Java 1.6).

Accuracy

The accuracy means, how precisely the method is able to set bounds to the heat-affected area. In Fig. 7 and Fig. 8, several testing samples and results of the heat-affected area detection are shown. In all figures, results of the clipping method are bordered with dashed line, for results of the spiral method dash-dotted rectangle was used and the statistical method is visualized by dotted line. The numbers of laser pulses engraved into the material are described directly in the figure labels.

For samples burned into cermet, more often the clipping method was successful. The reason is the higher roughness of the material surface which complicates thresholding phase in the statistical methods. On the other hand, the difference between the original surface and the heat-affected are is higher and so, the clipping method is able to distinguish them well.

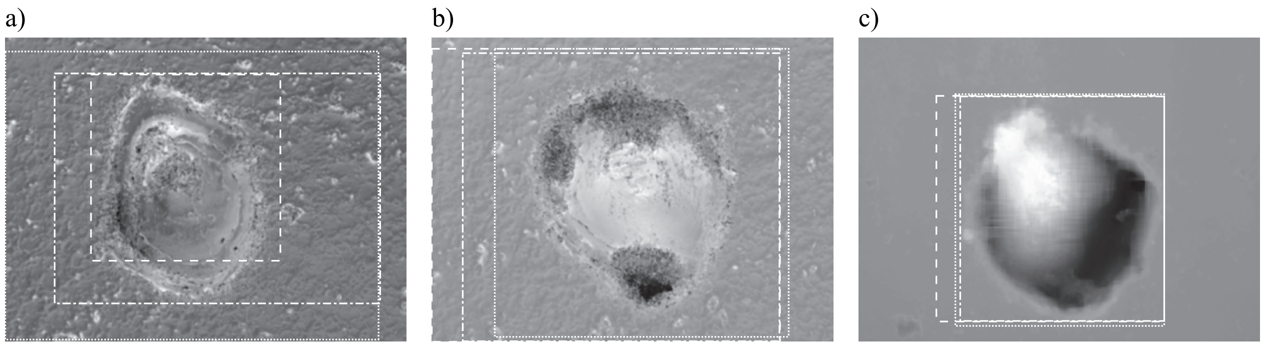


Fig. 7: Samples engraved into cermet with various numbers of laser pulses: a) 2; b) 50; c) 90.

As can be seen especially in Fig. 8a-b, for steel the detection is complicated especially for the small number of laser pulses, because there is only slight difference between the changes of the material caused by the engraving process and the irregularities of the original surface.

Majority of samples were detected well with the statistical method, while the others are more influenced by various defects and surface irregularities. Its main disadvantage lies in the manual experimental setting of the statistical grid cell size. Fortunately, for all samples engraved into the same material it has to be set only once, because this value depends especially on the used material. Clipping and spiral methods work fully automatically.

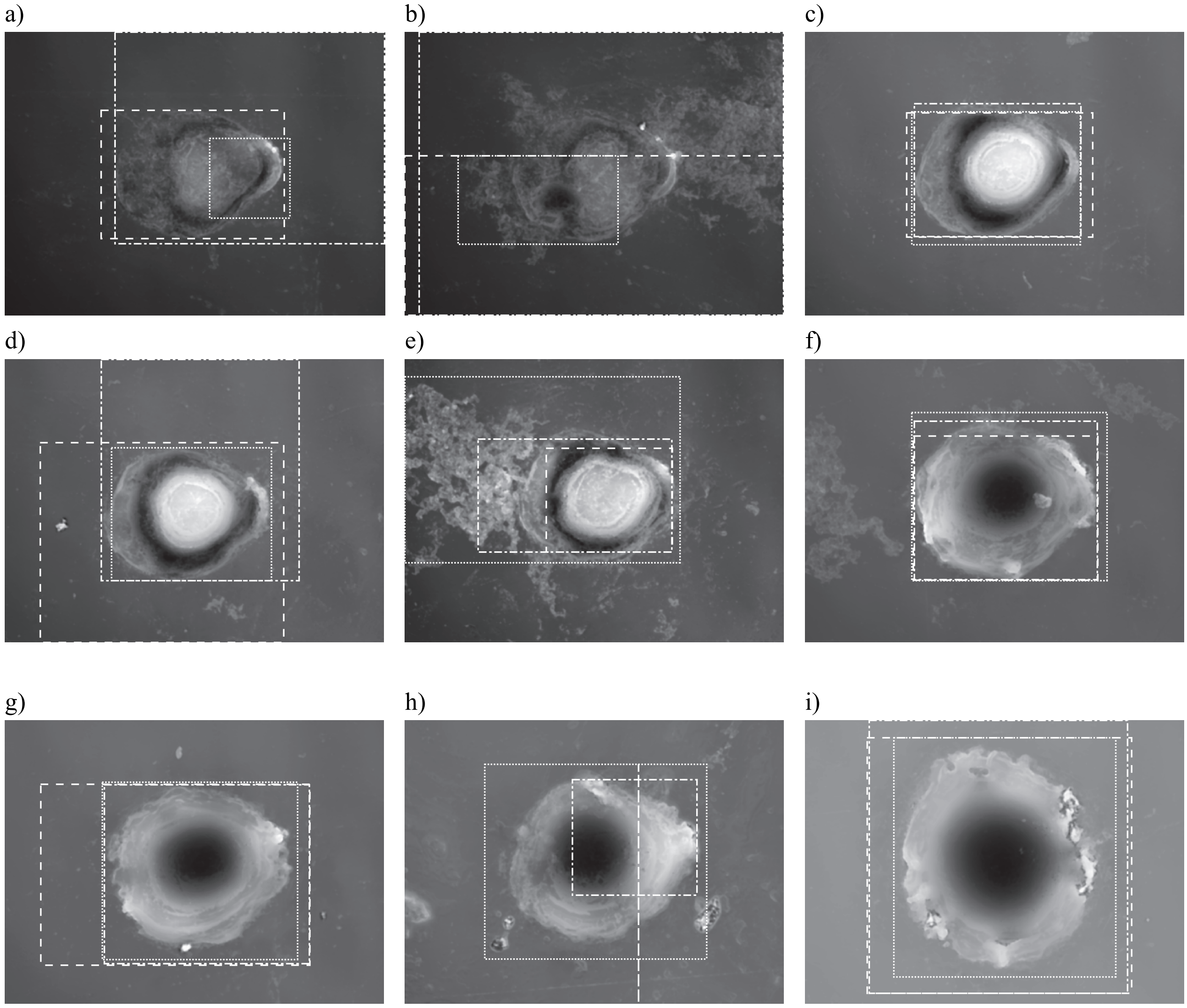


Fig. 8: Samples engraved into steel with different numbers of laser pulses: a-b) 1; c-e) 2; g) 10; h) 20; h) 40; i) 90.

Universality of Usage

The testing of methods universality was also provided for both materials. For each tested sample all three methods were used and gained results were analyzed. If the heat-affected area was detected well (as e.g. in Fig. 8c), the analyzed method gained a point. If the result of automatic detection was tolerable (e.g. spiral and clipping methods in Fig. 8g or clipping method in Fig. 8i), but not optimal, only half of a point was counted. For all similar samples (i.e. samples engraved into the same material with the same laser pulse number), the results were averaged. Results for cermet can be seen in Fig. 9a, results for steel in Fig. 9b.

Unfortunately our data set for cermet is not as large as in the case of steel and so for cermet only one experiment for each number of pulses engraved into the sample surface was tested. Nevertheless it seems, for cermet the clipping method (and partly also the spiral one) give better results than the statistical detection. On the other hand, for samples engraved into steel (where for each number of laser pulses five similar samples were analyzed), the best results were given by the statistical method, especially for samples with 10 and more laser pulses engraved into one point.

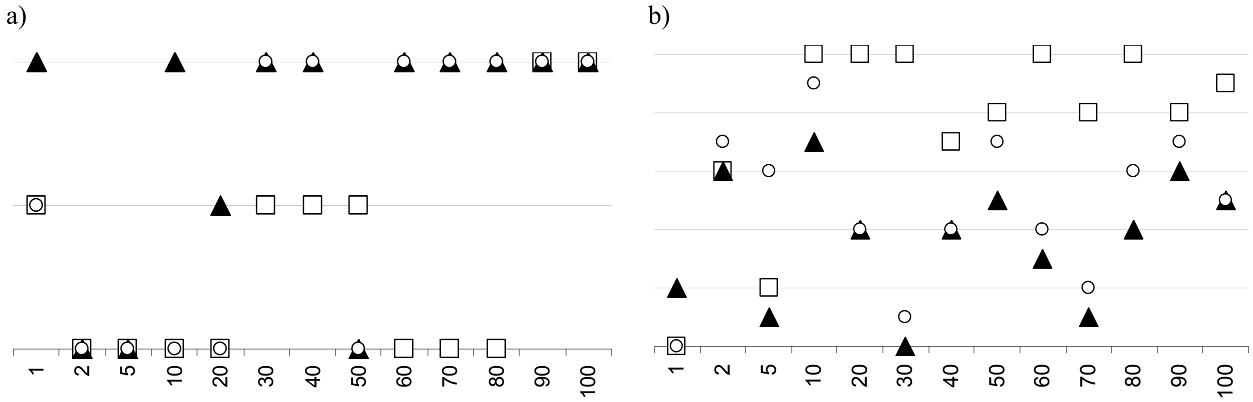


Fig. 9: Comparison of the successful heat-affected area detection for all three methods and samples engraved into: a) cermet; b) steel (☐ statistical, ▲clipping, o spiral method).

Conclusion

Our results show, that for different material (especially if they differ in the basic material surface roughness), different approaches are applicable. Another aspect influencing the usable method is the high difference between the basic material level and extremes of the heat-affected area. Tested methods give promising results, hence we will continue in their development and also testing on various materials and samples, so that they could be integrated into the real data pre-processing as the foolproof, exact and self-contained processing step.

References

[1.] DAHOTRE, N. B., HARIMKAR, S. P.: *Laser Fabrication and Machining of Materials*, Springer, New York, USA, 2008.

[2.] HÁJKOVÁ, J.: *LASER SIMULATION - Methods of Pulse Detection in Laser Simulation*, Proceedings of the 3rd International Conference on Software and Data Technologies ICSOFT 2008, INSTICC, Porto, Portugal, 2008, pp. 186-191.

[3.] HÁJKOVÁ, J.: *Data Processing for Simulation of Laser Beam Impact—Statistical Method for the Heat-Affected Area Detection*, Proceedings of the First International Conference on Computational Intelligence, Modelling, and Simulation CSSim 2009, Brno, Czech Republic, 2009, pp 69 – 74.

[4.] HARALICK, R. M., SHAPIRO G. L.: *Computer and Robot Vision, Volume I*. Addison-Wesley, Longman Publishing Co., Inc. Boston, MA, USA. 1992.

[5.] KOTÁSEK, J.: *Metody porovnávání výškových map*, Bachelor Thesis, University of West Bohemia, 2010.

[6.] SHAPIRO, L., Stockman, G.: *Computer Vision*. Prentice Hall, New Jersey, USA. 2001.

[7.] STEEN, W. M.: *Laser Material Processing*. Springer-Verlag, New York Berlin Heidelberg, 1991.

Current address

Ing. Jana Hájková, Ph.D.
University of West Bohemia in Pilsen,
Faculty of Applied Sciences,
Department of Computer Science and Engineering,
Univerzitní 22, 306 14 Pilsen, tel.: +420377632434,
e-mail: hajkovaj@kiv.zcu.cz

Journal of Applied Mathematics

volume IV (2011), number I

MATHEMATICAL MODELL USED IN DECISION-MAKING PROCESS WITH RESPECT TO THE RELIABILITY OF THE DATABASE

HOŠKOVÁ-MAYEROVÁ Šárka, (CZ), TALHOFER Václav, (CZ), HOFMANN Alois (CZ)

Abstract. Tle first aim of the article is to show how it is possible---thanks to the use of sophisticated analytical tools for evaluation of data quality - to better understand geo-space data. Another aim is to assess the influence of data quality on the results of space analyses that are made of them and that are the basis for such decision-making processes, in which it is necessary to take into account the influence of geographical environment.
Organizations that are engaged in creating geo-space databases usually define the content of these databases (i.e. listing of geographical objects and their features) and quality of the data being saved (e.g. geometric, topological and thematic accuracy, level of standardization etc.). As the area of the land that is described with the use of geodata is usually significantly larger than the capacity and technological possibilities of the responsible organization, it is not possible to keep the defined content and its quality in the entire secured area on the same level. When creating the space analysis it is therefore necessary to take into account the immediate quality level of data in the particular area and to have the technologies for finding out the reliability of the result of the particular analysis available. From the real practice a request of commanders is known, that is to have not only the result of their own analysis available as basics for their qualified decision (decision-making process) but also relevant information about its reliability.

Key words. treliability, decision making process, mathematical modelling, spatial data, GIS, quality assessment, utility value

Mathematics Subject Classification: Primary 00A71, 00A72; Secondary 93AE30.

1. Introduction

Command and control systems used in various branches of rescue systems and in the army as well uses spatial data and information more and more. Spatial data are collected from various sources and using various technologies. It results into position and thematic properties in homogeneity. In spite of this situation data are stored and used in a common spatial database or they are used for various kinds of spatial analyses. Obtained information can be apply in a decision making process and its precision and reliability has an influence over final solution. The goal of this paper is a proposal of a method focused on data and spatial information precision and reliability evaluation and continues the task from [5]. The resulting characteristics of data reliability can be applied in a command and control system. The proposed methods will contribute into increasing of quality of decision making process.

In more details in the Czech Republic rather extensive databases of area-localized data utilized in a number of fields are created. Data model objects and phenomena of both natural and social character (water courses, settlement structure, atmospheric pressure, etc.). The created and utilized data always encompass a position element, which localizes objects and phenomena in a given reference coordinate system, and a thematic element, which describes qualities of the given objects and phenomena (e.g. the speed of a water course, number of inhabitants in a settlement, actual atmospheric pressure readings). The actual data may then be of both geographic and non-geographic character (data on water courses contrasted with data on the transported cargo). The following text therefore uses predominantly the general term “spatial data” or “spatial information”.

Basic localization databases are created by state administration bodies (Czech Office for Surveying, Mapping and Cadastre - COSMC, Army of the Czech Republic – ACR) and are intended for activities related to state functions, including command and control systems implemented in armed forces and in crisis management of individual components of the Integrated Rescue System (IRS). Spatial data are used not only for basic orientation in space but also as data for solving tasks connected with actual decisions, e.g. geographic impact on combat and non-combat army activities in given environments, predictions of landscape damage under extreme meteorological conditions or emergencies, in cases of military threats to the state, etc. In a number of tasks the source data combine and based on mathematically or procedurally described processes, new data are created.

If during decision making process the spatial data and information are used a complex knowledge of the characteristics is essential for evaluating the reliability and accuracy of decisions. Using the method of value analysis and mathematical modeling is possible to create a comprehensive system for evaluation of spatial data usability. Based on the input characteristics of the used spatial data and databases, quality characteristics and their changes can be calculated with the help of analytical methods.

Via comparison of the quality improvement or modification of databases is possible to optimize both the overall usability, and costs incurred on its security.

Through mathematical modelling it is then possible to solve tasks of the following types:

- how a change in a given partial parameter or several parameters of a database is reflected in its total usability;
- which parameters need to be changed to achieve the required product functionality;
- which parameters may be “degraded” owing to the fact that the product’s functionality is unnecessarily high.

Above described procedure was used in the task **“**Finding the most beneficial route“. The main aim of this task was to judge several possible calculated paths for the vehicle Tatra 815, for more details see [4]. These paths were calculated in the „cost map (CM)“ based on the patency parameters for this type of vehicle. In the experiment two versions of cost map were used. Map versions were created based on usable property changes of the database. Particularly, in this case we mean changes in topographic database, soil type database and digital elevation model. The pictures illustrate changes in route selection received after fill in the database.

Fig. 1 Impact of building factor to the route selection

The proposed solution aims to streamline the activities associated with the use of inhomogeneous data and information systems, command and control so that operational components should be available not only its own database, as well as relevant documentation about the quality and reliability of data used. Based on this information they can in their decisions to work with those documents and where appropriate their decisions correct.

2. General assessment of spatial data utility

The product or a part of the product resultant function utility degree may be assessed based on the criteria for the spatial geodatabase utility value evaluation using a suitable aggregation function F [4]:

$$F = p_3 k_3 p_4 k_4 (p_1 k_1 + p_2 k_2 + p_5 k_5) \tag{1}$$

The chosen form of the aggregation function concerns also the case the user gets data on an area beyond his interest or data obsolete so that their use could seriously affect or even disable the *digital geospatial information* (DGI) functions. The weight of each criterion is marked as p_i, where $i = 1,\dots,5$. The mentioned aggregation function proves the product status at the questioned instant and its utility rate. It is applicable also to experiments to find the ways of how to increase product utility at minimum cost increment.

2.1. Individual DGI benefit cost assessment structure

The organisation, such as the Geographic Service of the Army of the Czech Republic or the Czech Office for Surveying, Mapping, and Cadastre, are usually responsible for DGI databases development continuously covering all the Czech Republic area or some parts of the World. Digital Landscape Model (DLM25 or DMU25 in the Czech language), Multinational Geospatial Co-

Production Program (MGCP) or Vector Map Level 1 (VMap1) can be mentioned as examples from military branch.

The DGI are usually developed and maintained by individual partial components of the complete database, such as save units, measurement units, map sheets etc. Therefore, it is quite a good idea to assess their utility value in the above-described system within the established the storing units introducing *individual benefit value*. Similarly the individual benefit value can be applied for the selected part of master databases from given *area of interest* which is used for certain task.
When assessing database utility, it is useful to define *ideal quality level* at first. The ideal level is used as a *comparison standard* to express each criterion compliance level. Using the comparison standard the individual criteria compliance level and consequently aggregate utility may be assessed.
The compliance level of each individual criterion $u_{n,s}$ is given as follows:

$$u_{n,s} = \frac{k_s}{k_s^*} \tag{2}$$

where

- k_s is for the value of s^{th} criterion compliance,
- k_s^* is for the level of compliance of s^{th} criterion or its group criterion of the comparison standard.

Than the aggregate individual benefit value (*individual functionality – U_n*) of the n^{th} save unit is defined by the aggregation function of the some type as (1). Therefore:

$$U_n = p_3 u_{n,3} p_4 u_{n,4} (p_1 u_{n,1} + p_2 u_{n,2} + p_5 u_{n,5}) \tag{3}$$

The individual criteria weights are identical with the weights in database utility value calculation. Particular criteria usually consist of several sub-criteria. The authors took 20 criteria into their consideration; hence the equation for calculation the aggregate individual utility value is therefore a function of 20 variables that characterise the levels of compliance for each individual criterion.

Any modification of selected criterion has an impact on the value of U_n. Individual variables are independent one to another, so the derivation of the function can model the changed utility values or individual utility values.

$$dU = \frac{dU_n}{du_{n,i}} \tag{4}$$

where $i = 1,\ldots, 5$, $n = 1,\ldots, N$, and N is number of all saved units in the database.
Determination of dU value is thus feasible in two ways regarding the desired information structure. When assessing *individual variables effects* on the individual functionality value, while the other variables keep constant values, it is necessary to differentiate U function as follows:

$$dU = \frac{dU_n}{du_{n,i}} \frac{du_{n,i}}{dx} \tag{5}$$

where x is one of the 20 mentioned variables.

In practice, however, such situations may arise that multiple factors may change at the sometime, e.g. the technical quality of database changes in all its parameters—the secondary data derivation methods will improve location and attribute accuracy and the data integrity will increase, and moreover the data are stored in a geodatabase accessible to all authorised users. In this database the data are maintained properly with respect to all topologic, thematic and time relations. In such a case it is suitable to define *dU* value as a total differential of all variables describing the modified factors.

Database functionality degree is comparable to the cost necessary for provisions—direct used material, wages, other expenses (HW, SW, amortisation, costs for co-operations, tax and social payments etc.), research and development cost, overhead cost and others. Functionality and cost imply *relative cost efficiency (RCE)* calculated as follows:

$$RCE = \frac{F}{\sum_{i=1}^{n} E_i} \tag{6}$$

where $i = 1,\ldots, N$.
Similarly to individual utility value, it is possible to consider the impact of particular variables of expenses E_i on final *RCE*. The goal is to find such solution as the functionality will be maximised and the expenses will be minimize.

The DGI benefit cost assessment including individual benefit cost is a task for a data manager or a geographer-analyst which is responsible to provide demanding project. The system enables him to consider which quality parameters are possible to improve in given time, with given technological conditions, with given sources, with given co-workers etc.

3. Pilot study

In order to verify the VAT methodology the task of Cross Country Movement (CCM) was chosen as an example. CCM can be solved as a common problem or with consideration of certain types of vehicles (the most frequent or the weakest in the unit, but in case of armed forces usually off road vehicles). The detailed theory of CCM is in [2].

The solution can offer to the commander not only one possibility, but the variants from which he can choose according to his intentions and the current situation at the given area.

3.1 Spatial database utility value evaluation

The master DGI database is usually utilised as a base for spatial data analyses. The national or international databases as DLM25, VMAP1or MGCP are very detailed, carefully maintained and used in many applications. But nobody can suppose that those databases contain all information he could need.
The task of CCM solution could require more information that is available in the master database. Geographer-analyst has to consider which information and in what quality can he obtain from master database. Further he has to find out all their properties and their accuracy or count how many characteristics are missing. Next step is the individual functionality value of given part of master database evaluation.

Attributes are usually defined as the characteristics or nature of objects. In geospatial sense, an attribute is regarded as a property inherent in a spatial entity [8]. In our case an attribute is a characteristics or variable constituting the base for the computation of basic coefficients. These attributes differ in their nature according to the real world phenomena they represent.

Not all attributes are available within the used thematic spatial databases. So far the incompleteness of attributes has been omitted. Thus the real state-of-the-art has not been taken into account and the resulting CCM path has been considered as 'certain'. One of the possibilities to make the resulting path closer to reality is to take the data attribute incompleteness into account and inform the decision maker (commander) about the uncertain parts of the path.

Two variants of the *Digital Landscape Model 25* (DLM25) database were utilised for the pilot project. The feature properties were defined according to the *Feature Attribute Coding Catalogue* (FACC) adapted as *Catalogue of the Topographic Objects* (CTO) (MTI, 2005) in the first variant updated in 2005. The missing values of object's attributes were marked as 0 in given domains. The 4th edition of CTO was transformed in accordance with the *DGIWG Feature Data Dictionary* (DGIWG-500, 2010) in 2010 and transformed edition (updated in 2010) was used in the second variants (MoD-GeoS, 2010). The missing properties were marked in several attribute categories as:

- -32767 for unknown variables,
- -32766 for unpopulated variables,
- -32765 for not applicable variables,
- -32764 for other variables, and
- -32768 for No or Null values.

The smaller personal database was created in the area of interest round Brno of the size approximately 400 km^2 and all objects and phenomena necessary for CCM evaluating was selected from DLM25 master databases of both variants. The individual utility value was counted for both variants, but with a small simplification. At the first step we didn't do any independent tests for position accuracy determination, further we didn't consider the software independency, and landscape importance. Then we suppose the whole database is complete, the position and thematical resolution corresponds to our task, and the data are properly protected.

On the base of statistical analyse 12.65% objects have any problems mainly incomplete attributes in the first variant of DML25 while 3.45% objects have any similar problems in the second one. The time difference is 5 years between both variants. Hence the individual utility value was calculated by the use of the formula **Error! Reference source not found.** as 0.6887 for the 2005 variant and 0.8825 for the 2010 variant. The ideal quality level is 1.0068. Both variants were used for CCM of TATRA 815 evaluation.

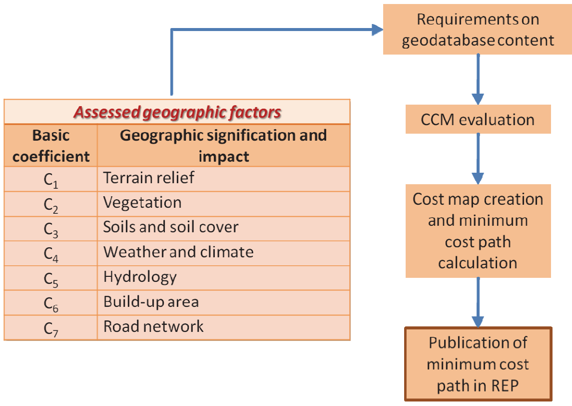
3.2 CCM of TATRA 815 analyses

For further simplification only data with full information (all attribute properties had to be filled) were considered in both variants. If some information was missing in any geographic object system didn't consider it and one reliable path was analyzed. Authors knew that e.g. narrow streams are probably passable for TATRA lorry, but their passing was allowed only over bridges if no information about depth and banks characteristics were in the database.

The process was in progress according to next schema (**Fig. 2**):

- CCM evaluation - only reliable information are considered,
- final cost map calculation,

- minimum cost path calculation from one initial point to three destinations placed in the forests as some hidden position.



Assessed geographic factors	
Basic coefficient	**Geographic signification and impact**
C_1	Terrain relief
C_2	Vegetation
C_3	Soils and soil cover
C_4	Weather and climate
C_5	Hydrology
C_6	Build-up area
C_7	Road network

Fig. 2 Spatial analyse without database quality evaluation

Only one solution is offered to commander. This solution seems to be appropriate, but geographer-analyst generally doesn't know details about situation on area of responsibility and commander's intentions. Problems should appear when the tactical (or other) situation doesn't make the published path possible to use. Commander usually requires a new solution in such a case to miss prohibited area. The quality characteristics of temporally database are than to be considered by geographer to be sure, where are the weakest points of a new analysis. The weak points of analysis have to be sent to commander together with own analysis and it is up to him what will be the final decision. Two tasks for commander appear in CCM example:

1. Use less reliable path and consider that vehicles could stay in front of some obstacle
2. Wait and order to GEO team to improve spatial database (e.g. required properties) as soon as possible and then use new reliable path

The second case was simulated by the second variant of database in which the quality parameters were improved.
ArcGIS 9.3 was used for all calculations and analyses. The main analyses were described using ModelBuilder to have simpler possibility to change the input parameters. In the next figures there are the main results – cost maps. The cost of each pixel is symbolized in the gray scale where darker tone signifies higher cost, higher speed in this case.

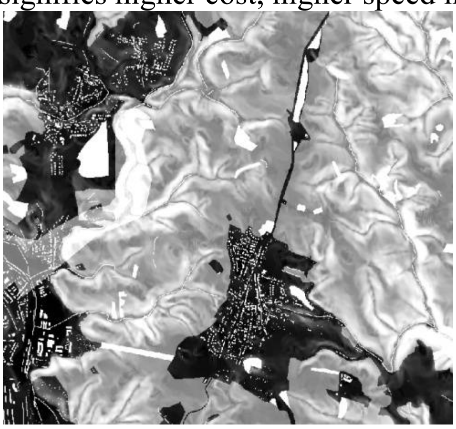


Fig. 3 The cut cost map created from DLM25 2005 version

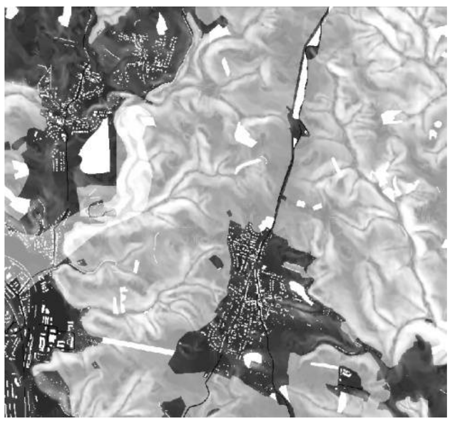


Fig. 4 The cut cost map created from DLM25 2010 version

The minimum cost paths were evaluated using both cost maps and the same process created in ModelBuilder were applied. The results are in the next figures.
The comparing of both results presented over the topographic situation is shown in the next picture (Fig. 5).

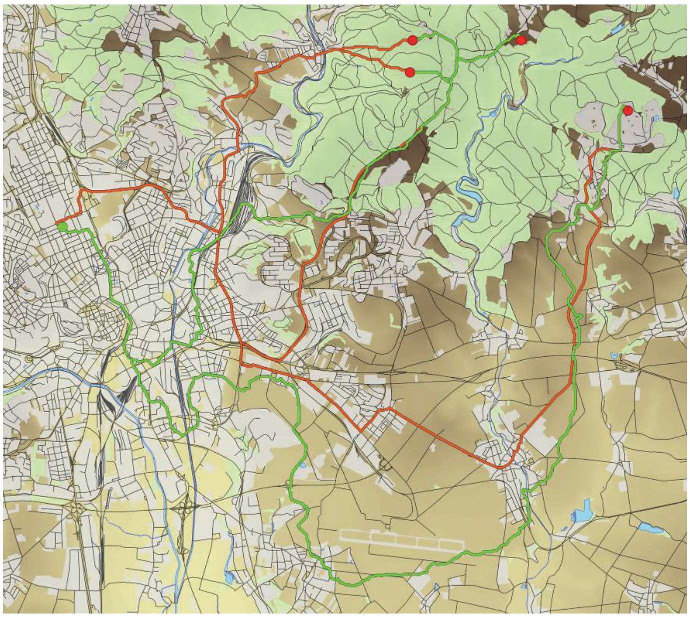


Fig. 5 Comparing of two variant of minimum cost paths. Red ones answer to 2005 version and the green ones to 2010 version.

Acknowledgement

Research results presented above were kindly supported by the project 'The evaluation of integrated digital spatial data reliability' founded by the Czech Science Foundation (Project code 205/09/1198) and also by the Project of Defence Research of the Ministry of Defence of the Czech Republic 'Geographic and meteorological factors on the theatre' (project code METEOR).

References

[1] MILES, L. D. (1989). *Techniques of Value Analysis Engeneering* (3rd. vyd.). USA: Eleanor Miles Walker,

[2] RYBANSKÝ, M. (2009). *Cross-Country Movement, The Imapct And Evaluation Of Geographic Factors* (First. vyd.). Brno, Czech Republic: Akademické nakladatelství CERM, s.r.o. Brno.

[3] TALHOFER, V. (2004). Digital Geographic Data: Potential Evaluation. *AGILE 2004, 7th Conference on Geographic Information Science, Conference proceedings*, 675-686, Heraclion, Crete, Greece: AGILE.

[4] TALHOFER, V., HOŠKOVÁ, Š., HOFMANN A., and KRATOCHVÍL, V. (2010) *Finding the most beneficial route,* MTISD, Italy.

[5] TALHOFER, V., HOŠKOVÁ, Š., HOFMANN A. (2010) *Mathematical model for spatial data properties assessment,* In.: 9th International Conference Aplimat 2010, Bratislava.

[6] TALHOFER, V., HOŠKOVÁ, Š., HOFMANN, A., & KRATOCHVIL, V. (2009). The system of the evaluation of integrated digital spatial data realibility. *6th Conference on Mathematics and Physics at Technical Universities* (pp. 281-288). Brno: University of Defence.

[7] SHI, W. (2010). *Principles of Moddeling Uncertainties on Spatial Data and Spatial Analysis.* Boca Raton: CRC Prass, Taylor and Francis Group.

[8] WILTSCHKO, T., & KAUFMANN, T. (2005). *Euroroads.* (U. Sandgren, Editor) Received 2009, from http://www.euroroads.org/php/Reports/D2.3.pdf

Current address

Šárka Hošková, Assoc. Prof., RNDr. Ph.D.
University of Defence, Faculty of Military Technology, Department of Mathematics and Physics, Kounicova 65, 662 10 Brno, Czech Republic,
sarka.hoskova@unob.cz

Václav Talhofer, COL, Assoc. Prof., Ing., CSc.
University of Defense, Faculty of Military Technology, Department of Military Geography and Meteorology, Kounicova 65, 61200 Brno, Czech Republic
vaclav.talhofer@unob.cz

Alois Hofmann, Ing., CSc.
University of Defense, Faculty of Military Technology, Department of Military Geography and Meteorology, Kounicova 65, 61200 Brno, Czech Republic
alois.hofmann@unob.cz

Journal of Applied Mathematics

volume IV (2011), number I

SOLUTION OF REAL BEAM ON ELASTIC FOUNDATION USING PROGRAM *MATHEMATICA*

JANČO Roland, (SK), KOVÁČOVÁ Monika, (SK)

Abstract. In real design of beam you can consider of properties of background. No all time you can neglecting of properties of background and consider rigid background. In this paper is short introduction to theory of solution of beam on elastic foundation. Result from this solution is differential equation of deflection curve. For solution this differential equation we used program *Mathematica*.

Key words and phrases. Beam, elastic foundation.

Mathematics Subject Classification. Primary 74A10, 74B05, 74M15 ; Secondary 74K10.

1 Introduction

Solution of frames and beams on elastic foundation are often occur in many practical case for example, solution of building frames and constructions, buried gas pipeline systems and in design of railway tracks for railway transport, etc.

Solution of beam on elastic foundation is statical indeterminate problem of mechanics. In this case we have beam with elastic foundation along whole of length and width or only some part of length or width. From theoretical solution of this type of problems exist only for some type of loads and beam of infinity, semi-infinity and finite length. Detailed explanation of theoretical solution can be find in [1], [2] and [3].

2 Theoretical background

Les us consider a prismatical beam with lenght L, which is supported along its length by a continuous elastic foundation. It is used Winkler elastic foundation. Such that when the beam is deflected, the intensity of continuously distributed reaction at every section is proportional to the deflection an that section. Under such conditions the reaction per unit length of the bar can be represent by the expression $k\,w$, in which w is the deflection and k is a constant usually called by [2] the *modulus of elastic foundation*. The exact solution of frames on elastic foundation we have to solve no only frames but also foundation, which is pretty complicated problem. Complicated is physical description of this problems, because complicate is description of elastic foundation, which have influence of lot of factor of foundation.

In studying the deflection curved of the beam we use the differential equation

$$EI\,\frac{d^4\,w}{d\,x^4} + q_R = q_0\,, \tag{1}$$

where EI is bending stiffness, q_R is denote the reaction of intensity of the load acting on the beam and q is the uniform load, which is the intensity of weight of beam. Hence

$$q_R = k\,w\,. \tag{2}$$

and equation (1) becomes

$$EI\,\frac{d^4\,w}{d\,x^4} + k\,w = q_0\,, \tag{3}$$

Let us consider of infinite length of beam with the uniform load q_0, which represent of weight of beam, and uniform load q_0, which represent of weight of beam and earth overlay, in the Fig. 1.

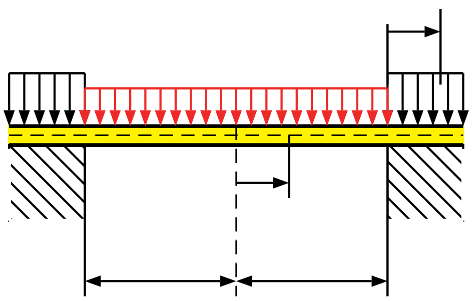


Figure 1: Real beam on elastic foundation.

For solution of problem in Fig. 1 we used method of superposition, which our problem divide to solution of two problems in Fig. 2.

Because we have axis of symmetry, we can solve only one half of beam, for example for Solution A in the Fig. 3.

Solution of differential equation (3) is in the form

$$w(x) = e^{-\beta x}(C_1 \cos\beta x + C_2 \sin\beta x) + e^{\beta x}(C_3 \cos\beta x + C_4 \sin\beta x) + w^*\,, \tag{4}$$

where C_1, C_2, C_3 and C_4 are integration constant, w^* is particular solution of differential equation and

$$\beta = \sqrt[4]{\frac{k}{4EI}}\,. \tag{5}$$

Integration constant for beam in Fig. 3 is finding for follow boundary condition

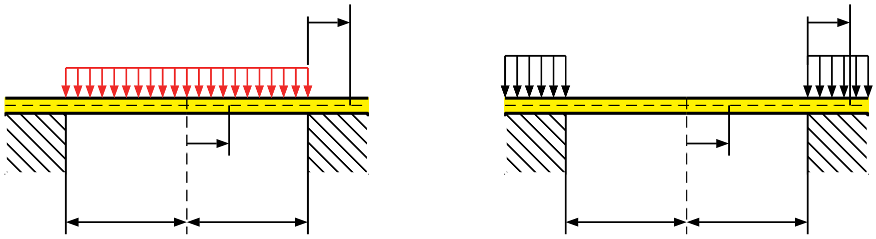


Figure 2: Method of superposition

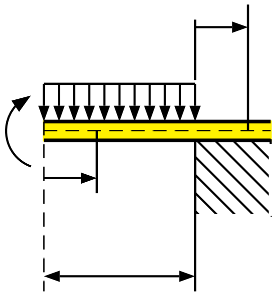


Figure 3: Model of beam on elastic foundation for Solution A

1. $x_1 = 0 \qquad w_1' = 0,$
2. $x_1 = 0 \qquad w_1''' = 0,$
3. $x_2 = \infty \qquad w_2 = 0,$
4. $x_2 = \infty \qquad w_2' = 0,$
5. $x_1 = \mathrm{L} \qquad x_2 = 0 \qquad w_1''' = w_2''',$
6. $x_1 = \mathrm{L} \qquad x_2 = 0 \qquad w_1'' = w_2'',$
7. $x_1 = \mathrm{L} \qquad x_2 = 0 \qquad w_1' = w_2',$
8. $x_1 = \mathrm{L} \qquad x_2 = 0 \qquad w_1 = w_2$

and for beam in Fig. 2b) (x_1 is start in middle of beam) are follow

1. $x_1 = 0 \qquad w_1' = 0,$
2. $x_1 = 0 \qquad w_1''' = 0,$
3. $x_2 = \infty \qquad w_2 = \dfrac{\mathrm{q}}{k},$
4. $x_2 = \infty \qquad w_2' = 0,$

5. $x_1 = \mathrm{L}$ $\quad x_2 = 0$ $\quad w_1''' = w_2'''$,

6. $x_1 = \mathrm{L}$ $\quad x_2 = 0$ $\quad w_1'' = w_2''$,

7. $x_1 = \mathrm{L}$ $\quad x_2 = 0$ $\quad w_1' = w_2'$,

8. $x_1 = \mathrm{L}$ $\quad x_2 = 0$ $\quad w_1 = w_2$.

3 Solution of real beam on elastic foundation using program *Mathematica*

For the beam on elastic foundation in the Fig. 1, let us consider following parameters: Young's modulus of beam $\mathrm{E} = 2.1\,10^5\,\mathrm{MPa}$, Poisson ratio $\nu = 0.3$, density $\rho = 7850\,\mathrm{kg/m^3}$, length of beam without foudation $L = 4\,\mathrm{m}$, outside diameter of pipe $\mathrm{D} = 1204\,\mathrm{mm}$, thickness of pipe $\mathrm{h} = 0.0135\,\mathrm{m}$, uniform load simulated of weight of pipe with gas $\mathrm{q}_o = 3888,228\,\mathrm{N/m}$, uniform load simulated weight of earth overlay $\mathrm{q}_{over} = 19800\,\mathrm{N/m}$ and modulus of elastic foundation $k = 7,9.10^6\,\mathrm{Nm^{-2}}$.

Differential equation described the first part without the elastic foundation $x_1 \in \langle 0, \mathrm{L}\rangle$ in fig.1 is following

$$w_1'''' = \mathrm{Q}_0\,, \tag{6}$$

where $\mathrm{Q}_0 = \dfrac{\mathrm{q}_0}{\mathrm{EI}}$ and I is moment of inertia of cross-section area, which is for the pipe defined by following equation

$$\mathrm{I} = \frac{\pi\mathrm{D}^4}{64} - \frac{\pi(\mathrm{D}-2\mathrm{h})^4}{64} = \frac{\pi\mathrm{D}^4}{64}\left[1 - \left(\frac{\mathrm{D}-2\mathrm{h}}{\mathrm{D}}\right)^4\right]. \tag{7}$$

Differential equation described the second part with the elastic foundation $x_2 \in \langle 0, \infty\rangle$ in fig.1 is following

$$w_2'''' + 4\beta^4\, w_2 = \mathrm{Q}\,, \tag{8}$$

where $\mathrm{Q} = \dfrac{\mathrm{q}}{\mathrm{EI}} = \dfrac{\mathrm{q}_0 + \mathrm{q}_{over}}{\mathrm{EI}}$ and β is given by equation (5).

Bondary conditions without using method of superposition are following

1. $x_1 = 0$ $\quad w_1' = 0$,

2. $x_1 = 0$ $\quad w_1''' = 0$,

3. $x_2 = \infty$ $\quad w_2 = \dfrac{\mathrm{q}}{k}$,

4. $x_2 = \infty$ $\quad w_2' = 0$,

5. $x_1 = \mathrm{L}$ $\quad x_2 = 0$ $\quad w_1''' = w_2'''$,

6. $x_1 = \mathrm{L}$ $\quad x_2 = 0$ $\quad w_1'' = w_2''$,

7. $x_1 = \mathrm{L} \qquad x_2 = 0 \qquad\qquad w_1' = w_2'$,

8. $x_1 = \mathrm{L} \qquad x_2 = 0 \qquad\qquad w_1 = w_2$.

Advantage of program *Mathematica* is follow, user have not know the shape of solution differential equation (6) or (8), the program calculate the shape of result using boundary condition by command DSolve, for example in Fig. 1 and given parameters and boundary condition given above. Syntax in program *Mathematica* is in the fig.4.

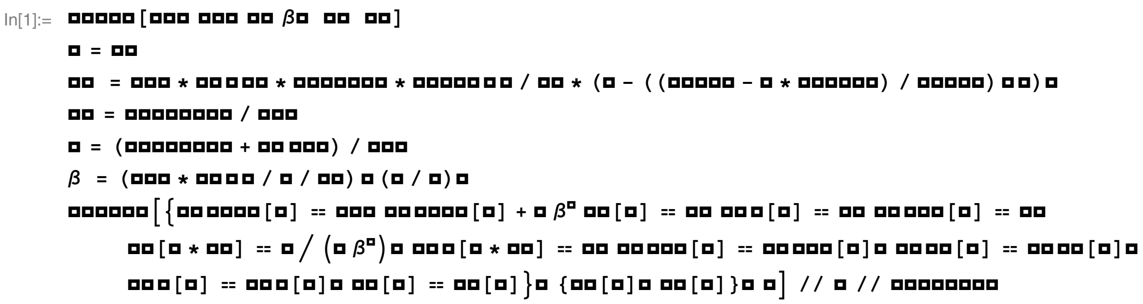


Figure 4: Input in program *Mathematica*

Result from program *Mathematica* is in the Fig.5.

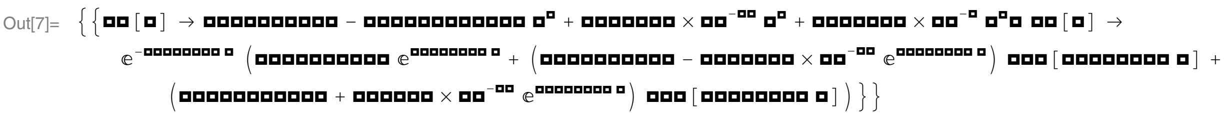


Figure 5: Output in program *Mathematica*

Maximum deflection of beam is at point $x_1 = 0$ with the value 0.00375301 m. At point $x_1 = 0$ and $x_2 = \mathrm{L}$ is deflection equal 0.0035727 m. Minimum value of deflection is 0.00299851 m.

Graphically the deflection in range $\mathrm{x}_1 \in \langle 0, \mathrm{L}\rangle$ is the Fig. 6. Graphically the deflection in second part of beam in range $\mathrm{x}_2 \in \langle 0, 10\,\mathrm{L}\rangle$ is the Fig. 7.

4 Conclusion

Solution of beam on the elastic foundation is statically indeterminate problem of mechanics. In this paper is solution of infinity beam load by external uniform load, which is equal to weight of beam with gas or weight of beam with gas and earth overlay. In real case is loading by more external load (bending moment, etc.) and elastic foundation can be nonhomogeneous. Solution of beam on elastic foundation lead to 4th order differential equation, where to finding a solution can significantly help software system *Mathematica*. The results can be applied to design of beam or parameters of elastic foundation.

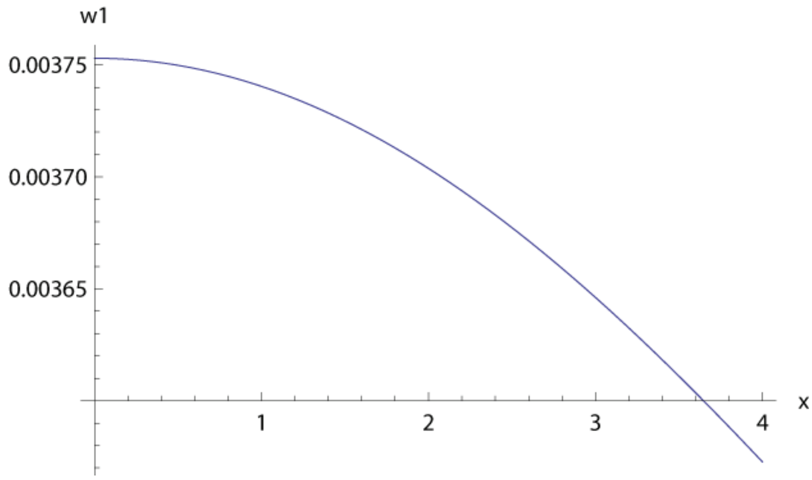


Figure 6: Graphically representation of deflection function in the range $x_1 \in \langle 0,\, \mathrm{L} \rangle$

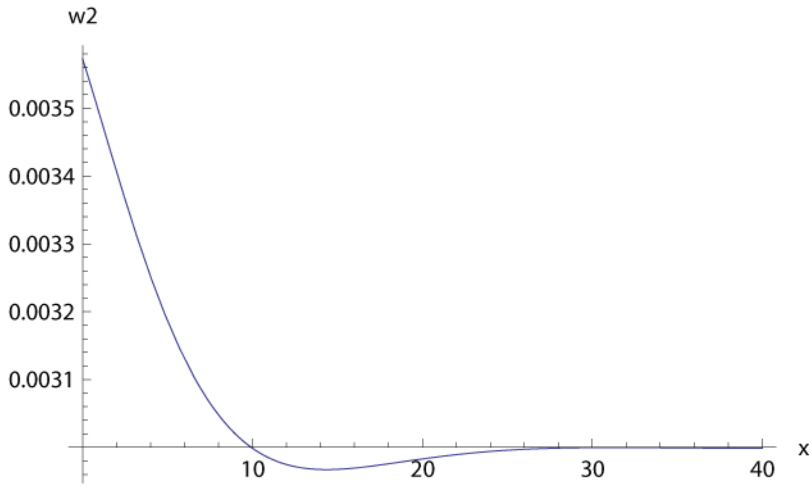


Figure 7: Graphically representation of deflection function in the range $x_2 \in \langle 0,\, 10\,\mathrm{L} \rangle$

Exact solution of problem described in chapter 4 [3], which is following: For the first part of pipeline in the range $x_1 \in \langle\, 0,\, \mathrm{L} \rangle$

$$w_1(x_1) \;=\; \frac{\mathrm{q}_0\, x_1^4}{24\mathrm{EI}} \,+\, \mathrm{A}_{11}\frac{x_1^2}{2} \,+\, \mathrm{A}_{13} \,+\, 2\frac{\mathrm{q}}{k}\,, \tag{9}$$

and for second part of pipeline in the range $x_2 \in \langle\, 0,\, \infty \rangle$

$$w_2(x_2) \;=\; \frac{\mathrm{q}_0}{k}\,, \tag{10}$$

where

$$\begin{aligned} \mathrm{A}_{11} \;&=\; -\frac{2\mathrm{L}\mathrm{q}_0\beta^3}{3k(1+\mathrm{L}\beta)}(3 \,+\, 3\mathrm{L}\beta \,+\, \mathrm{L}^2\beta^2)\,, \\ \mathrm{A}_{13} \;&=\; \frac{\mathrm{L}\mathrm{q}_0\beta}{6k}(6 \,+\, 6\mathrm{L}\beta \,+\, 4\mathrm{L}^2\beta^2 \,+\, \mathrm{L}^3\beta^3)\,. \end{aligned} \tag{11}$$

Graphically representation of deflection function from eqn.(9) and egn. (10) is in the fig.8. Comparing result from this paper in the fig.6 and fig.7 with the exact solution in the fig.8 we have the same result from using command DSolve in program *Mathematica* and exact solution in [3].

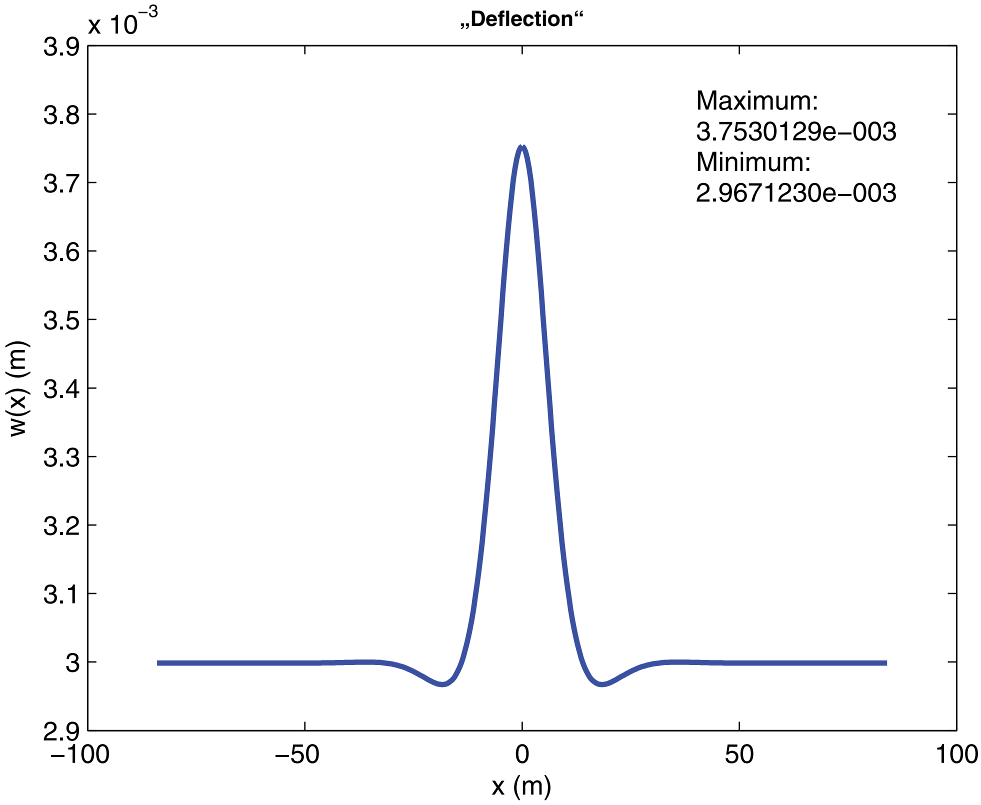


Figure 8: Graphically representation of deflection function from exact solution [3]

References

[1] FRYDRÝŠEK, K.: *Nosníky a rámy na pružnom podkladu 1.* VŠB-TU Ostrava, Ostrava, 2006.

[2] TIMOSHENKO, S.: *Strength of Materials, Advanced Theory and Problems, Part II.* 2nd Edition, D. Van Nostrand Co., Inc., New York, NY, 1947.

[3] JANČO, R.: *Numerical ananlysis of gas pipelines digging of during operation*, Inaugural thesis, SjF STU Bratislava, 2010.

Current address

MSc. Roland Jančo, PhD. ING-PAED IGIP
Institute of Applied Mechanics and Mechatronics, Section of Strength of Material,
Faculty of Mechanical Engineering, Slovak University of Technology Bratislava,
Nám. slobody 17, 812 31 Bratislava, Slovak Republic,
e-mail: roland.janco@stuba.sk .

Mgr. Monika Kováčová, PhD.
Institute of natural sciences, humanities and social sciences,
Faculty of Mechanical Engineering, Slovak University of Technology Bratislava,
Nám. slobody 17, 812 31 Bratislava, Slovensk Republika,
e-mail: monika.kovacova@stuba.sk .

Journal of Applied Mathematics

volume IV (2011), number I

PATTERN RECOGNITION AND SYSTEM ADAPTATION

JANOŠEK Michal, (CZ), KOCIAN Václav, (CZ),
KOTYRBA Martin, (CZ), VOLNÁ Eva, (CZ)

Abstract. In this paper we would like to introduce the pattern recognition approach based on neural networks. Patterns are an effective way to describe system's behaviours because we do not need to cover it every moment but we only search for certain patterns that appear from time to time. There are many ways how to recognize a pattern. In this study we use neural networks to do the job. Based on the pattern we recognize, it is possible to expect system's behaviour and adapt our acting desired way. There are many problems related to our research so in this paper we will concentrate on pattern recognition part mainly and present our current pattern recognition utility written in Java language and based on neural networks.

Key words. Pattern recognition, adaptation, prediction, neural networks, relative model, absolute model

Mathematics Subject Classification: Primary 68T10, 92B20; Secondary 68U35.

1 Patterns as system description

Using patterns we would like to describe particular system. It is important to realize that everything we observe is relative from our point of view. When we search for the pattern, we want to choose such pattern which represents the system reliably and define its important properties. Every pattern we find is always misrepresented with our point of view. Gershenson [1.] proposes two types of model, absolute and relative. The absolute model (*abs-model*) refers to what the thing actually is, independently of the observer. The relative model (*rel-model*) refers to the properties of the thing as distinguished by an observer within a context. Since the observer is finite and cannot gather complete information, *rel-models* are limited, whereas *abs-models* have an unlimited number of features. Since new observers can contemplate any *abs-model* from new contexts, there exists an infinity of potential *rel-models* for any *abs-model*. We can say that the *rel-model* is a model, while the abs-model is the modelled. Since we are all limited observers, it becomes clear that we can speak about reality only with *rel-beings/models* [2.]. It is obvious that we are not able to describe the whole system, all of its properties and behaviour. Using patterns, which only describe even smaller part of the system, we use much less information to describe our relative model. So it is important to choose and recognize patterns very carefully.

We can imagine a pattern as some object with same or similar properties. There are many ways how to recognize and sort them. Generally, we assign to some input value a pre-defined output value. For our purpose we use particular pattern recognition algorithm, which is classification [3.]. In this case we try to assign to some input value one of the output sets of values. So we have several sets with similar properties and assign each input value or values to the one of the pre-defined sets. Input value can be any data regardless it was originally text, audio, image or any other data. Such approach can work nearly with any system we would like to describe. But that is a very wide frame content.

Before we start it is crucial to choose particular system on which we will to able to conduct desired experiments. We need a system, where we can very easily observe certain patterns and where we can obtain observed data detail and without delay. For that reason we use Forex market. Patterns in the market (so as Forex) are not exact, they are slightly different every time they appear. They can have different amplitude and different duration, albeit visually the same pattern can look differently despite being the same. So it seems for us that Forex market should be very good in inexpensive working place.

Of course, similar approach can also work with other systems. The only thing that is necessary to change would be input format and expected outputs.

1.1 Forex market

Forex, or foreign exchanges, also known as the international foreign exchange market is a world foreign exchange trading. Simply it is a global network that is connected by modern means of communication. It is inter-connecting banks, insurance companies, investment funds and brokerage companies, which bring together individual investors. Their objective is the use certain mechanisms to valorise the investments based on the movements of currency. So Forex trader is a businessman who decides to speculate on the movements of exchange rates between currencies.

There are several reasons why to choose Forex market. There is a lot of data freely available in real time and in very good quality. Almost every broker company offers a demo platform and free online data for study. And there exist many described patterns whose emerge in the market as well.

Market systems are chaotic systems from their nature. For example, stocks are being sold or bought based on their prices. The price depends on how much has been bought or sold. The feedback loop has both positive and negative effects. The law of supply and demand implies a negative feedback loop, because the higher the price, the lower the demand, which, in fact, causes a lower price in the future. However, a parallel speculation mechanism implies a positive feedback loop, because an increasing price makes an assumption that the price will increase in the future and thus motivate the traders to buy more stocks. As we do not know delay the between these two effects we are not able to predict anything well. These nonlinear effects are common in the markets [2].

Market patterns can be classified into two categories: the continuation pattern and the reversal pattern. Continuation pattern indicates that the market price is going to keep its current movement trend; while the reversal pattern indicates that the market price will move to the opposite trend. Moreover reversal pattern does not necessarily suggest a complete reversal in trend, but merely a change or pause in direction. Patterns can be seen as some sort of maps which helps us to orientate in certain situations and navigate us to profitable trades. In the next two chapters we present important patterns that can be recognized in Forex market.

1.2 Support, Resistance pattern

One of the key points in the market is supports and resistances, or simply S/R. These are the points where the price has resisted increasing or decreasing its value. It means that traders are not willing to buy too high or sell too low. In fact, it is some sort of traders' psychical barrier. Let's define support and resistance. Support is a price below no one wants to sell; it's like a floor or bottom barrier which majorities of traders accept as minimal. Resistance is a price above no one wants to buy; it's like a ceiling or top barrier which majorities of traders accept as maximal.

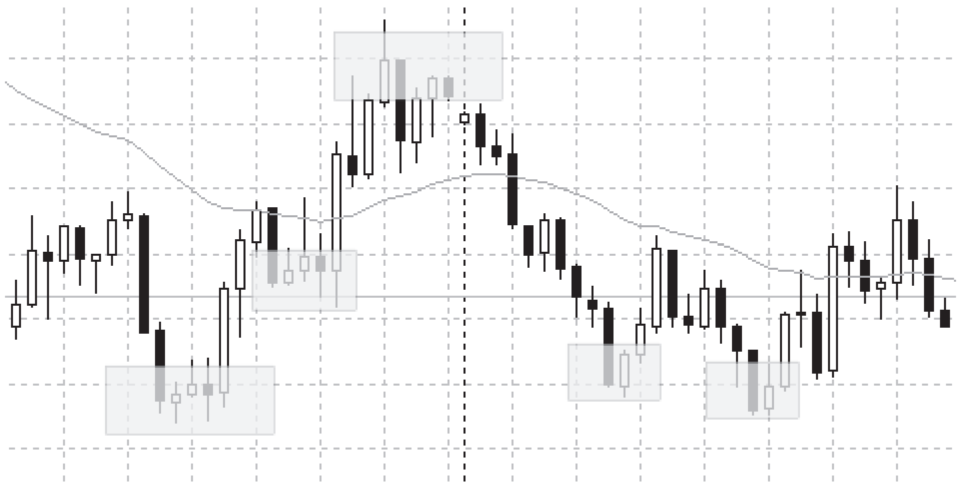


Fig. 1: Supports and resistance

In Fig. 1 we can see an example of four supports in the lower part of the image and one resistance in the upper part of the image. We can mention that supports quite often stop on previous support in the history, which is another important phenomenon that we will discuss in the next chapter.

1.3 Double top, Double bottom pattern

This pattern is based on support and resistance pattern. As it is mentioned above, supports and resistances emerge sometimes on same levels as in the history. This means that traders hesitate on that same price again and if the level is not breached is it quite strong patterns for us to expect what can happen. Double top means that traders refuse to buy at higher price twice and we can expect that price will decrease in nearby future. Contrary, double bottom means that traders refuse to sell at lower price twice and we can expect that price will increase in nearby future.

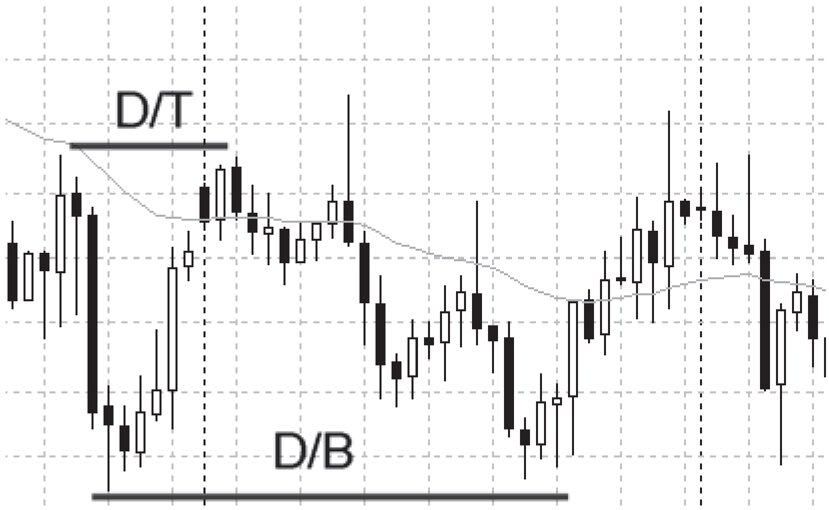


Fig. 2: Double top and double bottom

In Fig. 2 we can see an example of Double top (or D/T) pattern, after the pattern emerged the price has decreased. On the lower part of the image we can see Double bottom (or D/B) pattern, after the pattern emerged the price has increased. In this figure we can see another basic pattern. It is Uptrend and Downtrend. Uptrend is price rise between local minimum and maximum. Downtrend is price fall between local maximum and minimum. Both patterns can be seen between D/T and D/B pattern.

2 Neurotask

For experimental purposes original software Neurotask was designed and created to identify different types of patterns. It is a framework to support plug-ins for different types of tasks. So far Neurotask includes these plug-ins:

- *Hebbian learning for classification*
- *Multi-perceptron based classification*
- *Kohonen Self Organizing Maps*
- *Adaptive Resonance Theory (ART1)*
- *Pattern extraction from an OHLC data sets*
- *Pattern extraction from a general multidimensional data sets*

Software modular structure allows fast and easy to extend it for new functions. Currently these types of tasks are available:

- *Extracting patterns from data. Currently, the patterns can be extracted only from the table (CSV) data, but nothing prevents to add support for other (e.g. graphics) resources in the future.*
- *Patterns editing to a format suitable for particular type of recognition algorithm (neural networks).*
- *Pattern recognition experiments using different types of neural networks.*

All plug-ins are developed based on MWC (Model View Controller) design pattern. That facilitates to execute the plug-in in interactive mode either in non-interactive mode. In interactive mode it is possible to watch all calculations step by step in a real time. Also it is possible (and necessary in fact) to set up every single parameter of the plug-in. Also every subtask of a computation (pattern's set loading or mode switching) has to be executed by user. On the other hand in non-interactive mode every parameter's settings and subtask execution is read from a script and performed automatically. Consecutive calculation can be logged into a file for further check-out. It is obvious, that non-interactive mode is suitable for cases when neural network is just a part of some complex system.

2.1 Pattern recognition via Neurotask

As mentioned in previous chapter there is a possibility to choose from several pattern recognition plug-ins. In the following experiments we will focus on Hebbian learning classification.

Neural network topology using Hebbian learning is displayed in Fig. 3. The network consists of *n* input neurons *Xi (i=1 to n)*, *m* output neurons *Yj (j=1 to m)* and one bias *B* (neuron with constant value 1). Neurons are connected with each other using weighted connections.

- Each input neuron *Xi* is connected with each output neuron *Yj (j=1 to m)*
- Bias is connected with each output neuron *Yj (j=1 to m)*. Bias can be considered as permanently active input neuron.

Signals between neurons are propagated in a direction of arrows, i.e. from the input neurons *Xi* and bias *B* to the output neurons *Yj*. The weights, which are assigned to the connections, determine the ratio and a polarity of the transmitted signal.

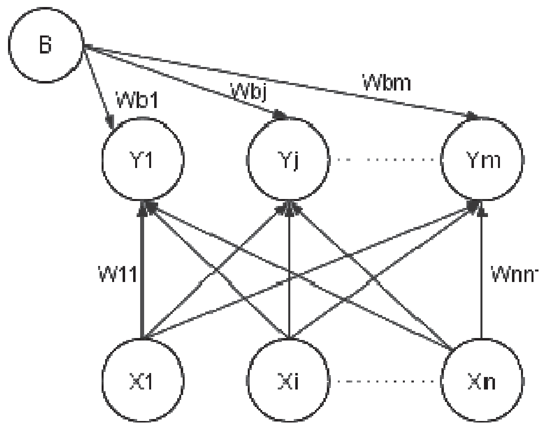


Fig. 3: Neural network using Hebbian learning

Network construction and initialisation

In the beginning all weight values *Wij* and *Wbj* are set to *0*, so as all input/output neuron values.

$$W_{ij} = 0, X_i = 0, T_j = 0 \qquad \text{for } i = 1...n \text{ and } j = 1...m \tag{2.1}$$

Learning

Following nomenclature will be used:

S – vector containing the original learning pattern (source)

X – vector containing input neuron values
T – vector containing expected network response
Y – vector containing output network values
Ec – error counter

The algorithm works with bipolar (1,-1) patterns (S and T vectors). During the learning period we repeat following steps for all learning patterns:

Network response calculation

1. submit learning pattern to a network ($X = S$)
2. calculate input values of every output neuron (2.2)

$$IN_j = \sum_{i=1}^{i=n} X_i . W_{ij} \tag{2.2}$$

3. calculate activation **Y** (2.3) of every output neuron according to the activation function in Fig. 4

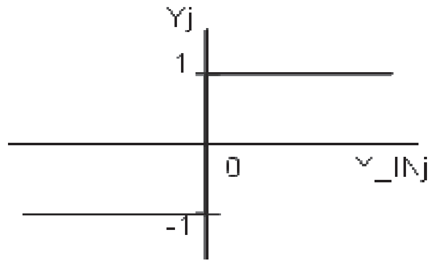


Fig. 4: Neural network activation function

$$(Yj = 1 \; pokud \; INj > 0, \; jinak \; Yj = -1) \tag{2.3}$$

Error counter and weights update

4. Compare network response **Y** with expected **T**. If **Y≠T**, increment error counter (2.4)

$$Ec = Ec + 1 \tag{2.4}$$

5. Perform connection weight update (2.5)

$$Wij\,(new) = Wij(old) + Xi.Tj \,, Wbj(new) = Wbj(old) + Tj \tag{2.5}$$

In the case that error counter value is *0* after learning patterns are submitted (i.e. every network response matched the expected) then learning is finished. In opposite case ***(Ec** > 0)* we have to repeat the learning cycle (steps 1-5) for every learning pattern.

Active mode

After learning process is finished, we can switch the network to the active mode. In this mode no weight changes are performed anymore. In the active mode we submit patterns to the network input and read the response from its output. In other words, steps 1-3 from previous chapter are performed. As imply from learning process ending condition ***(Ec = 0)***, the network will correctly classify every pattern which has been a member of the learning set. However, an interesting feature is the ability to generalize learned set of patterns that are similar to already learned patterns.

2.2 Experiment

In this experiment we would like to prove the network's ability to learn the basic representative pattern in Forex market. We have used two sets of patterns. Each set contains four patterns. The patterns from both sets have exactly the same meaning (uptrend, downtrend, resistance and support), but have been binarized using slightly different method. Patterns bitmaps (Fig. 5, Fig. 6), 8x8 pixels, have been transformed to 64 bit long vectors **S** (Tab. 1, Tab. 2), so that rows of the bitmap have been laid one after one in line. In output vectors **T** only one bit is active at the moment. The order of the active bit determines the number of pattern class:
1st – Uptrend
2nd – Downtrend
3rd – Resistance
4th – Support

By that we mark the input patterns from training set by proper classification class – desired response from neural network. We expect that number representing the classification class from the neural network. This process is necessary for supervised learning when inputs and desired response are submitted to the network. After the adaptation the network is able to classify submitted patterns.

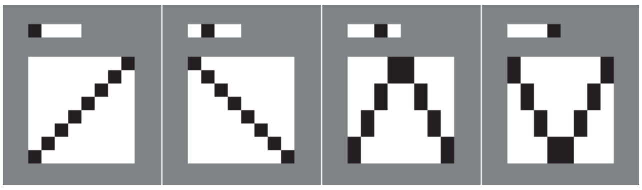


Fig. 5: Patterns 1-4, patterns bitmap (lower rectangle), desired response bitmap (upper rectangle)

p	**T**	**S**
1	1000	00000001|00000010|00000100|00001000|00010000|00100000|01000000|10000000
2	0100	10000000|01000000|00100000|00010000|00001000|00000100|00000010|00000001
3	0010	00011000|00011000|00100100|00100100|01000010|01000010|10000001|10000001
4	0001	10000001|10000001|01000010|01000010|00100100|00100100|00011000|00011000

Tab. 1: Patterns 1-4, vectors **S** and **T**

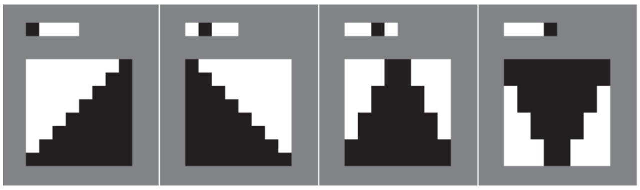


Fig. 6: Patterns 5-8, patterns bitmap (lower rectangle), desired response bitmap (upper rectangle)

p	**T**	**S**
5	1000	00000001|00000011|00000111|00001111|00011111|00111111|01111111|11111111
6	0100	11111111|01111111|00111111|00011111|00001111|00000111|00000011|00000001
7	0010	00011000|00011000|00111100|00111100|01111110|01111110|11111111|11111111
8	0001	11111111|11111111|01111110|01111110|00111100|00111100|00011000|00011000

Tab. 2: Patterns 5-8, vectors **S** and **T**

In the first round first four patterns (1-4) has been submitted to the neural network. After 5 learning cycles the network gave correct response only for two patterns from the four *(**Ec** = 2)*. In the next learning cycles the network did not converge (some weights oscillated around a low fixed value, other raised systematically.

In the second round second four patterns (5-8) has been submitted to the “clean” unlearned neural network. After 4 learning cycles the network converged and learned the patterns 5-8 correctly, so network could be switched to the active mode *(**Ec** = 0)*.

After that, patterns (1-4) were submitted to this second network and the network was able to recognize them. So the second network was able to recognize all eight patterns (1-8) even if it was taught using patterns (5-8) only. Original motivation for creating patterns (5-8) was to prove assumption that if the network can learn these “spatial” patterns network it will gain an ability to recognize patterns 1-4 as sub-patterns. Furthermore it allowed the network to gain more general knowledge about patterns characteristics and was able to recognize not only same but similar patterns as well. The experiment confirmed that if we submit “correct” training set to the network we are able to recognize patterns that the networks wasn’t able to learn, specifically the ability to recognize patterns (1-4) as same as patterns (5-8). It was shown that for this type of neural network it seems to be appropriate to use general like patterns for training. It presents a challenge in the theoretical field and requires more research.

3 System adaptation vs. prediction

Let's say we have built pattern recognition system and it is working properly to meet our requirements. We are able to recognize certain patterns reliably. What can we do next? Basically, we can predict systems behaviour or we can adapt to any change that emerge.

It is possible to try to predict what will happen, but more or less it's a lottery. We will never be able to predict such systems' behaviour completely. This doesn't mean it is not possible to build a system based on prediction [5.]. But there is another approach that tries to adapt to any change by reflecting current situation. To adapt on any change (expected or unexpected) it should be sufficient to compensate any deviation from desired course. In case that response to a deviation comes quickly enough that way of regulation can be very effective. It does not matter how complicated system is (how many factors and interactions has) in case we have efficient means of control [2]. To respond quickly and flexible it is desirable to have some expectation what can happen and what kind of response will be appropriate. We can learn such expectation through experiences.

4 Conclusion

We have presented the design of the pattern recognition application based on neural networks. Using patterns, which are only part of whole system characteristics, we can describe system or its behaviour. It is very important to choose patterns very carefully because every observer can see the system from his or her point of view. For our purpose we have used classification pattern recognition algorithm. Because of easy access to real time data we have decided to use Forex market data for experiments. To help with pattern recognition we have created pattern recognition software Neurotask which allows us to define and recognize patterns using different algorithms. It is Model View Controller design pattern based and all functionality is plug-in based.

For our experiment using Hebbian learning we choose four patterns - uptrend, downtrend, support, resistance. The network has to learn these patterns using binarized version of bitmap pattern representation. The most interesting discovery was that the neural network was able to learn "spatial" patterns much quicker than "line" patters (some of which was not able to learn). "Spatial" patterns can be considered as some sort of "line" patterns generalization. But that will require more research.

Among other open tasks to do are input data binarization, detail patterns descriptions and expectations what can happen after certain pattern appear.

Acknowledgement

The paper was supported University of Ostrava grant SGS/PřF/2011. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1.] GERSHENSON, C.: *Complex philosophy*. In *Proceedings of the 1st Biennial Seminar on Philosophical, Methodological & Epistemological Implications of Complexity Theory*. La Habana, Cuba. URL http://uk.arXiv.org/abs/nlin.AO/0109001, 2002.

[2.] GERSHENSON, C.: *Design and Control of Self-organizing Systems*. Dissertation, Vrije Universiteit Brussel, 2007.

[3.] CISKOWSKI, P. and ZATON, M.: *Neural Pattern Recognition with Self-organizing Maps for Efficient Processing of Forex Market Data Streams*. In Artificial Intelligence and Soft Computing, Volume 6113/2010, pp. 307-314, DOI: 10.1007/978-3-642-13208-7_39, 2010.

[4.] FAUSETT, L.: *Fundamentals of neural networks, architectures, algorithms and applications,* Prentice-Hall, ISBN 0-13-042250-9, 1994.

[5.] K.K. LAI, LEAN Yu and SHOUYANG Wang: *A Neural Network and Web-Based Decision Support System for Forex Forecasting and Trading*. In Data Mining and Knowledge Management, Volume 3327/2005, pp. 243-253, DOI: 10.1007/978-3-540-30537-8_27, 2005.

Current address

Michal Janošek, Mgr.
Department of Informatics and Computers, Faculty of Science,
University of Ostrava, 30. dubna 22,
701 03 Ostrava, Czech Republic,
+420 597 092 184,
e-mail: michal.janosek@osu.cz

Kocian Václav, Mgr.
Department of Informatics and Computers, Faculty of Science,
University of Ostrava, 30. dubna 22,
701 03 Ostrava, Czech Republic,
+420 597 092 184,
e-mail: vaclav.kocian@osu.cz

Kotyrba Martin, Mgr.
Department of Informatics and Computers, Faculty of Science,
University of Ostrava, 30. dubna 22,
701 03 Ostrava, Czech Republic,
+420 597 092 184,
e-mail: martin.kotyrba@osu.cz

Volná Eva, doc. RNDr. PaedDr., PhD.
Department of Informatics and Computers, Faculty of Science,
University of Ostrava, 30. dubna 22,
701 03 Ostrava, Czech Republic,
+420 597 092 184,
e-mail: vaclav.kocian@osu.cz

Journal of Applied Mathematics

volume IV (2011), number I

AGE STRUCTURE IN HISTORICAL POPULATIONS: MATHEMATICAL AND STATISTICAL APPROACH

KAČEROVÁ Eva, (CZ), HENZLER Jiří, (CZ)

Abstract.
The models analysed in this paper are based on data excerpted from one of the most cited sources of demographical data on the middle of the 17th century in Bohemia, the List of Inhabitants according to Faith of 1651. The List yields extensive data on serfs in several areas in Bohemia: on their age, profession, religion and marital status. Helas, data concerning the age of listed persons are in general rounded according to various rules. The aim of this article is to cope with this problem by means of mathematical modelling of age distribution and statistical analysis via moving averages.

Key words and phrases. Age structure in 17th century populations, way of rounding data, rounded data, missing data, mathematical models in demography.

Mathematics Subject Classification. Primary 62P02.

1 Introduction

A basic source for the study of the population structure according to the age and genders in the middle of the 17th century is the List of Serfs According to Faith of 1651 (Pazderova, 2002), whose origin has a connection with recatolisation efforts after the Thirty Years War. Immediate cause of the elaboration of the List was the edict of the vice-governors in Bohemia, made public in 4th of February, 1651. It stated shortcomings of the existing lists of noncatholic serfs and ordered to all suzerains to list, according to the enclosed form, all their serfs of both sexes, and to expedite it in the six weeks term to the vice governors office.

2 Analysed population

The data analysed in this article represent the section of the List concerning the Chrudim area. According to the List, 46 626 people lived in the Chrudim area, including 24 860 women.

2 104 persons missed the age record. Children of less than 12 years of age (age of the first confession) were listed rarely. The listing of these children was not complete in any dominion of the Chrudim area, so the children under 12 are not presented in the following analysis. For the demographic structure study, the age record is the most important. The exact age of the listed persons played rather a secondary role in the aims of the originators of the List and therefore it is often rounded or even missed. Scribes tended to round the age of those, who did not know their age exactly, to multiples of 10 or 5. Even numbers were in it more popular then the odd ones. Rounding of age was dependent on the social categories of the serfs, too. Widows were often ascribed the age of 40, or 60 if older. Alone women with children, i.e. probably unmarried mothers were frequently ascribed the age of 30, and so on. The measure of distortion can be measured by the index of cumulative age *ik*:

$$ik^p = (5 \cdot \sum_{x=0}^{7} S^p_{25+5x})(\sum_{x=23}^{62} S^p_x) \tag{1}$$

where S_x represents the number of persons in a given age group and the upper index p describes sex. Nevertheless, despite obvious inaccuracies and rounding, the age data from the List represent valuable information on the population at that time (Fig. 1).

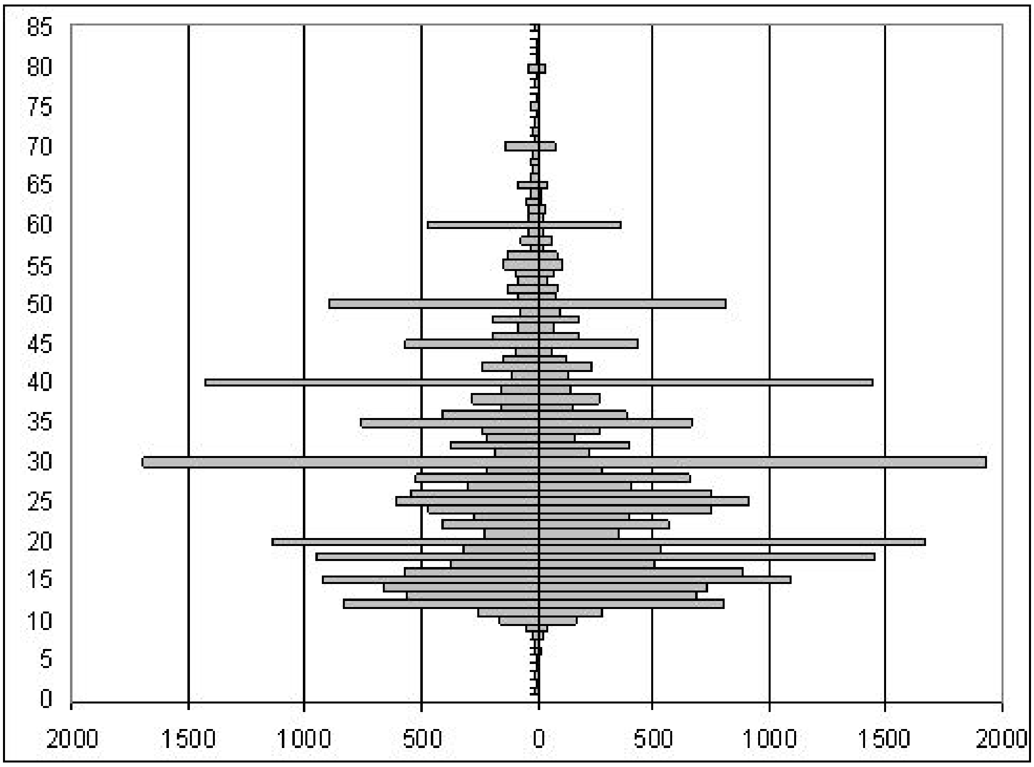


Fig. 1. Age structure of population of the Chrudim area according to the List of serfs according to faith of 1651

Rounding of age leads to distorted results concerning the population age structure. This distortion could be diminished by applying decennial age intervals, in which the most frequented value will be always in the middle, i.e. intervals 5−14, 15−24, etc. (Fig. 2). For the sake of comparability with the works of other authors dealing with the mentioned List, five-year intervals are also taken into account.

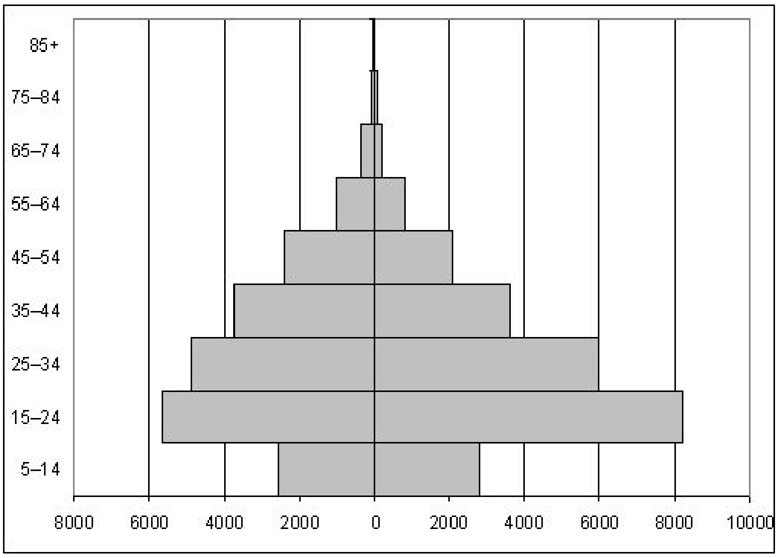


Fig. 2. Age structure of population of the Chrudim area according to the List of serfs according to faith of 1651 (decennial intervals)

The commonly used age intervals 0−4, 5−9, 10−14, etc. (Fig. 3) are also given. In some studies, decennial intervals 0−9, 10−19, etc. could be found. The choosing of type of age intervals could influence the weight of rounding in resulting age structure. In any case, it leads to the loss of information.

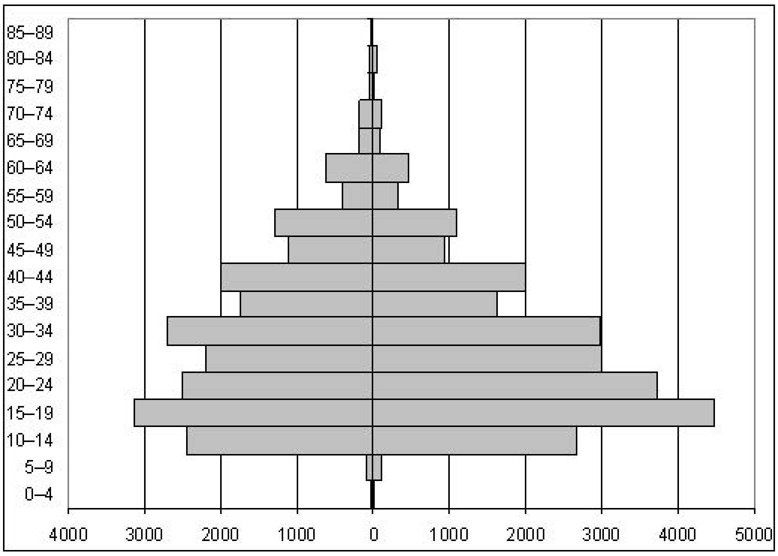


Fig. 3. Age structure of population of the Chrudim area according to the List of serfs according to faith of 1651 (five-year intervals)

3 Statistical smoothing of the data

Smoothing of the data using moving averages is a standard statistical technique. As the data were the most frequently rounded to the whole tens, the most suitable moving averages are those of ten subsequent values. In the mentioned List, age under 12 is significantly undervalued for two reasons: children under 12 were not included in the List or their data were included into the category of the confession age of 12 years. Therefore, the data concerning the age under 12 were not included into the processing. For avoiding distortion of the "smoothed" age pyramid in subsequent age categories by absence of the data under 12, are the moving averages corresponding to the years following the age of 12 computed on the base of shorter time intervals: for the age of 12 the original data are taken, the data for the age of 13 are smoothed by the average of three values for age 12, 13 and 14, the data for the age of 14 are smoothed by the average of five values for age 12, 13, 14, 15 and 16, the data for the age of 15 are smoothed by the average of seven values for age 12, 13, 14, 15, 16, 17 and 18, and finally the data for the age of 16 are smoothed by the average of nine values for age 12, 13, 14, 15, 16, 17, 18, 19 and 20. From the age of 17 onwards the data for a given age are smoothed by the moving average of ten subsequent values. As to the even number of the averaged values, the data of the age K, $K \geq 17$, will be smoothed by the arithmetic average of the two subsequent moving averages:

$$\frac{1}{2}\left(\frac{1}{10}\sum_{i=k-5}^{k+4} d_i + \frac{1}{10}\sum_{i=k-4}^{k+5} d_i\right), \tag{2}$$

where d_i is the original value from the List concerning the age i.

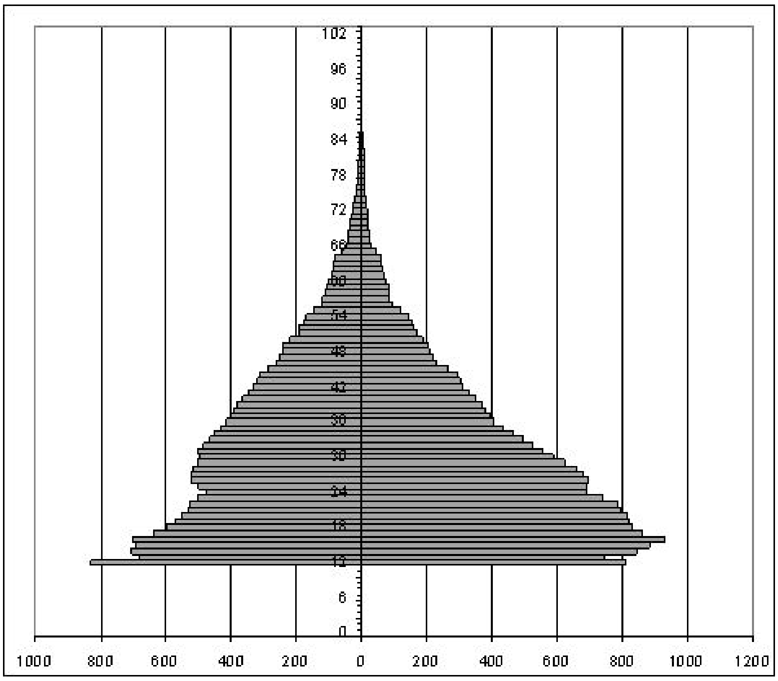


Fig. 4. Smoothing data by moving averages

4 Mathematical models

Another option to smooth the data is to create simple mathematical model supposing that the rounding process exerts some regularity. If the theoretical values computed on the base of such a model will not be systematically biased, a hypothesis on the possible way of rounding can be formulated. In this paper, two mathematical models are presented. In both models we suppose that rounding to whole tens concerns the ages not more than 4 years up and not more than 4 years down the rounded value and rounding to values ending by 5 concerns the ages not more than 2 years up and not more than 2 years down the rounded value. Rounding to even numbers concerns the neighbouring odd numbers. In both models an interesting phenomenon was taken into account: From the original age structure (Fig. 1) one can easily see that from the age of 24 are the frequencies concerning the ages ending by 6 significantly bigger (somewhere more than two times bigger) than the frequencies concerning the ages ending by 4. The explanation of this feature may be as follows:

1. rounding the age of 23, of 33 etc. took place directly up to 25, 35 etc., not to neighbouring even numbers 24, 34, etc.,

2. rounding the age of 25, 35 etc. took place up to 26, 36 etc. and not down to 24, 34 etc.

To estimate the values h_{10}, h_5, h_{even} that in original frequencies of ages ending by 0 or 5 and of even ages could be ascribed to rounding, and by using those values to estimate theoretical values Y_i from the original frequencies y_i , in both models we supposed that neighbouring theoretical values Y_{28}, Y_{29}, Y_{30}, Y_{31}, Y_{32} differ by the same value Δ; the same supposition holds for values Y_{33}, Y_{34}, Y_{35}, Y_{36}, Y_{37}, etc. Data concerning the age under 12 were from the same reasons as in the preceding chapter excluded from our models.

5 Fixed numbers' model

Fixed numbers' model consists in supposition that numbers of cases rounded to age ended by 0 and 5 and to even age are the same for all neighbouring ages. E.g. the same number of persons having a real age of 26, 27, 28, 29, 31, 32, 33, 34 was rounded to the age of 30, the same number of $h_5(35)$ persons having a real age of 33 or 34 or 37 was rounded to the age of 35 (as to the age of 36 see the the preceding chapter), etc.

From the equations

$$y_{28} = Y_{30} + 2\Delta - h_{10} + 2h_{\text{even}}(28), \tag{3}$$
$$y_{29} = Y_{30} + \Delta - h_{10} - h_{\text{even}}(28), \tag{4}$$
$$y_{30} = Y_{30} + 8h_{10}, \tag{5}$$
$$y_{31} = Y_{30} - \Delta - h_{10} - h_{\text{even}}(32), \tag{6}$$
$$y_{32} = Y_{30} - 2\Delta - h_{10} + 2h_{\text{even}}(32), \tag{7}$$
$$\tag{8}$$

we can compute unknowns Y_{30}, h_{10}, $h_{\text{even}}(28)$, $h_{\text{even}}(32)$ and Δ and the remaining values Y_{28}, Y_{29}, Y_{31} and Y_{32} and the same is valid in the similar equations for original data y_{38}, y_{39}, y_{40}, y_{41}, y_{42}, etc.

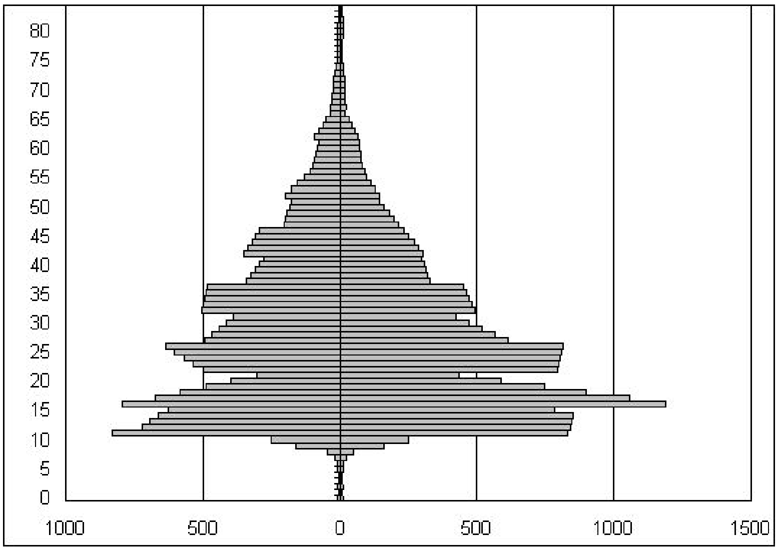


Fig. 5. Fixed numbers' model

In a similar way we can compute the system of equations

$$y_{33} = Y_{35} + 2\Delta - h_{10}(30) - h_5 - h_{\text{even}}(34) - h_{\text{even}}(32), \tag{9}$$

$$y_{34} = Y_{35} + \Delta - h_{10}(30) - h_5 - h_{\text{even}}(34), \tag{10}$$

$$y_{35} = Y_{35} + 4h_5 - h_{\text{even}}(36), \tag{11}$$

$$y_{36} = Y_{35} - \Delta - h_{10}(40) - h_5 + 2h_{\text{even}}(36), \tag{12}$$

$$y_{37} = Y_{35} - 2\Delta - h_{10}(40) - h_5 - h_{\text{even}}(36) - h_{\text{even}}(38) \tag{13}$$

and similar systems of equations for y_{43}, y_{44}, y_{45}, y_{46}, y_{47}, etc.

6 Linearly changing numbers' model

In linearly changing numbers' model we suppose that that numbers of cases rounded to age ended by 0 and 5 and to even age are the bigger, the closer is rounded age to the age, to which the rounding takes place. E.g. by rounding to an age of 30 a number of $h_{10}(30)$ persons of real age of 26 and 34 would be rounded but double number $2h_{10}(30)$ of persons of real age of 27 and 33, triple number $3h_{10}(30)$ of persons of real age of 28 and 32, and quadruple number $4h_{10}(30)$ of persons of real age of 29 and 31, and similarly by rounding to an age of 35. In the latter case, the effect of phenomenon 34, 36 will take place (see the chapter 4), etc.

Similarly as in the fixed numbers' model we solve the following system of equations

$$y_{28} = Y_{30} + 2\Delta - 3h_{10} + 2h_{\text{even}}(28), \tag{14}$$
$$y_{29} = Y_{30} + \Delta - 4h_{10} - h_{\text{even}}(28), \tag{15}$$
$$y_{30} = Y_{30} + 20h_{10}, \tag{16}$$
$$y_{31} = Y_{30} - \Delta - 4h_{10} - h_{\text{even}}(32), \tag{17}$$
$$y_{32} = Y_{30} - 2\Delta - 3h_{10} + 2h_{\text{even}}(32). \tag{18}$$

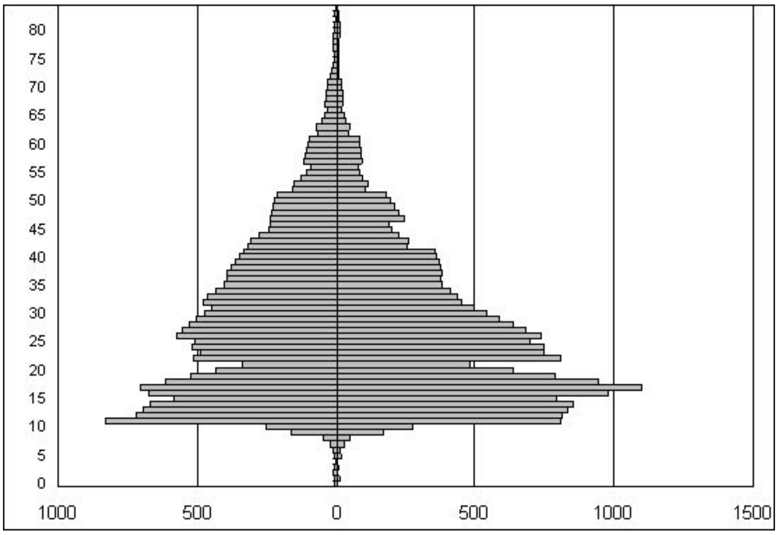


Fig. 6. Linearly changing numbers' model

and the system of equations

$$y_{33} = Y_{35} + 2\Delta - 2h_{10}(30) - h_5 - h_{\text{even}}(34) - h_{\text{even}}(32), \tag{19}$$
$$y_{34} = Y_{35} + \Delta - h_{10}(30) - 2h_5 - h_{\text{even}}(34), \tag{20}$$
$$y_{35} = Y_{35} + 4h_5 - h_{\text{even}}(36), \tag{21}$$
$$y_{36} = Y_{35} - \Delta - h_{10}(40) - 2h_5 - 2h_{\text{even}}(36), \tag{22}$$
$$y_{37} = Y_{35} - 2\Delta - 2h_{10}(40) - h_5 - h_{\text{even}}(36) - h_{\text{even}}(38). \tag{23}$$

7 Other possibilities in modelling

Supposing the same difference between the 5 consecutive theoretical frequences, we may variate the wages of quantities h_{10}, h_5, h_{even}; in the fixed numbers' model they were all equal to 1, in the linearly changing numbers' model they were 1,2,3,4 in the case of h_{10} and 1,2 in the case of h_{10}.

We may also abandone the supposition of linearity in the series if 5 consecutive theoretical frequences. The shape of curve fitting the age structure may be quessed from the above given moving averages model.

8 Conclusions

From the theoretical bihistograms (Fig. 5 and 6) there is obvious that for the studied real population the linearly changing numbers' model exerts smaller systematic deviation (periodicity of theoretical values) than the fixed numbers' model. On the other hand, the fixed numbers' model seems to be better in assessing the female part of population. Bihistogram constructed on the basis of moving averages yields global view on the population only. The crucial difference between the statistical approach based on the moving averages and mathematical approach using fixed numbers' model or linearly changing numbers' model consists in the following fact. Whereas mathematical models bring into the data hypotheses based partly on the present knowledge on construction such 17th century lists, the statistical approach is based on existing data only.

Acknowledgement

The paper was supported by grant no. 2D06026 with title "Reproduction of Human Capital" financed by the Ministry of Education, Youth and Sport within the framework of National Research Program II.

References

[1] DUCHÁČEK, K., FIALOVÁ, L., HORSKÁ, P., REPASOVÁ, M., SLÁDEK, M.: *On using the 1661-1839 lists of subject of the Trebon dominion to study the age structure of the population.* Historická demografie, Vol. 13, pp. 59-75, 1989.
[2] FIALOVÁ, L., KUČERA, M., MAUR, E., HORSKÁ, P., MUSIL, J., STLOUKAL, M.: *Dějiny obyvatelstva českých zemí [History of the population of the Czech Lands].* Mladá fronta, Praha, 1998.
[3] MAUR, E.: *Problémy demografické struktury Čech v polovině 17. století [Problems of the demographic structure of Bohemia in the middle 17th century].* Praha, Československý časopis historický, Vol. 19, pp. 849-850, 1971.
[4] PAZDEROVÁ, A.: *List of serfs according to faith - The Chrudim area.* Národní archív, Praha, 2002.

Current address

RNDr. Eva Kačerová
University of Economics Prague, Departement of Demography, nám. W. Churchilla 4, 130 67, Praha 3, Czech Republic, tel. no. +420224095273 $e-mail:$ *kacerova@vse.cz*

doc. RNDr. Jiří Henzler, CSc.
University of Economics Prague, Departement of Mathematics, Ekonomická 957, 148 00, Praha 4, Czech Republic, tel. no. +420224094246 $e-mail:$ *henzler@vse.cz*

Journal of Applied Mathematics

volume IV (2011), number I

STABILITY ANALYSIS FOR ROBUST CONTROL IN MATLAB

KVAPIL David, (CZ)

Abstract. Stability analysis is a very important task in the mathematical control theory and its practical application. Stability and performance are two of the fundamental issues in the design, analysis and evaluation of control systems. Robust systems are designed to preserve stability and performance in various classes of uncertainties.
During modeling and identification of the heat distribution and consumption we build models of elements of heat plants and of distribution networks. We need to check or guarantee the stability of state matrices and polynomials. Several own created m-files in MATLAB®, which can be useful for the robust stability analysis of systems, will be presented. And their using in concrete examples will be demonstrated.

Key words. Mathematical Control Theory, Stability, Robustness, Mathematical Modeling, MATLAB

Mathematics Subject Classification: Primary 93D09, 93D21; Secondary 93A30.

1 Introduction

Consider continuous time (CT) or discrete time (DT) linear time-invariant (LTI) state-space system, respectively

$$\dot{x}(t) = Ax(t) + Bu(t) \qquad\qquad x(k+1) = Ax(k) + Bu(k)$$
$$y(t) = Cx(t) + Du(t) \qquad\qquad y(k) = Cx(k) + Du(k)$$

where $t \in \mathbf{R}$, $k \in \mathbf{Z}$. System is asymptotically stable iff (i.e. if and only if) (see [7.][8.][12.][13.][14.])

$$\operatorname{Re}(\lambda_i) < 0 \quad \text{ or } \quad |\lambda_i| < 1,$$

where $i = 1,\dots,n$, respectively, for CT or DT case. For $\operatorname{Re}(\lambda_i) = 0$ or $|\lambda_i| = 1$ is the system on stability domain boundary. Matrix A is called Hurwitz or Schur (convergent), respectively, if all their eigenvalues have negative real part (for CT case) or their absolute values are less than one (for DT case). Similarly for polynomials, all their roots must lie in the open left complex half-plane or in the open unit disc.

If we have a system, then we must use different methods for Hurwitz stability in CT case or Schur stability in DT case. Alternatively, the DT task can be converted to the CT problem using the bilinear transformation

$$w = \frac{z+1}{z-1}$$

which transform the DT stability region (unit circle) to the CT stability region (open left half-plane).

Robust control deals with uncertainties in systems, robust methods aim to achieve performance and stability in the presence of perturbations and bounded modeling errors. Important class of uncertain systems is parametric uncertainty, where we assume that we have an accurate model but not exact value of one or more parameters. Uncertain (perturbation) parameters are expressed by a vector of uncertainty. Bounded set of parameters is often known in advance, usually takes the form of ball in an appropriate norm. If we add to the uncertain system a restrictive set of parameters we get a family of systems.

2 Quasipolynomials

For control of the systems with time delays we deal with characteristic equation in the quasipolynomial form

$$\delta(s) = d(s) + \mathrm{e}^{-sT_1} n_1(s) + \mathrm{e}^{-sT_2} n_2(s) + \ldots + \mathrm{e}^{-sT_m} n_m(s),$$

where we assume $\deg\left[d(s)\right] = n$, $\deg\left[n_i(s)\right] < n$ and $0 < T_1 < T_2 < \ldots < T_m$. The quasipolynomial family has a form

$$Q(s,\lambda) = d(s,\lambda) + \mathrm{e}^{-sT_1} n_1(s,\lambda) + \mathrm{e}^{-sT_2} n_2(s,\lambda) + \ldots + \mathrm{e}^{-sT_m} n_m(s,\lambda),$$

where $\lambda \in [a,b]$. The stability of $\delta(s)$ can be checked using the interlacing condition (Hermite – Biehler Theorem, [1.]). We can check it using own created MATLAB m-file **prolozeni2.m**, where the inputs are polynomials. For our concrete quasipolynomial $\delta(s) = d(s) + \mathrm{e}^{-sT_1} n_1(s) + \mathrm{e}^{-sT_2} n_2(s)$, where

$$d(s) = s^9 + 5s^8 + 20s^7 + 100s^6 + 200s^5 + 100s^4 + 100s^3 + 50s^2 + 15s + 1,$$

$$n_1(s) = 3s^8 + 10s^7 + 10s^6 + 15s^5 + 100s^4 + 50s^3 + 50s^2 + 10s + 2,$$

$$n_1(s) = 2s^8 + 22s^7 + 35s^6 + 51s^5 + 131s^4 + 130s^3 + 55s^2 + 24s + 3,$$

and $T_1 = 0.1$, $T_2 = 0.3$, we have:

```
>> [ZN,MatKor] = prolozeni2(-1.5,1.5,D,N1,T1,N2,T2);
Stejna znamenka pilotnich prvku
>>
```

We can see in the next figure that our quasipolynomial satisfies the interlacing condition and it is stable.

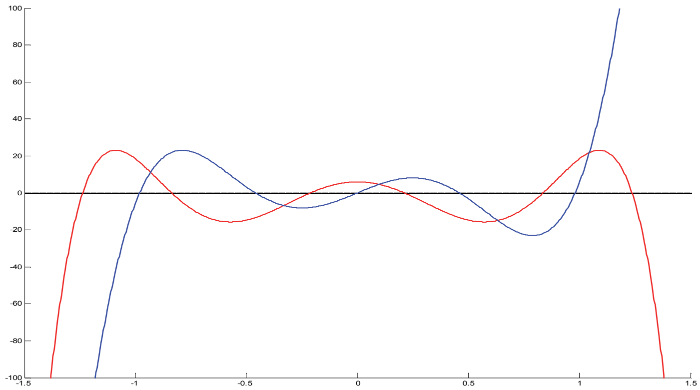


Fig.1: Interlacing property of quasipolynomial

3 Tsypkin – Polyak Theorem

Consider a ball of polynomials specified by a weighted l_p norm in the coefficient space for an arbitrary $p \in \mathbf{Z}^{+}$. Let us parametrized the real polynomial

$$A(s) = a_0 + a_1 s + ... + a_n s^n$$

by its coefficient vector $a = [a_0, a_1, ..., a_n]$ and consider a family of polynomials $A(s)$ centered at a nominal point $a^0 = [a_0^0, a_1^0, ..., a_n^0]$ with the coefficient lying in weighted l_p ball

$$B_p(a^0, \rho) := \left\{ a : \left[\sum_{k=0}^{n} \left| \frac{a_k - a_k^0}{\alpha_k} \right|^p \right]^{1/p} \leq \rho \right\},$$

where $\alpha_k > 0$ are given weights, $1 \leq p \leq \infty$ is a fixed integer and $\rho \geq 0$ is a prescribed common margin for the perturbations. Three special cases are considered: $p = 1$ (octahedral uncertainty), $p = 2$ (ellipsoidal uncertainty) and $p = \infty$ (interval uncertainty). Robust stability means that a ball of prescribed radius in a certain norm contains only Hurwitz polynomials of degree n. Tsypkin – Polyak theorem determines the Hurwitz stability in the ball $B_p(a^0, \rho)$ for prescribed value ρ and finds the maximal ρ which guarantees robust stability.

Let $A_0(j\omega) = U_0(\omega) + j\omega\, V_0(\omega)$, $U_0(\omega) = a_0^0 - a_2^0\omega^2 + a_4^0\omega^4 - ...$, $V_0(\omega) = a_1^0 - a_3^0\omega^2 + a_5^0\omega^4 - ...$.

For $1 < p < \infty$ we define

$$S_p(\omega) := \left[\alpha_0^q + (\alpha_2\omega^2)^q + (\alpha_4\omega^4)^q + ... \right]^{1/q}, \; T_p(\omega) := \left[\alpha_1^q + (\alpha_3\omega^2)^q + (\alpha_5\omega^4)^q + ... \right]^{1/q},$$

where $\frac{1}{p} + \frac{1}{q} = 1$. For $p = 1$ we have $S_1(\omega) := \max\limits_{k \text{ sudé}} \alpha_k \omega^k$, $T_1(\omega) := \max\limits_{k \text{ liché}} \alpha_k \omega^{k-1}$, and for $p = \infty$ is $S_\infty(\omega) := \alpha_0 + \alpha_2\omega^2 + ...$, $T_\infty(\omega) := \alpha_1 + \alpha_3\omega^2 + ...$. For each p is

$$z(\omega) := \frac{U_0(\omega)}{S_p(\omega)} + j\frac{V_0(\omega)}{T_p(\omega)} = x(\omega) + jy(\omega).$$

The theorem says that each polynomial in the ball $B_p\left(a^0,\rho\right)$ is Hurwitz iff the plot of $z(\omega)$: [1.]

a) goes through n quadrants counterclockwise direction,

b) does not intersect l_p disc $D_p(\rho) := \left\{ z = x + jy : \left[|x|^p + |y|^p \right]^{\frac{1}{p}} \le \rho \right\}$,

c) its boundary points $z(0)$ and $z(\infty)$ have coordinates with absolute values greater than ρ.

From this theorem the maximal ρ preserving Hurwitz stability of the ball can be found by finding the radius of the maximal l_p disc that can be inscribed in the frequency plot $z(\omega)$.

For $p = \infty$ the frequency plot $z(\omega) = x(\omega) + jy(\omega)$ is given by

$$x(\omega) = \frac{a_0^0 - a_2^0\omega^2 + a_4^0\omega^4 - \dots}{\alpha_0 + \alpha_2\omega^2 + \alpha_4\omega^4 + \dots},\; y(\omega) = \frac{a_1^0 - a_3^0\omega^2 + a_5^0\omega^4 - \dots}{\alpha_1 + \alpha_3\omega^2 + \alpha_5\omega^4 + \dots}$$

and must not intersect the square $|x| \le \rho$, $|y| \le \rho$. This procedure is used in the m-file **Tsy_Poly_locus.m**. If we have the polynomial

$$A(s) = s^6 + 14s^5 + 80.25s^4 + 251.25s^3 + 502.25s^2 + 667.25s + 433.5$$

with $\alpha = \left[0.1,\ 1.4,\ 5.6175,\ 15.075,\ 25.137,\ 33.36,\ 43.35\right]$, we have:

```
Zadej koeficienty nominalniho polynomu: [1 14 80.25 251.25 502.25 667.25 433.5]
Zadej koeficienty polynomu vah: [0.1 1.4 5.6175 15.075 25.137 33.36 43.35]
>>
```

and we obtain stability margin

```
>> rho = 1.23
```

In the following figure is graphical test of stability margin.

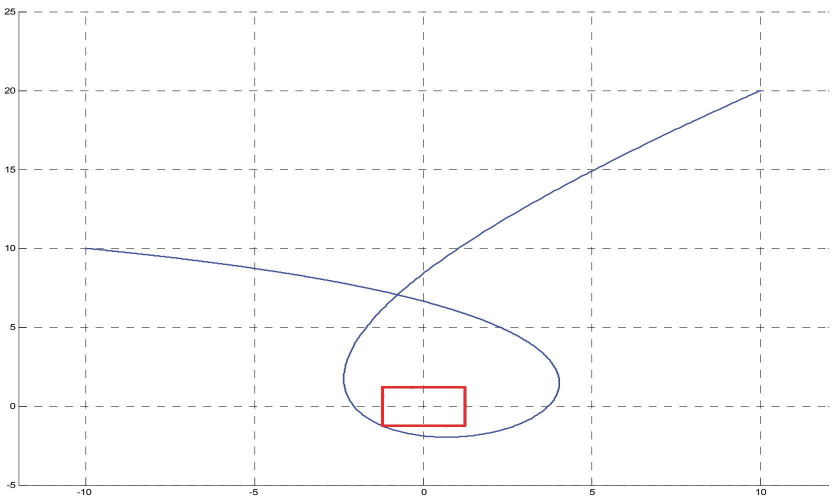


Fig.2: Tsypkin – Polyak locus (l_∞ stability)

4 Disc Polynomials

Consider now an alternative model of uncertainty – a set of disc polynomials, which are characterized by the fact that each coefficient can be any complex number in an arbitrary but fixed disc of the complex plain. Consider the stability region S and $n+1$ arbitrary discs D_i, $i = 0,1,...,n$, in the complex plane. Each disc is centered at the point β_i and has radius $r_i \geq 0$. Now let F_D be the family of all complex polynomials $\delta(z) = \delta_0 + \delta_1 z + ... + \delta_n z^n$ such that $\delta_j \in D_j$ for $j = 0,1,...,n$, i.e. every coefficient δ_j of the polynomial $\delta(z)$ in F_D satisfies $|\delta_j - \beta_j| \leq r_j$. We assume that every polynomial in F_D is of degree n.

Let $\beta(s) = \beta_0 + \beta_1 s + ... + \beta_n s^n$ be the center polynomial that is the polynomial whose coefficients are the centers of the discs D_j. It can be shown that each member of the family of polynomials F_D is Hurwitz iff

a) $\beta(s)$ is Hurwitz, and

b) $\|g_1\|_\infty < 1$ and $\|g_2\|_\infty < 1$,

where $g_1(s) = \dfrac{\gamma_1(s)}{\beta(s)}$, $g_2(s) = \dfrac{\gamma_2(s)}{\beta(s)}$, and $\gamma_1(s) := r_0 - jr_1 s - r_2 s^2 + jr_3 s^3 + r_4 s^4 - jr_5 s^5 - ...$,

$\gamma_2(s) := r_0 + jr_1 s - r_2 s^2 - jr_3 s^3 + r_4 s^4 + jr_5 s^5 - ...$.

Note that for proper Hurwitz complex rational function $g(s) = n(s)/d(s)$ is the H_∞-norm define

$$\|g(s)\|_\infty := \sup_{\omega \in \mathbb{R}} \left| \frac{n(j\omega)}{d(j\omega)} \right|.$$

The own created MATLAB procedure **Disc_polynom.m** is able to check the stability of disc polynomials. The inputs are vector of polynomials coefficients and vector of disc radii, the output is decision about stability. For concrete complex disc polynomial we have:

```
Zadej koef.nominalniho polynomu: [-1-j*11 3.5-j*18 9-j*27 1.5-j*6 2-j*3.5]
Zadej koeficienty polomeru: [1 3 8 1 2]

Nominalni polynom je stabilni
ans =
-0.4789 - 1.5034i
-0.9792 + 1.0068i
-0.1011 + 0.4163i
-0.0351 - 0.3828i

!!! Rodina diskovych polynomu NENI stabilni !!!

Stabilita_Rodiny = 0
>>
```

And we obtain the decision that our disc polynomial is not stable.

5 Polytopic Systems – Bialas Test

Consider a segment of stable polynomials p_1, p_2

$$p_\lambda(s) := \lambda p_1(s) + (1-\lambda) p_2(s), \text{ where } \lambda \in [0,1].$$

Bialas test [2.][17.] is for polynomials p_1, p_2 which are strictly Hurwitz and their leading coefficients are nonnegative and the remaining coefficients are positive. Let P_1 and P_2 be their Hurwitz matrices and define the matrix

$$W = -P_1 P_2^{-1}.$$

Then each of the polynomials $p(s) = \lambda p_1(s) + (1-\lambda) p_2(s)$, $\lambda \in [0,1]$, is strictly Hurwitz iff the real eigenvalues of W are all strictly negative.
Using own created m-file **PT_Bialas.m** we can check stability of the polynomial segment created by the polynomials $p_1 = s^3 + 3s^2 + 4s + 6$ and $p_2 = s^3 + 5s^2 + 7s + 2$.

```
Zadej koeficienty prvniho polynomu: [1 3 4 6]
Zadej koeficienty druheho polynomu: [1 5 7 2]

Nominalni polynom je stabilni
Nominalni polynom je stabilni

W =
   -0.4545   -0.7273         0
   -0.0909   -0.5455         0
   -0.3636    1.8182   -3.0000

lam =
   -3.0000
   -0.7611
   -0.2389

Oba polynomy tvori stabilni usecku
>>
```

Thus our segment of polynomials is stable.

6 Generalized Kharitonov Theorem

Consider the controller C and the plant P, respectively,

$$C(s) = \frac{F_1(s)}{F_2(s)}, \; P(s,q,r) = \frac{P_1(s,q)}{P_2(s,r)} = \frac{\sum_{i=0}^{n} q_i s^i}{\sum_{j=0}^{d} r_j s^j},$$

where $F_1(s)$, $F_2(s)$ are fixed polynomials, $q_i \in \left[\underline{q}_i, \overline{q}_i\right]$, $i = 0,1,\ldots,n$ $r_j \in \left[\underline{r}_j, \overline{r}_j\right]$, $j = 0,1,\ldots,d$.
Then the characteristic equation of the closed-loop system is $p(s) = F_1 P_1 + F_2 P_2$. The generalized Kharitonov segments are

$\Delta_E^1(s) = F_1\left(\lambda K_1^i(s) + (1-\lambda) K_1^j(s)\right) + F_2 K_2^k(s)$, $\Delta_E^2(s) = F_1 K_1^k(s) + F_2\left(\lambda K_2^i(s) + (1-\lambda) K_2^j(s)\right)$,

where $i, j, k \in \{1,2,3,4\}$, $i \neq j$, $\lambda \in [0,1]$, and Kharitonov polynomials of the numerator and denominator, respectively, are

$$K_1^1 = \underline{q}_0 + \underline{q}_1 s + \overline{q}_2 s^2 + \ldots \qquad K_2^1 = \underline{r}_0 + \underline{r}_1 s + \overline{r}_2 s^2 + \ldots$$

$$K_1^2 = \underline{q}_0 + \overline{q}_1 s + \overline{q}_2 s^2 + \ldots \qquad K_2^2 = \underline{r}_0 + \overline{r}_1 s + \overline{r}_2 s^2 + \ldots$$

$$K_1^3 = \overline{q}_0 + \underline{q}_1 s + \underline{q}_2 s^2 + \ldots \qquad K_2^3 = \overline{r}_0 + \underline{r}_1 s + \underline{r}_2 s^2 + \ldots$$

$$K_1^4 = \overline{q}_0 + \overline{q}_1 s + \underline{q}_2 s^2 + \ldots \qquad K_2^4 = \overline{r}_0 + \overline{r}_1 s + \underline{r}_2 s^2 + \ldots$$

Now we have 32 characteristic equations. Generalized Kharitonov Theorem says [1.][17.] that the robust stability of the closed loop system is equivalent to stability of all 32 generalized Kharitonov segments, than we can speak that the controller robustly stabilizes the interval plant.

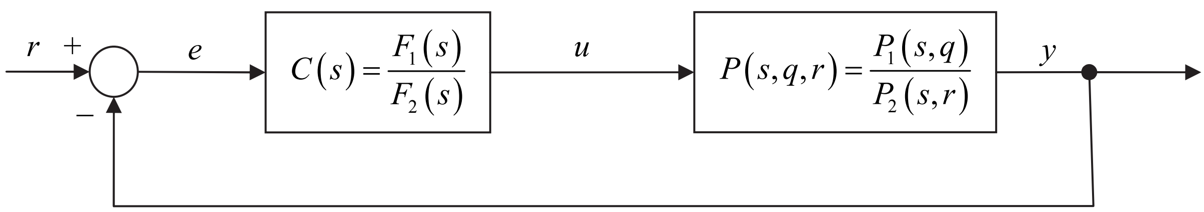


Fig.3: Generalized Kharitonov Theorem – stabilization of the interval plant

If we have the controller and interval plant

$$C(s) = \frac{s^2 + 2s + 1}{s^3 + 2s^2 + 2s + 1}, \quad P(s) = \frac{[0.1\ 0.2]s + [0.9\ 1.0]}{[0.9\ 1.0]s^2 + [1.8\ 2.0]s + [1.9\ 2.1]},$$

we write

```
>> [ZN,MatKor] = gkt(Nc,Dc,Np,Dp);
```

where

```
>> Nc = [1 2 1];
>> Dc = [1 2 2 1];
>> Np = [0.1 0.2; 0.9 1];
>> Dp = [0.9 1; 1.8 2; 1.9 2.1];
```

and we obtain

```
Regulator stabilizuje intervalovou soustavu

Stabilizace = 1
>>
```

thus the controller robustly stabilizes our interval plant.

Conclusion

Stability of systems is very important in analysis of practical problems. Presented own created MATLAB m-files could be very useful for many process tasks and real situations. Mathematical modeling of robust systems is useful and practical branch of the control theory. This advanced and perspective approach plays a significant role in applied mathematical control theory.

Acknowledgement

The paper was supported by grant from Ministry of Education Czech Republic Nr. 2C06007 with title "Intelligent Control System for Municipal District Heating Networks".

References

[1.] BHATTACHARYYA, S.P., CHAPELLAT, H., KEEL, L.H.: *Robust Control: The Parametric Approach*. Prentice Hall, 1995.
[2.] BOSGRA, O. H., KWAKERNAAK, H., MEINSMA, G.: *Design Methods for Control Systems*. DISC, Delft, 2007
[3.] HAVLENA, V., ŠTECHA, J.: *Moderní teorie řízení*. ČVUT, Praha, 1999.
[4.] HEIJ, C., RAN, A., van SCHAGEN, F.: *Introduction to Mathematical Systems Theory*. Birkhauser Verlag, Basel, 2007.
[5.] IOANNOU, P.I., SUN J.: *Robust Adaptive Control*. Prentice Hall, 1996.
[6.] KALAS, J.: *Analýza v komplexním oboru*. MU, Brno, 2006.
[7.] KALAS, J., Ráb, M.: *Obyčejné diferenciální rovnice*. MU, Brno, 1995.
[8.] KURZWEIL, J.: *Obyčejné diferenciální rovnice*. SNTL, Praha, 1978.
[9.] MACKENROTH, U.: *Robust Control Systems*. Springer Verlag, 2004.
[10.] PÍRKO, Z., VEIT, J.: *Laplaceova transformace*. SNTL, Praha, 1972.
[11.] REKTORYS, K.: *Přehled užité matematiky*. Prometheus, Praha, 1995.
[12.] SONTAG, E.D.: *Mathematical Control Theory*. Springer Verlag, 1998.
[13.] ŠTECHA, J., HAVLENA, V.: *Teorie dynamických systémů*. ČVUT, Praha, 2005.
[14.] STREJC, V.: *Stavová teorie lineárního diskrétního řízení*. Academia, Praha, 1978.
[15.] ŠULC, B., VÍTEČKOVÁ, M.: *Teorie a praxe návrhu regulačních obvodů*. ČVUT, Praha, 2004
[16.] ŠVARC, I., ŠEDA, M., VÍTEČKOVÁ, M.: *Automatické řízení*. CERM, Brno, 2007.
[17.] VESELÝ, V., HARSÁNYI, L.: *Robustné riadenie dynamických systémov*. STU, Bratislava, 2008

Current address

David KVAPIL, RNDr.
UNIS, a.s.
Department of Optimisation and Advanced Control
Jundrovská 33, 624 00 Brno, CZ
e-mail: dkvapil@unis.cz

Journal of Applied Mathematics

volume IV (2011), number I

TRUST AFFECTION MODEL APPLICATION FOR SOCIAL ISSUES

NETRVALOVA Arnostka, (CZ), SAFARIK Jiri, (CZ)

Abstract. The paper deals with application of Trust Affection Model Framework for modelling social issues. Terms trust, phenomenal trust as a modification of impersonal trust, and trust representation are introduced and model of trust affection is presented. Model design using multi-agent system is described and applied to real data. These data deal with the public opinion poll of social impact assessment of economic crisis. Survey was acquired from websites articles of the Institute of Sociology of the Academy of Sciences of the Czech Republic.

Key words. Trust, trust modelling, impersonal trust, trust affection

Mathematics Subject Classification: Primary 93A30, 94A17; Secondary 03B42.

1 Introduction

Many studies coming from psychological or social sciences describe the meaning and characteristics of trust [7], [2], and [3]. Computational models for exploration of trust formation were created e.g. in [8] and [5]. Wide-spreading of e-service [6], e-commerce, e-banking, etc., raise question of human machine trust. Further, trust plays an important role in peer-to-peer networks, and multi agent systems [9], where humans and/or machines have to collaborate. The aim of our work is simulation of the trust evolution under intentional trust affection applied to social issues.

2 Trust and Trust Representation

The acceptation of the term trust is wide [1]. Based on Gambetta [3], we interpret trust as a confidence in the ability or intention of a person to be of benefit to trustworthy something or someone at sometime in future. Trust in our model is represented by a value from continuous interval $\langle 0, 1 \rangle$. Value 0 represents complete distrust and value 1 means blind trust. Trust evolves not only within personal relations, i.e. personal trust, but person can trust to a phenomenon, so called phenomenal trust, that is the modification of impersonal trust. In this case, trust is formed towards the exclusive values of a given phenomenon, e.g., into possible products of the same kind from different producers.

Consider a group of n subjects represented as the set $S = \{s_1, s_2, \ldots, s_n\}$, and a phenomenon of m products represented as the set $P = \{p_1, p_2, \ldots, p_m\}$. Trust of subject s_i, $i = 1, \ldots n$, to product p_k, $k = 1, \ldots, m$, is denoted as follows

$$t_i^k = t(s_i, p_k),\ t_i^k \in \langle 0, 1\rangle, \text{ and } \sum_{k=1}^{m} t_i^{\,k} = 1 . \tag{1}$$

3 Intervention Model

The general model of information intervention effect is depicted in the Figure 1 (Vavra F., University of West Bohemia, personal communication) will be applied.

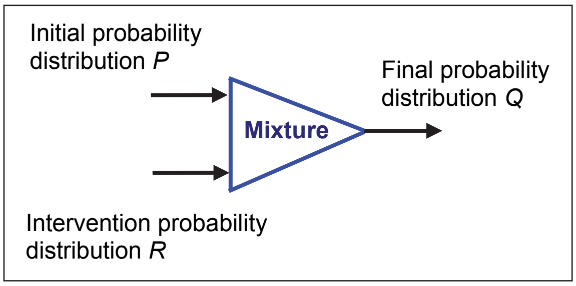


Figure 1 Trust Probability Distribution Mixture

Suppose finite set of events X with the probability distribution mass function $P(x)$, $x \in X$ on the input represents the state before intervention, e.g. initial probability of specific product preferences from a set of products of some kind. Probability distribution $Q(x)$ on the output describes the state after intervention activity. The intervention is modeled by probability distribution $R(x)$. The simple method for joining initial probability and intervention probability is their mixture

$$Q(x) = (1-\lambda)P(x) + \lambda\, R(x) \tag{2}$$

where $0 < \lambda \leq 1$, represents intensity of the intervention. Given probability mass functions $P(x)$, $R(x)$, $Q(x)$, the intensity λ can be found by the method of the least squares when all of probability mass functions are known

$$\lambda = \frac{\sum_{x \in X} (Q(x) - P(x))\,(R(x) - P(x))}{\sum_{x \in X} (R(x) - P(x))^2} \tag{3}$$

4 Phenomenal Trust Affection Model

Further, the presented intervention model will be applied for trust affection. We will define the dominant product p_d as a product a subject $s \in S$ trusts mostly. This trust is called t_d^s, $t_d^s = \max t(s, p_k)$, $k = 1, \ldots, m$. Values of subject's trust in other products are modeled by t_i^s=*(1-*

$t_d^s)/(m\text{-}1)$, $i \neq d$. Obviously, each value $t_d^s > 1/m$ makes a product dominant. The higher is value of t_d^s, the higher is trust of the subject in the dominancy of a product. Values t_d^s in each product are supposed to have approximately normal distribution in the population. Then, the population S can be divided into preferential classes according to the dominant product. Population's dominant product is the product, which is trusted mostly by the whole population.
Consider affection of trust in favor of some product in order to gain or even increase dominancy. This is modeled by mixture of intervention distribution I and initial trust distribution to the products of individuals. So, new trust probability distribution of an individual is given by values t'^k_i

$$t'^k_i = (1 - \lambda_i^k)\, t_i^k + \lambda_i^k\, I_i^k, \tag{4}$$

where $0 \le t_i^k \le 1$, $0 \le I_i^k \le 1$, $\sum_{k=1}^{m} t_i^k = 1$, and $\sum_{k=1}^{m} I_i^k = 1$.

Affection of trust distribution in the population is modeled using multiagent system. The model is hierarchical and covers four sets of subjects. The first set is called Consumers, second Producers, next Analyzer (set of one or more agents), and the last is Dominator (one agent). Dominator is the highest element in the hierarchical structure, has the control function of the whole intervention process, sets the input parameters, and evaluates the impact of the intervention. Analyzer and Producer represent the next lower hierarchic level. Intervention is realized through chosen producers (authorized by Dominator) on the whole set of the consumers or its subset. Analyzer is advisory service agent, which requests and collects data on trust changes of the consumers, analyzes the intervention process, and sends the results to Dominator. Consumer is the lowest element in hierarchy that is able to change his phenomenal trust distribution to products depending on intervention, and sends the messages about trust changes to Analyzer.

5 Case Study

To demonstrate trust evolution under affection, we used data obtained from the reports on the portal websites of the Institute of Sociology of the Academy of Sciences of the Czech Republic [4]. The data deal with an opinion on effect of economic depression (Economic crisis by Czech public eyesight published in April and May 2009). The respondents answered the question: "Tell us, please, what are your view of economic crisis effect and its impact on your personal situation?" The answers of the respondents are written down in Table 1.

Table 1 Opinion on Effect of Economic Crisis from April to May 2009 [%]

ANSWER	04/2009	05/2009
DECIDEDLY AFFECTION	25	29
SUSPICIOUSLY AFFECTION	37	40
SUSPICIOUSLY NO AFFECTION	23	21
DECIDEDLY NO AFFECTION	4	4
DON'T KNOW	11	6

In the model, the phenomenon values, i.e. products, are respondent's individual views on crisis having five values.

For simplicity, data are reduced into three following phenomenon values - first two answers in “affects”, second two ones in “doesn’t affect”, and the last one in “don’t know” (Table 2). As the initial probability distribution, the values from April 2009 and as the final probability distribution the values of trust from May 2009 are taken for the study introduced below (Subsection 4.1). The dominant product value in the population is “affects” in both cases, what is obvious from data in Table 2.

Table 2 Phenomena Values from April to May 2009 [%]

PHENOMENA VALUE	04/2009	05/2009
AFFECTS	62	69
DOESN’T AFFECT	27	25
DON’T KNOW	11	6

5.1 Study of Population Trust Dependence on Intervention Intensity

Several studies of model behavior have been performed. Presented study is accomplished for all products. Number of consumers, i.e. n = 1000, is chosen proportionally to the number of respondents which was 1038. The data acquired in April 2009 are used at the beginning of trust development simulation and the situation in May 2009 should be reached by intervention. This means that in the beginning 620 individuals dominantly believe the crisis affected their personal situation, 270 individual dominantly believe the crisis does not affected their personal situation a 110 individuals do not have dominant opinion.
The values of dominant trust of individuals’ t_d^s were generated with approximately normal probability distribution using mean value 0.5 and deviation 0.05 keeping $t_d^s > 1/3$ ($m=3$). Values of intervention probability distribution for intended increase of population trust into dominant product are $I_{\text{AFFECTS}} = 0.7$, $I_{\text{DOESN'T AFFECT}} = 0.15$, and $I_{\text{DON'T KNOW}} = 0.15$ in this study. The population trust dependence on intervention intensity λ produced by the simulation is shown in the Figure 2.

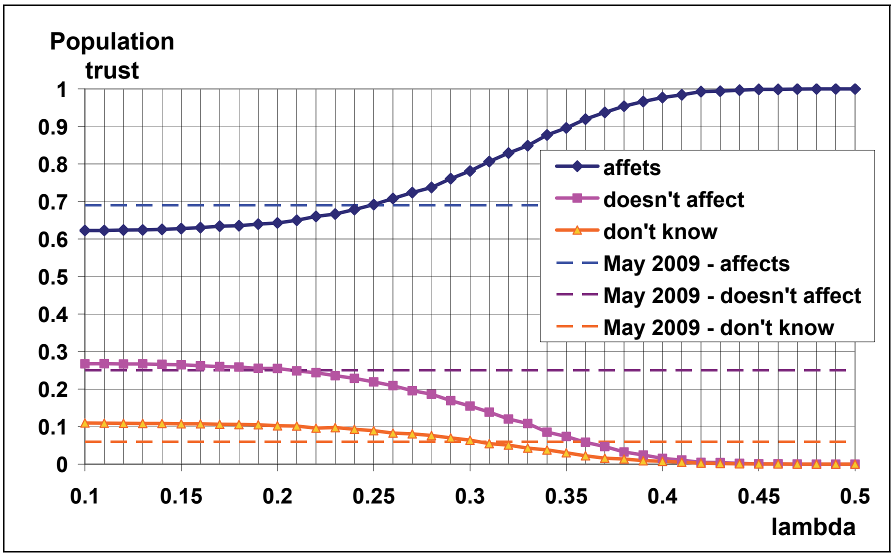


Figure 2 Dependence of Population Trust on Intervention Intensity λ

In the graph, a curve connects the computed discrete values denoted by different marks. The values acquired from the questionnaire are depicted by dashed lines. Value of intervention intensity λ needed to reach final population trust distribution determined by simulation is $\lambda = 0.25$. The maximum increase rate of population trust is for $\lambda \in \langle 0.2; 0.4 \rangle$. The graph can be utilized e.g. to forecast population trust evolution if similar situation in the future occurs.
The values for other products gained by the simulation are 22% for "doesn't affect" and 9% for "don't know". The values acquired by the questionnaire were 25% for the first one and 6% for the second one. The fact that caused these divergences is the uniform distribution trust probability between two reminded products in the model and this lead to slight undervaluation and overvaluation of trust into these products.

5.2 Model Results Verification

To validate trust intervention simulation results, we computed intervention intensity λ using formula (3) from the data of the questionnaire. The computed results are shown in Table 3. Thus, computed value of intervention intensity ($\lambda_{COMP} \approx 0.27$) is close to that gained by the simulation. Small dissimilarity can be explained by the fact that initial in the simulation, population trust was generated randomly.
In addition to intervention intensity computation, the values of probability distribution entropies $H(T_{\mathrm{APRIL}})$, $H(I)$, and $H(T_{\mathrm{MAY}})$are shown in Table 3. These entropies are computed by classic information entropy formula. Entropy decreased from April to May, i.e. from value 1.2879 to 1.1129 one, what indicates that trust probability distribution among the products in April is more uniform than in May. Indeed, population trust into dominant product grew in May and population trust into other products dropped.

Table 3 Computed Results: Data of Questionnaire (from April to May 2009)

PHENOMENA VALUE	T_{APRIL}	I	T_{MAY}	$H(T_{\mathrm{APRIL}})$	$H(I)$	$H(T_{\mathrm{MAY}})$
AFFECTS	0.62	0.70	0.69	0.4276	0.3602	0.3694
DOESN'T AFFECT	0.27	0.15	0.25	0.5100	0.4105	0.5000
DON'T KNOW	0.11	0.15	0.06	0.3503	0.4105	0.2435
Σ	1.00	1.00	1.00	**1.2879**	1.1813	**1.1129**
$\boldsymbol{\lambda_{\mathrm{COMP}} = 0{,}2679}$						

6 Conclusion

We developed the trust affection model for trust evolution. The model itself was deployed in the agent based trust management model. We demonstrated its exploitation for real data from social domain. The model confirmed expected sociologic behavior, moreover some its aspects can be quantified.

Acknowledgement

The work was granted by the Ministry of Education, Youth and Sport of the Czech Republic - University spec. research – 1311.

References

[1.] FETZER, S.: *The World Book Dictionary*. The World Book Encyclopedia, World Book Inc., Chicago, 1988.

[2.] FUKUYAMA, F: *Trust: The social virtues and the creation of prosperity*. The Free Press, New York, 1995.

[3.] GAMBETTA, D.: *Can We Trust Trust?* In Gambetta, Diego (ed.) Trust: Making and Breaking Cooperative Relations, electronic edition. Department of Sociology, University of Oxford, chapter 13, 213-237, 2000.

[4.] Institute of Sociology of the Academy of Sciences of the Czech Republic. 2009. Available at: http://www.soc.cas.cz/articles/cz/20011/Tiskove-zpravy.html, [Cit. 2009-06-26]

[5.] LIFEN, L.: *Trust Derivation and Recommendation Management in a Trust Model*. In: Proc. International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 219-222, Harbin, 2008.

[6.] LIU, Y., YAU, S., PENG, D., YIN, Y.: *A Flexible Trust Model for Distributed Service Infrastructures*. In Proceedings of the 2008 11th IEEE Symposium on Object Oriented Real-Time Distributed Computing, Orlando, 108-115, 2008.

[7.] LUHMANN, N.: *Trust and Power*. New York, John Wiley, 1979.

[8.] MUI, L.: *Computational Model of Trust and Reputation: Agents, Evolutionary Games, and Social Networks*. Ph.D. Thesis, Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2002.

[9.] SATHIYAMOORTHY, E., IYENGER, S., RAMACHANDRAN, V.: *Agent Based Trust Management Framework in Distributed E-Business Environment*, International Journal of Computer Sciences & Information Technology (IJCSIT), vol. 2, No. 1, 2010.

Current address

Arnostka Netrvalova, Ing, Ph.D.
University of West Bohemia, Faculty of Applied Sciences,
Department of Computer Science, Univerzitni 22,
306 14 Plzen, Czech Republic
Direct Tel. +420-37-763-2425
e-mail: netrvalo@kiv.zcu.cz

Jiri Safarik, prof., Ing., CSc.
University of West Bohemia, Faculty of Applied Sciences,
Department of Computer Science, Univerzitni 22,
306 14 Plzen, Czech Republic
Direct Tel. +420-37-763-2439
e-mail: safarikj@kiv.zcu.cz

Journal of Applied Mathematics

volume IV (2011), number I

DOUBLE SYSTEM PARTS OPTIMIZATION

PELIKÁN Jan, (CZ), HENZLER Jiří, (CZ), VALENTOVÁ Eva, (CZ)

Abstract. An optimization model proposed in the article is solving the question of the reserves for the functional components-parts of mechanism in order to increase its reliability. It presupposes the knowledge of the probabilities of these parts' failure and the estimation of the losses caused by this failure. The model is the problem with 0-1 variables. The solution of the model divide the parts into those which are to be doubled and those which are not. The model is illustrated by an example.

Key words and phrases. Doubling of system parts, optimization model, probability of failure, mean value of losses, system reliability

Mathematics Subject Classification. Primary 90B02, 93E02; Secondary 49N02.

1 Introduction

Large and complex mechanisms are composed of a great number of components, aggregates, partial machineries. Each of the parts is responsible for the right functioning of the whole system and vice versa, each part's failure can disturb the system or completely put it out of operation and cause damages in its effect.

One of the possibility of eliminating or at least diminishing these damages is the doubling of some important parts. Having these parts doubled, there is a possibility to exchange immediately a non-functioning part by a functioning one (or in other words, the failure is reduced only to a necessary time of a switch-over).

On the other hand, when the part is not doubled, it has to be removed from the system and then replaced by a new one. The example can be a power-distribution network composed of the electric line, switches, fuses, transformers and other parts. If, for example, a transformer fails out, the consumers depending on this particular transformer are without the power supply for the certain time and the losses as an effect are obvious.

This period depends on the time of removing the transformer and replacing it by another one. If there is, at the same location, another transformer as a reserve, then the period of switching over the reserve is much shorter than the period of the transformer's replacement.

Similar problem can emerge in the projects of a regulating system or a communication network and so on.

On the other side there are costs of doubling, that is the price of a doubled part. For that reason not all of the parts can be doubled, especially the expensive ones and also those whose failure is not bringing so expensive damages.

2 Double system parts optimization model

If we want to know which parts are to be doubled, the following optimization model can be used. First we introduce the presumption of the model.

Let us suppose n parts of the system (aggregates, components) $Z_1, Z_2, \ldots, Z_n$. Each of these parts is characterized by:

p_i probability of the failure-free run of Z_i without a reserve,

$\overline{p}_i$ probability of the failure-free run of Z_i with a reserve,

q_i the mean value of losses caused by Z_i's disorders without a reserve,

$\overline{q}_i$ the mean value of losses caused by Z_i's disorders with a reserve,

c_i costs of the purchase and maintenance of the reserve for Z_i.

Obviously:

$$p_i \leq \overline{p}_i \quad \text{and} \quad q_i \geq \overline{q}_i. \tag{1}$$

Next we suppose:

- a statistical independence of the failures of those parts,
- the costs of the parts' doubling are limited to the amount K.

Let us introduce 0-1 variables $x_1, x_2, \ldots, x_n$, the variable x_i involves the decision between the doubling of Z_i ($x_i = 1$) or not ($x_i = 0$). Total costs of the reserves for the parts are $\sum_{i=1}^{n} c_i x_i$ and since the resources for reserves are limited by K, so it has to be valid

$$\sum_{i=1}^{n} c_i x_i \leq K. \tag{2}$$

On that conditions we can:

a) maximize the reliability of the system, that is failure-free run,

b) minimize the mean value of the sum of losses caused by the part's disorders.

In the case a) the probability of the failure-free state of the system is the product of the probabilities of the failure-free states of all the parts.

The part Z_i will be failure-free with the probability $\overline{p}_i$, if it has a reserve ($x_i = 1$). If the part Z_i is without reserve ($x_i = 0$) then the failure-free probability is p_i. Altogether the probability of the part Z_i's failure-free state can be put in the form $p_i + (\overline{p}_i - p_i)x_i$. Hence the total probability of the failure-free state is

$$\prod = \prod_{i=1}^{n} [p_i + (\overline{p}_i - p_i)x_i]. \tag{3}$$

After logarithming in order to make the objective function linear we get the objective function in the form

$$z(x) = \log(\Pi) = \log[p_i + (\overline{p}_i - p_i)x_i]. \tag{4}$$

This function will be maximized.

Since the expression $\log[p_i + (\overline{p}_i - p_i)x_i]$ for $x_i = 0$ equals $\log(p_i)$ and for $x_i = 1$ equals $log(\overline{p}_i)$, we can write the expression $\log[p_i + (\overline{p}_i - p_i)x_i]$ in the form $(1 - x_i)\log(p_i) + x_i \log(\overline{p}_i)$.

The function

$$\begin{aligned} z(x) = \sum_{i=1}^{n} [(1 - x_i)\log(p_i) + x_i \log(\overline{p}_i)] &= \sum_{i=1}^{n} [\log(p_i) + x_i \log(\overline{p}_i/p_i)] = \\ &= \sum_{i=1}^{n} \log(p_i) + \sum_{i=1}^{n} x_i \log(\overline{p}_i/p_i) \end{aligned} \tag{5}$$

expresses the logarithm of the whole system's reliability.

Maximizing reliability model is:

$$z(x) = \sum_{i=1}^{n} [\log(p_i) + x_i \log(\overline{p}_i/pi)], \qquad \sum_{i=1}^{n} c_i x_i \leq K, \quad x_i \in \{0, 1\}, \quad i = 1, 2, ..., n. \tag{6}$$

In the case b) the mean value of the losses caused by the part Z_i's failure is without the reserve q_i and with the reserve $\overline{q}_i$.

The mean value of the total losses is then:

$$z(x) = \sum_{i=1}^{n} [(1 - x_i)q_i + x_i\overline{q}_i] = \sum_{i=1}^{n} [q_i - x_i \Delta q_i], \qquad \text{where } q_i = q_i - \overline{q}_i. \tag{7}$$

This function will be minimized.

The minimal losses model is:

$$z(x) = \sum_{i=1}^{n} [q_i - x_i \Delta q_i] \longrightarrow \min, \qquad \sum_{i=1}^{n} c_i x_i \leq K, \quad x_i \in \{0, 1\}, \quad i = 1, 2, ..., n. \tag{8}$$

3 Two models

Now, two models can be distinguished. First one is a one-case situation, when the function operates for a short period - static case. The second one is a long-time model for a longer time period - dynamic model.

3.1 Static model

The probability of the components' failure Z_i will be denoted as Π_i. Consequently the probability of the failure-free run of the part without reserve is $p_i = 1 - \pi_i$ and the probability of the failure-free run of the part with a reserve is $\overline{p}_i = 1 - \pi_i^2$.

If the losses caused by one failure of the part Z_i is denoted by Q_i, the mean value of the losses is:

$q_i = \pi_i Q_i$ in case when there is no reserve,

$\overline{q_i} = \pi_i^2 Q_i$ in case when there is a reserve for the part Z_i.

3.1.1 Example

Let us have parts Z_1, Z_2, Z_3, Z_4, Z_5, table 1 contains the main characteristics of those parts.

The costs of the parts' doubling are limited by the amount $K = 100$.

	Z_1	Z_2	Z_3	Z_4	Z_5
p_i	0.9	0.8	0.9	0.93	0.91
$\pi_i = 1 - p_i$	0.1	0.2	0.1	0.07	0.09
π_i^2	0.01	0.04	0.01	0.0049	0.0081
$\overline{p}_i = 1 - \pi_i^2$	0.99	0.96	0.99	0.9951	0.9919
c_i	80	30	35	50	3.20
Q_i	1666	250	333	1613	989
$q_i = Q_i \pi_i$	166	50	33	113	89
$\overline{q}_i = Q_i \pi_i^2$	16.6	10	3.3	8	8
$\Delta q_i = q_i - \overline{q}_i$	150	40	30	105	81

Table 1: Characteristics of the model

Reliability model which maximizes the failure-free probability is:

$$\log(z(x)) = \log(0.548402) + x_1 \log(0.99/0.9) + x_2 \log(0.96/0.8) + \\ + x_3 \log(0.99/0.9) + x_4 \log(0.9951/0.93) + x_5 \log(0.91/0.9919) \longrightarrow \max, \tag{9}$$

$$80x_1 + 30x_2 + 35x_3 + 50x_4 + 20x_5 \leq 100, \tag{10}$$

$$x_1, x_2, x_3, x_4, x_5 \in \{0, 1\}. \tag{11}$$

By using standard software LINGO we get the optimal solution $x = (0, 1, 1, 0, 1)$ with the failure-free probability equal 0.789041, which is maximal. From the result follows that it has to double Z_2, Z_3, Z_5.

Model which minimizes the mean value of the total losses is:

$$z(x) = 451.9 - 150x_1 - 40x_2 - 30x_3 - 105x_4 - 81x_5 \longrightarrow \min, \tag{12}$$

$$80x_1 + 30x_2 + 35x_3 + 50x_4 + 20x_5 \leq 100, \tag{13}$$

$$x_1, x_2, x_3, x_4, x_5 \in \{0, 1\}. \tag{14}$$

solution x	reliability	losses	doubling cost
(0,0,0,0,0)	0.578402	451.9	0
(0,1,1,0,1)	• 0.789041	300.3	85
(1,0,0,0,1)	0.657534	• 220.6	100
(1,1,1,1,1)	0.934507	45.9	215

Table 2: Different solutions

When we use again LINGO system, we get the optimal solution $x = (1, 0, 0, 0, 1)$ with the minimal value of losses 220.9 in the mean value. According to this solution only the parts Z_1 and Z_5 will be doubled. The differences in the solutions obtained above we can explain by great influence of the amount of the losses in the optimal solution in the second model. First solution $x = (0, 1, 1, 0, 1)$ means the most reliable system, but the losses are not minimal. Second solution $x = (1, 0, 0, 0, 1)$ gives us less reliable system, but the losses are minimal.

We can observe the values of reliability, mean losses and costs of reserves for diferent solutions in the table 2. The values with a bullet are optimal for $K = 100$.

3.2 Dynamic model

Let us suppose that the system's reliability should be optimized within a period $< 0, T >$ and the periods between the failures of the parts are exponentially distributed. Let the mean value of the period between two failures of the component Z_i be $1/\lambda_i$. The period of replacing the failed component Z_i by new one is fixed number t_i.

If the Z_i is without a reserve, then the probability of the failure-free run is $p_i = \exp(-\lambda_i T)$.

Since the number of the failures within the period $< 0, T >$ is Poisson distributed with the mean value $\lambda_i T$, the mean value of the losses in the case $x_i = 0$ is equal to $q_i = \lambda_i T Q_i$, where the Q_i are the losses caused by the part Z_i's failure.

In case that component Z_i has a reserve ($x_i = 1$), the t_i is period necessary for the reserve components' exchange, where $t_i << T$. During the period t_i the failure-free probability is $\exp(-it_i)$.

Since the mean value of the number of the component Z_i's failures within $< 0, T >$ is $\lambda_i T$, hence we get the formula describing the reliability in the form

$$\overline{p}_i = \exp(-\lambda_i T) + (1 - \exp(-\lambda_i T)) \exp(-t_i \lambda_i^2 T). \tag{15}$$

When the part is doubled, the number of failures is Poisson distributed and the mean value of the number of failures is λ_i times the period, which is approximately $(\lambda_i T)t_i$. Hence the mean value of losses is

$$\overline{q}_i = \lambda_i^2 T t_i Q_i. \tag{16}$$

In the formulas for $\overline{p}_i$ and $\overline{q}_i$ was not taken in account:

- reduction the time period T of fail parts' exchance time,
- the case when both original and reserve parts fail at the same time.

Acknowledgement

The research was supported by grants no. GAČR 402/09/0041 and no. GAČR 402/09/0405 of the Grant Agency of the Czech Republic.

References

[1] NEMHAUSER, G.L., WOLSEY, L.A.: *Integer and Combinatorical Optimization.* J.Wiley, New York, 1988.

Current addresses

prof. RNDr. Jan Pelikán, CSc.
University of Economics Prague, Departement of Econometrics, nám W. Churchilla 4, 130 67, Praha 3, Czech Republic, tel. no. +420 224 095 444, $e-mail:$ *pelikan@vse.cz*

doc. RNDr. Jiří Henzler, CSc.
University of Economics Prague, Departement of Mathematics, Ekonomická 957, 148 00, Praha 4, Czech Republic, tel. no. +420 224 094 246, $e-mail:$ *henzler@vse.cz*

RNDr. Eva Valentová
University of Economics Prague, Departement of Mathematics, Ekonomická 957, 148 00, Praha 4, Czech Republic, tel. no. +420 224 094 243 , $e-mail:$ *evalenta@vse.cz*

Journal of Applied Mathematics

volume IV (2011), number I

A MODEL FOR THE PORTUGUESE TOURISM MARKET

SILVA, Albino, J. (P), FERREIRA, Manuel Alberto M. (P), AMARAL, J. Ferreira (P), FILIPE, José António (P), COELHO, M. (P)

Abstract: The discussion of tourism problems is classical and many developments have brought tourism models to the actual debate. In a moment when global crisis is painful for many economies mainly in many developed countries, this topic is now particularly relevant. In this study, factors explaining demand, supply and prices are discussed. At macroeconomic level, it is seen how they contribute to model the Portuguese tourism market. A relationship among the variables is analyzed and its modeling is represented mathematically. The model allows us to conclude about the contribution of this kind of model to show the importance of these variables and relationships to the determination of the macroeconomic aggregates in the Portuguese tourism market.

Keywords: Tourism, Portuguese Market, Tour Operator, Tourism Demand, Tourism Supply

Introduction

The tourism in Portugal is very important in terms of macroeconomic aggregates. And very particularly, there are several Portuguese regions in which tourism is the main contributor sector for Gross National Product.

A Report of World Tourism Organization (UNWTO, 2008) showed how tourism is important as economic activity in Portugal. According to this Report, Portugal is one of the 20 most visited countries. In that year, more than 12 million people visited Portugal, more than the Portuguese population. In 2008, according to INE preliminary data (Instituto Nacional de Estatística, 2009), tourism has generated about 5% of GVA of Economy, which is approximately 7.3 billion Euros. According to the Report on Competitiveness of Travel and Tourism, 2008 (World Economic Forum (2008)), Portugal occupied the 15th place that year, in a list of 130 countries in the ranking of competitiveness of the tourism sector. Overall climbed seven positions compared to 2007 and four positions in all 27 EU countries (Portugal Digital, 2008). Amador and Cabral (2009) present a detailed analysis of the evolution of the services sector in Portugal and show that this favorable trend is verified in this sector in general and in particular that Portugal reveals a comparative advantage in the sector of Travel and Tourism.

These facts explain why a model like the one developed in this study for tourism market in Portugal is so relevant. The model implemented considers a set of variables and relationships that will be explained in the next chapter.

Main Features of the Model and Theoretical Concepts

The model main features for the Portuguese tourism market are the following aspects:

- The study of the market of the tour operator (travel agency);
- A strong support on microeconomic theory;
- It is supported on 3 items: demand, supply and prices;
- It permits to obtain results associated to exogenous variables and to tourism rents.

In fact, the macroeconomic building of a model like this one is supported on a set of suppositions around the tourism product, economic agents and the several different types of existing markets.

Tourism Product

Tourism Product is the trip plan of a tourist, self-made or not, being the overnight staying/revenue per bed the variable to be quantified. If a tourist is not national, this product is obtained in the market buying a “package” supplied by the tour operator. Taking into account that plans and programs are well diversified, it is considered that the variable “overnight staying” is good enough to be an approximation to the Homogeneous Product.

Economic Agents

In the Portuguese tourism market it is relevant to consider three economic agents:

- The tour operator/travel agency
- The accommodation company/hosting company (hotels, hostels, etc)
- The tourist.

Tour Operator

Headquartered in the emitters, the tour operator is responsible for a big part of tourists that come to Portugal from abroad.

The tour operator is the main agent on this market. Tour operators are near the tourist and have lots of possible destinations available to send tourists to. Besides, several tour agents have a significant part of market share and this is quite relevant to be considered in the analysis.

Their positioning in the market and their economic capacity allow them to be very accurately vigilant to the features of the markets and to be aware of changes in the habits and economic conditions of the tourists.

This shows how important is their great bargaining power among firms that offer services in the country, especially in companies for accommodation (hotels and similar).

It is particularly impressive the power of intervention on prices and the possibility of withdrawal of accommodation services if they do not have enough demand for the programs they offer. The costs of this operator will thus be closely linked to the acquisition of services to include in the package.

Company for accommodation / hosting company

The hosting company emerges as the main supporting infrastructure for tourists. Its short-term, goals are to increase occupancy and to improve the revenue per bed. The tour operator is its main interlocutor, that makes it a price taker (can' not enforce the rack rate, which is the price that would maximize its utility function).

The company has significant fixed costs. The "variable costs" are the variable that has a direct and proportional connection with "overnight stays".

The supply of accommodation is in turn a function of the stock of fixed capital, which combined with the staff, ensures the service to be provided to tourists.

Tourist

Moving from his/her place of habitual residence, whether abroad or in Portugal, for a period exceeding 24 hours, the tourist accommodation is the main support infrastructure in place he/she is visiting.

The tourist is a consumer who seeks the tourism product because he needs it and because he has financial capacity to acquire it.

This "good" or tourism product results from a concrete plan that tourists draw up, or results from one plan accepted by the tourist that is drawn up by the operator. It includes the location or destination (possibly several), and the activities he wants to accomplish and yet the cost travel.

It is considered that tourists from abroad primarily use the services of tour operators ('packages' or holiday programs). National tourists, in turn, project (produce) their holiday program, which leads them to directly contact the housing companies.

Tour operator, hosting company and tourist represent the major economic players in the Portuguese tourism market. Each one of them wants to maximize its respective utility function and will have its own restrictions. The tour operator may not sell more "packages" than the acquired overnights. The hosting company has the housing capacity as the main obstacle to an increased supply of overnight stays. Tourists will naturally be dependent on either their needs or their ability to purchase travel.

The Design of the Model and Theoretical Discussion

Some blocks for the model

In this model of the tourism market in Portugal there are three blocks: the demand, supply and prices.

Demand

It is considered that tourism demand is explained by exogenous variables, by policy and by prices.

Demand functions are developed for national tourists, for foreign tourists by emitter countries, and a global demand function.

Demand function for foreign visitors by country

For emitter countries of foreign tourists in Portugal (Spain, France, Holland, Germany and UK are the main tourists' suppliers for the Portuguese market), the main determinants of overnights (variable to be explained) relate to the following variables:

- non-essential consumption of households in the emitters;
- Purchasing power of the currency of the emitters;
- Price of accommodation in Portugal corrected by the exchange rate of the emitter country, if it is the case;
- Price of goods and services in Portugal weighted by competition from other countries of destination;
- Overnights spent in the previous year,
- Time (trend effect).

For each of these variables the following explaining factors are listed:

a) Purchasing power of the currency of the emitters ($PCME$);

This variable relies on the assumption that tourists are sensitive to changes in exchange rates and price increases in tourist destinations, compared with prices at the place of residence or in the emitter country. Its calculation is derived from the knowledge of inflation rates in the country of destination adjusted by the existing ones in the country of origin. This will deflate the exchange rates, showing how much the holiday in a tourist destination costs in real terms, compared with its cost at the place of residence.

b) non-essential consumption of households in the emitters ($CPNE$)

For establishing the values of this variable was considered as non-essential consumption the one that does not include expenditure on (statistical aggregates):

- Food, beverages and tobacco;
- Clothing and Footwear;
- Housing;
- Medicines and medical care.

These items appear in the National Accounts as part of final consumption of households on the economic territory.

This represents the hypothesis that the families will travel in tourism after the guaranteed expenditure on goods and basic services.

c) the price of accommodation in Portugal adjusted by the exchange rate of the emitter country ($PHTC$)

This variable is considered as associated with the average price of "packages" sold in the UK, weighted by the exchange rate of each emitter country (if considering just the main non-euro emitter country).

d) Price of goods and services (IPC) in Portugal weighted by the prices of goods and services (IPC) of competing countries.

The purpose of this variable is not only to measure the relationship between the prices of goods and services between sender and receiver country weighted by their exchange rate, but also to include the competitiveness of other destinations comparatively to Portugal, via weights which result from the market share for these countries (destinations) to the main flows of tourists travelling to Portugal.

e) Overnight stays in the previous year (D_{-1})

The purpose of a visit that took place recently, either to repeat or to inform others, influences the flow of tourists away.

f) Trend effect (T)

Broadening the scope of the previous variable, this effect has structural characteristics that, in tourism, in many situations can' not be neglected (habits, knowledge to be extended or the pleasure of seeing landscapes and climates can be explained by the trend).

Mathematically tourism demand for the emitter country can be expressed as follows:

$$DEST = D_1 + D_2 = f(CPNE, PCME, PHTC, PR, DEST(-1), T, DU)$$

Being:

$DEST$ - a measure of demand for Portuguese tourism services, here represented by the number of overnight stays of foreigners (for the emitter country) annually considered;

$PCME$ - purchasing power of the currency of the emitters;

$CPNE$ - non-essential consumption of households in the emitters;

$PHTC$ - the price of accommodation in Portugal adjusted by the exchange rate of the emitter country;

PR - the price of goods and services in Portugal weighted by the prices of goods and services by major competitors in tourism;

$DEST(-1)$ - a measure of demand, represented by the number of overnight stays of foreigners in the previous period;

T - the tendency or effect "trend-time";

DU - the "DUMMY" variable.

It is expected that variables $PCME$, $CPNE$, $DEST(-1)$ and T are directly related with dependent variable $DEST$, this is $\frac{\partial(DEST)}{\partial(PCME)} > 0$, $\frac{\partial(DEST)}{\partial(CPNE)} > 0$, $\frac{\partial(DEST)}{\partial(DEST(-1))} > 0$ and $\frac{\partial(DEST)}{\partial(T)} > 0$. Otherwise, it is expected that variables $PHTC$ and PR are inversely related to dependent variable $DEST$, as follows: $\frac{\partial(DEST)}{\partial(PHTC)} < 0$ and $\frac{\partial(DEST)}{\partial(PR)} < 0$.

To estimate the impact of changes in the variables in tourism demand, the coefficients associated with this impact need to be estimated, and the previous equation be expressed in the following way:

$$DEST = a_0 CPNE^{a_1} PCME^{a_2} PHTC^{a_3} PR^{a_4} [DEST(-1)]^{a_5} T^{a_6} E$$

Using the method of least squares regression, and after a logarithm, this equation can be estimated as presented below, introducing DUMMY variable (just now for reasons of mathematical operating of the model):

$$\ln DEST = \ln a_o + a_1 \ln CPNE + a_2 \ln PCME + a_3 \ln PHTC + a_4 \ln PR + a_5 \ln DEST(-1) + a_6 \ln T + \lambda DU + \ln E$$

So it is allowed the coefficients to be interpreted as the elasticity of the dependent variable for each of the explanatory variables.

Demand global function for foreign visitors

The explanatory variables in global demand for foreign visitors presented in this model are:

• Private consumption in EU countries;

• Housing Prices in Euro;

• Purchasing power of the Euro (once United Kingdom is an important issuing country - one of the five most important - it is used the relationship Euro *versus* British Pound); and just considered the exchange rate Euro / Pound).

The approach made to the variables used in the demand function by the issuing country has some limitations, which are related to the failure to achieve overall series for those variables and not considering appropriate to proceed to a mere sum of the grades obtained for each issuing country.

In its algebraic presentation, this function comes with the following form:

$$DEG = f(CP, PHEURO, PCEURO)$$

And
DEG is a measure of demand for Portuguese tourism services, represented by the number of overnight stays of foreigners;
CP is Private consumption in EU countries.
$PHEURO$ is the housing price in Euros, running the Euro as reference currency for all emitting countries (excluding UK; it will be considered its exchange rate with the pound sterling).
$PCEURO$ is the purchasing power of the Euro in relation to the pound.

It will be used the same method of the previous equation.

Demand function for the National Tourists

To explain the nights spent by Portuguese Tourists in national hotels (D_p), the variable private consumption (CP) and trend (T) are considered as the major determinants.

Algebraically this function will be:

$$D_p = f(CP, T)$$

It will be used the same method of the previous functions.

Supply

Function of Tourism Production

The tourism production consists of the nights spent, considering the maximum capacity of companies. The potential production will result from considering the effective overnight bed occupancy.

This potential production is correlated with investment which changes will be reflected in the stock of capital, and with staff admitted in a situation of rigidity.

Moreover, it is admitted the existence of constant returns on scale, with a Cobb-Douglas function, and it will be established a relationship between the average productivity of labour and the coefficient of capitalistic intensity, emerging the capital as an explanatory variable.

Thus

$$X = f(L, K)$$

And

X is the potential tourism production that is represented by tourists' overnights;

L is the labour factor, given by the average of workers in service during the year;

K is the effective capital stock.

So, not forgetting the trend,

$$X = AL^{1-a}K^{a}e^{T^{t}}$$

or

$$\frac{X}{L} = A(\frac{K}{L})^{a}e^{T^{t}}$$

after the application of logarithms, it comes

$$\ln(\frac{K}{L}) = \ln A + a\ln(\frac{K}{L}) + T^{t}$$

In this model, tourism production can be compared to the tourism demand.

Prices

It is considered that the price will influence demand, through supply side.

At first, we have the price of the "package" offered by tour operator (linked to exogenous variables such as inflation in the emitters) to establish the international demand for tour "packages". This price also influence the supply of housing.

In turn, the hosting company submit a higher price (the price of "counter") for taking up excess capacity. For the definition of this price, this company will carry out an analysis of the costs.

An estimate of the cost-function in the hosting company will assume the dependent variable (variable costs) will be explained by overnights and by the prices of goods and services purchased by this company (intermediate consumption).

Formalization

The Demand

Be

$$D = D_1 + D_2 + D_3$$

Where

D is the total number of overnight stays in housing companies;

D_1 is the number of nights spent by foreigners, in group;

D_2 is the number of nights spent by foreigners, individually;

D_3 is the number of nights spent by nationals;

$D_1 + D_2$ is the total number of nights spent by foreigners in the accommodation companies.

$D_1 + D_2 = DEST = f(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)$

and

Y_1 - Nonessential consumption of foreign tourists

Y_2 - Purchasing power of the currency of the country of residence of foreign tourist

Y_3 - Price of the "package" for Portugal weighted by the exchange rate

Y_4 - Price of goods and services in Portugal weighted by the price of goods and services from competing countries of Portugal

Y_5 - Nights of the previous year

Y_6 - Trend

$D_3 = f(Y_7)$

Being

Y_7 the private consumption in Portugal.

Supply

Be

$$X = f(L, K)$$

In which

L is the number of people working in hosting companies;

K is the fixed capital stock for hosting companies.

And

$$C = f(D_1 + D_2 + D_3)$$

Being C the total variable costs of hosting companies.

The market and the determination of prices

Tour Operator and Its Utility Maximization (Profit)

Be

$$L = p_v q_1(p_v) - p_c q_2(p_c)$$

In which

L is the profit of the foreigner tour operator

p_v is the selling price of touristic product (D) to the tourist, sold by tour operator;

p_c is the buying price of the touristic product (D) by the tour operator to the hosting company ;

q_1 is the touristic product (D) sold by the tour operator to the foreigner tourists – the quantities are measured in terms of overnights;

q_2 is the touristic product (D) bought by foreigner tour operator to the hosting company.

For the maximization of profit, we will have the following model:

$$MaxL(p_v, p_c) = p_v q_1(p_v) - p_c q_2(p_c)$$

Sub. to

$$q_1(p_v) - q_2(p_c) \leq 0\,,$$

$$p_v\,, p_c \geq 0\,.$$

The Lagrange function is

$$Z(p_v, p_c, s, \lambda) = p_v q_1(p_v) - p_c q_2(p_c) + \lambda(-q_1(p_v) + q_2(p_c) - s)$$

Being s an auxiliary variable and λ the Lagrange multiplier.

After considering Kuhn-Tucker conditions, and if $\lambda = 0$:

$$q_1 = A\frac{1}{p_v}, A \in R$$

$$q_2 = B\frac{1}{p_c}, B \in R$$

Once

$$q_1(p_v) + p_v \frac{\partial q_1}{\partial p_v} = 0 \ ,$$

$$q_2(p_c) + p_c \frac{\partial q_{12}}{\partial p_c} = 0 \text{ and}$$

$-q_1(p_v) - q_2(p_c) - s = 0$, being $s \geq 0$ the auxiliary variable.

So,

$$-A\frac{1}{p_v} + B\frac{1}{p_c} - s = 0 \text{ and so}$$

$$-A\frac{1}{p_v} + B\frac{1}{p_c} \leq 0 .$$

Finally,

$$A\frac{1}{p_v} \leq B\frac{1}{p_c} \quad \text{or} \quad p_c \leq \frac{B}{A} p_v .$$

Let be $\frac{B}{A} = D$, considering $0 < D < 1$,

$p_c = Dp_v$ and consequently

$$p_c < p_v$$

Now, if $s = 0$ (the case in which the restriction equals 0),

$$Z(p_v, p_c, \lambda) = p_v q_1(p_v) - p_c q_2(p_c) - \lambda(q_1(p_v) - q_2(p_c))$$

After first order conditions,

$$\lambda = \frac{q_1(p_v)}{\frac{\partial q_1}{\partial p_v}} + p_v$$

$$\lambda = \frac{q_2(p_c)}{\frac{\partial q_2}{\partial p_c}} + p_c$$

And so,

$$\frac{q_1(p_v)}{\frac{\partial q_1}{\partial p_v}} + p_v = \frac{q_2(p_c)}{\frac{\partial q_2}{\partial p_c}} + p_c\,,$$

Resulting, after some operations,

$$\frac{p_v}{p_c} = \frac{1+\frac{1}{e_{q_2}}}{1+\frac{1}{e_{q_1}}}$$

If p_v is considered fixed, p_c will depend on the demand functions.

With the conditions of saturation $(q_1 = q_2)$, and $p_v > 0$ and $p_c > 0$, each elasticity will be studied through

$$\frac{p_v}{p_c} = \frac{1+\frac{1}{e_{q_2}}}{1+\frac{1}{e_{q_1}}} > 0$$

Five scenarios are possible:

- $e_{q_2} < -1$ and $e_{q_1} < -1$
- $e_{q_2} < -1$ and $e_{q_1} > 0$
- $e_{q_2} > 0$ and $e_{q_1} < -1$
- $e_{q_2} > 0$ and $e_{q_1} > 0$
- $-1 < e_{q_2} < 0$ and $-1 < e_{q_1} < 0$

Considering the first scenario ($e_{q_2} < -1$ and $e_{q_1} < -1$), $p_c < p_v$ and conditions of saturation $q_1 = q_2$,

$$\frac{\partial q_1}{\partial p_v} > \frac{\partial q_2}{\partial p_c}.$$

For the model purposes, this result shows a bigger variation rate in the market of products (overnights, bought by tourists to the tour operators after prices' changes) than in the market of factors (overnights bought by tour operators to the hosting houses).

An example can be found when $q_1 = \frac{A}{p_v} + B$. On this situation there is an inverse relationship between price (p_v) and overnights (q_1) but also an important representative variable of, for instance, the fidelity rate in relation to a destination region.

And now for e_{q_1} (the same procedure for e_{q_2}):

$$\frac{\partial q_1}{\partial p_v}\frac{p_v}{q_1} = -\frac{A}{(p_v)^2}\frac{p_v}{\frac{A}{p_v}+B} = -\frac{A}{A+p_vB} < -1\,.$$

Considering a lesser elasticity for overnights demanded by tourists and the values for each elasticity, for example:

$e_{q_1} = -1.2$ and $e_{q_2} = -1.7$,

$$\frac{p_v}{p_c} = \frac{1+\frac{1}{e_{q_2}}}{1+\frac{1}{e_{q_1}}} = \frac{0.41}{0.17} \cong 2.41$$

This means that there is a relationship between prices of about 2.41 independently of the value of each one.

Hosting Company and Its Utility Maximization (Profit)

After fixing p_v by tour operator and being p_c determined, it is now just necessary that hosting company use this price to the maximization of its profit. So,

$$\Pi = \overline{P_1}\,\overline{D_1} + P_2D_2 + \overline{P_3}\,\overline{D_3} - C(\overline{D_1} + D_2 + \overline{D_3})\,.$$

Being

$P_3 = P_2$ and $\overline{D_1} + D_2 = f(P_2)$, and so $D_2 = f(P_2) - \overline{D_1}$, in which

P_1 = Price of foreigners' overnights, in group. It is given to the company, once it is determined by tour operator.

P_2 = Price of foreigners' overnights, individual

P_3 = Price of nationals' overnights

So,

$$\Pi = P_2(f(P_2) - \overline{D_1}) + \overline{P_3}\,\overline{D_3} + \overline{P_1}\,\overline{D_1} - C(\overline{D_1} + D_2 + \overline{D_3})$$

The hosting company will maximize its profits and deplete its offer capacity, having by restriction the installed capacity.

$$Max\Pi = P_2(f(P_2) - \overline{D_1} + \overline{D_3}) + \overline{P_1}\,\overline{D_1} - C(f(P_2) + \overline{D_3})$$

Subj. to

$$\overline{D_1} + D_2 + \overline{D_3} = \overline{X}$$

(Or $D_2 = \overline{X} - \overline{D_1} - \overline{D_3}$ or yet $\overline{D_1} + D_2 + \overline{D_3} - \overline{X} = 0$).

Considering the Lagrange function and using conditions for maximization, then

$$P_2 = -\frac{\overline{X}}{\frac{\partial f}{\partial P_2}} + \frac{\overline{D_1}}{\frac{\partial f}{\partial P_2}} + \frac{\partial C}{\partial D}$$

$$P_2 = -\frac{\overline{X} - \overline{D_1}}{\frac{\partial f}{\partial P_2}} + \frac{\partial C}{\partial D}$$

Finally,

$$P_2 = -\frac{1}{\frac{\partial f}{\partial P_2}}(\overline{X} - \overline{D_1}) + \frac{\partial C}{\partial D}, \text{ being } -\frac{1}{\frac{\partial f}{\partial P_2}} = K,\ K > 0\,.$$

This means that an hypothesis for resolution can be presented having in account a first order linear differential equation with constant coefficients (see Ferreira and Amaral, 1988)[1].

Besides,

$$P_2 = -\frac{\partial P_2}{\partial D_2}(\overline{X} - \overline{D_1}) + \frac{\partial C}{\partial D}$$

or

$$\frac{\partial P_2}{\partial D_2}(\overline{X} - \overline{D_1}) + P_2 = C'$$

being

$$\frac{\partial C}{\partial D} = C'.$$

Solving the homogeneous equation $\frac{\partial P_2}{\partial D_2}(\overline{X} - \overline{D_1}) + P_2 = 0$, which solution can give us some clues about the stability of the complete solution, it is obtained:

$$P_2 = Ce^{-\frac{D_2}{(\overline{X} - \overline{D_1})}}$$

A particular solution is:

$P_2 = K$. With $K = C'$ will be

$P_2 = C'$.

The general solution:

[1] The solution is not presented in this paper, because it is very difficult to interpret it in the considered practical situation.

$$P_2 = \overline{P_2} + P_2^* = Ce^{\frac{D_2}{(\overline{X}-\overline{D_1})}} + C'$$

And

$\overline{X} = \overline{D_1} + D_2 + \overline{D_3} + D^*$, and D^* is the non-used capacity

- If $D^* = 0$ and $\overline{D_3} = 0$, it comes

$$\frac{D_2}{(\overline{X}-\overline{D_1})} = 1$$

- If $\overline{D_3} \neq 0$

$$\frac{D_2}{(\overline{X}-\overline{D_1})} < 1$$

- If $D^* \neq 0 \wedge \overline{D_3} \neq 0$, it comes

$$0 < \frac{D_2}{(\overline{X}-\overline{D_1})} \leq 1$$

Conclusion

A model for Portuguese tourism market has been made considering demand, supply and prices. The model is a mathematical formulation of Portuguese tourism and shows how these factors can contribute to explain the market and the impact on hosting rents. It also permits to make a design of the Portuguese tourism market considering tour operators.

The proposal of this model it to be workable by using the total rents of hosting companies, aiming to analyze the quantitative economic impact of these rents in the Portuguese National Product.

Such a model like the one developed represents a tool for tourism companies to improve their results and the services they provide. It has been tested successfully in Algarve, one of the most important Portuguese touristic regions.

References

[1] AMADOR, J. and CABRAL, S. (2009). “O Comércio Internacional de Serviços na Economia Portuguesa, Banco de Portugal”, *Boletim Económico*, Outono 2009, 229-249.

[2] FERREIRA, M. A. M. and AMARAL, I. (1988). “*Integrais Múltiplos. Equações Diferenciais*”. Lisboa: Edições Sílabo.

[3] Instituto Nacional de Estatística (2009). *Conta Satélite do Turismo (2007 – 2009)*, Destaque: Informação à Comunicação Social”.

[4] UNWTO (2008). “UNWTO World Tourism Barometer” , Volume 6, nº2, June 2008

[5] World Economic Forum (2008). “The Travel and Tourism Competitiveness Report 2008”, http://www.weforum.org/ttcr08browse/index.html

[6] SILVA, ALBINO J. (1991), “O Turismo Em Portugal. Uma Análise de Integração Micro-Macroeconómica”, Tese de Doutoramento não publicada, ISEG, Universidade Técnica de Lisboa.

Current addresses

João Albino M. da Silva, Professor Catedrático
Faculdade de Economia, Universidade do Algarve
Centro de Investigação em Turismo e Lazer
Campus de Gambelas , 8005-139, Faro, Portugal
Tel. +351 289817571
E-mail : jsilva@ualg.pt

Manuel Alberto M. Ferreira, Professor Catedrático
ISCTE – IUL
UNIDE – IUL
Av. Forças Armadas 1649-026 Lisboa, Portugal
Tel. +351 217 903 000
E-mail: manuel.ferreira@iscte.pt

João Ferreira Amaral, Professor Catedrático
ISEG- UTL
Rua do Quelhas, 6, 1200-781
LISBOA, Portugal
Tel. + 351 213922808
Email : famaral@iseg.utl.pt

José António Filipe, Professor Auxiliar
ISCTE – IUL
UNIDE – IUL
Av. Forças Armadas 1649-026 Lisboa, Portugal
Tel.+351 217 903 000
E-mail: jose.filipe@iscte.pt

Manuel Coelho, Professor Auxiliar
ISEG- UTL
SOCIUS
Rua do Quelhas, 6, 1200-781
LISBOA, Portugal
Phone: +(351) 213925800.
Fax: +(351) 213922808.
Email: coelho@iseg.utl.pt

Journal of Applied Mathematics

volume IV (2011), number I

THE EFFECTS OF THE PROPOSAL FOR PLANNING THE EVACUATION WITH LINEAR MODEL EXPERIMENTAL EVALUATION

TEICHMANN Dušan, (CZ), DORDA Michal, (CZ), GROSSO Alessandra, (I), IVAN Martin (CZ)

Abstract. The evaluation of the development of calculations for the optimal planning of area evacuation of citizens is presented in this article. The calculations are implemented by linear programming methods. The linear models are designed for two main criteria. The optimization criteria are the maximum evacuation time and vehicles utilization. The article contains two solution procedures that are mutually compared.

Key words. Evacuation, optimization, transport

Mathematics Subject Classification: Primary 90C05

1 Introduction

The evacuation of citizens is generally a process of moving the citizens from more dangerous places to secure places. The evacuation can proceed as a point or as the area one. During the point evacuation we evacuate objects (buildings). During the area evacuation we evacuate larger areas (regions or parts of them). In our article we will deal with public transport organization during the area evacuation.

The area evacuation consists of the following three steps:

1. the vehicles (buses) leave the established points to the evacuated communities.
2. in the evacuated communities, the citizens board the prepared vehicles.
3. the vehicles transfer the citizens to the secure places - the evacuation centers.

The area evacuation of citizens proceeds both as a single evacuation (by cars) and as a process executed by the public transport.

We will be interested in the planning of evacuation by the public transport.

We can represent the transport networks by a graph in which every edge will have an assigned value. Afterwards we have to decide, what the values will represent. In the case of evacuation the

kilometers are not important but the timing is. That is why the weight of edge will represent the time needed for passing the communication.
During the solution we have to fulfill some restrictions. We must not surpass the number of vehicles we have reserved, we have to remove all citizens from the dangerous places and we must not get over the capacity of the evacuation centers. Furthermore we want the evacuation to execute in the fastest and cheapest possible way.
Finding an optimal solution of the problems that have larger range is a difficult and laborious work. We would have to analyze all the possible solutions and follow keeping many restrictions. That is why we choose another approach. Our work will be easier thanks to the linear programming methods. It deals with planning problem the linear programming is the proper mean. Because we use the linear programming methods we can expect to get the optimal solution.
But the first step we have to do is rightly to formulate the assigned problem, to describe the decisions and expectations and the optimization criteria. In the linear programming all we said means - to determine important constants for the calculations and to create variables, domains of variables, a system of restricting conditions and the objective function.

2 Brief overview of the state of evacuation modeling knowledge

The first basic vision related to the problem of the modeling process of the evacuation was realized in the work [Southworth, 1991]. It joins together the fundamental information on the evacuation process, and advises the direction to take on its resolution.
The approaches for the mathematical modeling published in the literature can be divided in some groups.
The first group is created by pure analytical approaches. Here we can present, for example, some foreign works done by [Daganzo, 1993] and [Bish 2006]. Daganzo in his work proposes to model for the solution of the area evacuation the movement of citizens as multi-commodity traffic flows. In his approach he tries to apply the theory of the kinematic wave used in the models for the traffic flows. Bish in his thesis uses linear programming and focuses on the organization of the evacuation of the coast before a tropical tempest. It deals with the evacuation planning for case of an expected exceptional event. From the published articles which are in the scientific magazines in the Czech Republic and the Slovak Republic we can mention the analytic approaches written by [Janáček, 2009] and [Teichmann, Peško, 2010]. They deal with the area evacuation for a case of expected and unexpected exceptional events. The second group of approaches contains purely heuristic approaches. In this case we can mention the article written by [Lu, Georgie, Shekhar, 2005]. This article focuses its attention on the heuristic approach, which substitutes the linear models that the evacuation plans cannot to solve because the calculations are too vast. To the same group for example the article written by [Zeng, Wang, 2009] belongs as well.
The third group is represented by the work [Lim, Zangeneh, Baharnemati, Assavapokee, 2009], which combines the mathematical programming with the heuristic algorithm. This approach enables to determine the concrete evacuation plans, the time of starting evacuations and the recommended paths of evacuation for the concrete exceptional event. Authors deal with approach for the expected exceptional event.
The fourth group of approaches for the solution of the problem of area evacuation is the group of the simulation approaches. In this case we can mention the thesis done by [El-Sbayti, 2008]. The author simulates the road traffic in the time of the area evacuation. This simulation model enables to test the capacity of road communications for case of the area evacuation.

A separate group of works that is related to the evacuation is created by works that model the behavior of citizens during the evacuation. Here we can speak about the articles written by [Simonovic, Ahmad, 2005] and [Pel, 2007], which deal with the study of the main factors influencing the behavior of the citizens during the evacuation process. In the articles it is shown how these aspects influence the choice of the secure place and the escape paths used during the evacuation.

3 Mathematical Model

We pay our attention to the situation, when we are planning the evacuation of the citizens for the case of unexpected exceptional event. The presented article wants to continue the theme published in the article [Teichmann, Peško, 2010]. In this article a linear model with two criteria was presented for the situation of the unexpected exceptional event, simple experiments were conducted in order to validate the functionality of the proposed model.
The main goal of our work is the validation of the usability of the model in larger examples. We assume that:

- we have more places from which the vehicles will leave for the evacuation,
- it is necessary to evacuate more communities,
- more evacuation centers (EC) are available.

In our article we describe the modification of the proposed model published in the article [Teichmann, Peško, 2010]. In the case of the modified model we will choose another approach for the formulation of the objective function.
Because the foundation of the decision process is the same as in the previous case, we can also take over original formulation of the problem.
We have to define the set of vehicles I, the set of stations J, from which the vehicles leave for the evacuation, the set of the communities K, from which we will evacuate the citizens and finally the set of the evacuation centers L.
We have the information about the number of the vehicles belonging to the type $i \in I$ that are prepared for the evacuation from the station $j \in J$ - a_{ij}, the information regarding the capacity of the vehicle of the type $i \in I$ - c_i, the information regarding the number of citizens evacuated from the community $k \in K$ - b_k, the information regarding the capacity of the evacuation centers $l \in L$ - q_l, the information regarding the time that the vehicles need to go between the station $j \in J$ and the evacuated community $k \in K$ - t_{1jk}, the information regarding the time that the vehicles need to go from the evacuated community $k \in K$ to the evacuation centers $l \in L$ - t_{2kl}.
The times t_{1jk} and t_{2kl} can be calculated on the base of the proposed traffic organization during the evacuation.
During the declaration of evacuation it can happen the situation in which some vehicles will not be in the station $j \in J$. So we can increase the time t_{1jk} by a value corresponding to the time necessary to move the vehicles to the station $j \in J$.
Then the time needed for preparing the vehicles in the station from which vehicles will leave is added into the time t_{1jk} and the time needed for the entry of the citizens inside the vehicle is added into the time t_{2kl}.
Until now all the values we talked about were the constants.

In further step we will add variables modeling the particular decisions.
For the modeling of the particular decisions in the solved problem we utilize the following variables:

- x_{ijk} - a nonnegative integer variable – is the number of the vehicles of the type $i \in I$ going from the station $j \in J$ into the evacuated community $k \in K$,
- y_{ikl} - a nonnegative integer variable – is the number of the vehicles of the type $i \in I$ going from the evacuated community $k \in K$ into the evacuation center $l \in L$,
- w_{jk} - it is the bivalent variable modeling the decisions about the routing of the vehicles from the station $j \in J$ into the evacuated community $k \in K$ (if $w_{jk} = 1$, it means that the routing will be realized, if $w_{jk} = 0$, it means that the routing will not be realized),
- z_{kl} - it is the bivalent variable modeling the decisions about realization of the routing of the vehicles from the evacuated community $k \in K$ to the evacuation center $l \in L$ (if $z_{kl} = 1$, it means that the routing will be realized, if $z_{kl} = 0$, it means that the routing will not be realized),
- h - it is the maximum time passing from the moment of declaration of the evacuation to the moment its end (which is the arrival of the last vehicle to the evacuation centre),
- r_k - it is the number of the unused places in the vehicles leaving the community $k \in K$.

The original model proposed in the article [Teichmann, Peško, 2010] is shown below:

$$min \quad f(h) = h \tag{1}$$

$$min \quad f(r) = \sum_{k \in K} r_k \tag{2}$$

under conditions

$$\sum_{k \in K} x_{ijk} \le a_{ij} \qquad \text{for } i \in I \text{ a } j \in J \tag{3}$$

$$\sum_{i \in I} \sum_{j \in J} c_i x_{ijk} \ge b_k \qquad \text{for } k \in K \tag{4}$$

$$\sum_{j \in J} x_{ijk} = \sum_{l \in L} y_{ikl} \qquad \text{for } i \in I \text{ a } k \in K \tag{5}$$

$$x_{ijk} \le T\ w_{jk} \qquad \text{for } i \in I,\ j \in J \text{ a } k \in K \tag{6}$$

$$y_{ikl} \le T\ z_{kl} \qquad \text{for } i \in I,\ k \in K \text{ a } l \in L \tag{7}$$

$$\sum_{l \in L} z_{kl} = 1 \qquad \text{for } k \in K \tag{8}$$

$$\sum_{k \in K} b_k z_{kl} \le q_l \qquad \text{for } l \in L \tag{9}$$

$$t_{1jk} w_{jk} + t_{2kl} z_{kl} \le h \qquad \text{for } j \in J,\ k \in K \text{ a } l \in L \tag{10}$$

$$\sum_{i \in I} \sum_{j \in J} c_i x_{ijk} = b_k + r_k \qquad \text{for } k \in K \tag{11}$$

$$x_{ijk} \in Z_0^+ \qquad \text{for } i \in I,\ j \in J \text{ a } k \in K \tag{12}$$

$$y_{ikl} \in Z_0^+ \qquad \text{for } i \in I,\ k \in K \text{ a } l \in L \tag{13}$$

$$w_{jk} \in \{0;1\} \qquad \text{for } j \in J \text{ a } k \in K \tag{14}$$

$$z_{kl} \in \{0;1\} \qquad \text{for } k \in K \text{ a } l \in L \tag{15}$$

$$h \ge 0 \tag{16}$$

$$r_k \geq 0 \qquad \text{for } k \in K \tag{17}$$

The function (1) represents the first objective function - maximum time of evacuation; the function (2) represents the second objective function – total number of the unused places in the vehicles leaving the communities. The restrictive condition of the type (3) ensures that the number of available vehicles for every station and every type of vehicles will not be exceeded. The restrictive condition of type (4) makes sure that for every community will be offered enough places in the vehicles and therefore it is possible to evacuate all citizens. The restrictive condition of type (5) ensures the continuity of the vehicles flows for every evacuated community. The restrictive conditions of type (6) and (7) ensure the needed relationships between the sizes of the vehicles flows in the particular degrees of evacuation process and the decisions about the mutual relationship of the places taking part in the evacuation of the citizens. The restrictive condition of type (8) ensures that every evacuated community will be attached only to single evacuation center. The restrictive condition of type (9) makes sure that during the evacuation process the capacity of any centre of evacuation can not be exceed. The restrictive condition of type (10) creates a relationship between the system of conditions and the objective function (1). The restrictive condition of type (11) creates the relationship between the variables that represent the number of the unused places in the vehicles that take part in the evacuation process and the objective function (2). The restricted conditions (12) – (17) are the obligatory conditions defining domains of variables.
In order to optimize the objective function (1) we need the conditions (3) – (10) and (12) – (16). To optimize the objective function (2) the conditions (3), (5), (7) – (9), (11) – (13), (15) and (17) are needed. For the optimization process of the two criteria it will be used the STEM method. Its theoretic description is published in [Fiala, 2008]. In this method the maximum deviation in the objective functions is minimized between the partial optimal solutions (the optimal solutions according to the partial objective functions) and the compromise solution.
It is possible to substitute the objective functions of the model described in the article [Teichmann, Peško, 2010] through aggregate function

$$\min f(h) = T\ h + \sum_{k \in K} r_k \tag{18}$$

solved under conditions (3) and (5) – (17).
The constant $T > 0$ is the rightly chosen value so that it can be possible to prefer the first objective function (the maximum evacuation time is always more important than the number of the unused places in the vehicles).
In the following chapter we will show the application of both methods on the concrete example. The solution of proposed model was realized analogously as in the published article [Teichmann, Peško, 2010] by using of the optimization software Xpress-IVE.
Let us notice that the experiments were executed in the demo version that is free available for academic purposes.

4 Executed experiments

In the paper [Teichmann, Peško, 2010] a simple case was proposed to verify models function. The model was simple because there were 2 types of vehicles, 2 stations, 2 evacuated communities and 2 evacuation centers. In the following experiments we have enlarged the extension and the

dimension of the problem. From all of realized experiments we show in this article the results of one of them. Our possibilities of experiments were restricted because the demo version of the optimization software Xpress – IVE does not permit us to work with a too great number of variables and conditions. Afterwards we show how the calculated results obtained by both methods can be differentiated. We remember that the first approach is based on the STEM method; the second approach contains the aggregate objective function (we write both criteria into one function). The number of types of vehicles is the same; the number of other values will be enlarged.
We have 4 stations, 11 evacuated communities and 3 evacuation centers.
The numbers of available vehicles for various types are declared in the table 1.

Table 1: The numbers of available vehicles.

Type of vehicles	Capacity of the vehicle [persons]	Number of vehicles			
		Station 1	Station 2	Station 3	Station 4
1	150	5	3	3	4
2	250	1	1	0	1

The referred information to the requests of the evacuated communities, the capacities of the evacuation centers and the times for the individual rides are included in the table 2. The times of rides are presented in minutes.

Table 2: Summary of input data.

	Number of the evacuated citizens	Time of ride						
		Station				Center of evacuation		
		1	2	3	4	1	2	3
Community 1	251	20	40	45	40	30	40	10
Community 2	300	30	10	15	20	50	60	40
Community 3	300	12	45	50	55	40	40	30
Community 4	258	45	30	35	40	30	50	20
Community 5	401	30	30	35	40	10	25	10
Community 6	80	10	15	10	15	15	20	5
Community 7	100	15	20	15	20	30	10	20
Community 8	200	20	10	5	10	20	20	10
Community 9	12	35	30	25	30	10	50	10
Community10	50	40	35	30	40	25	25	15
Community11	100	50	40	35	40	45	45	35
Capacity of the evacuation center						700	600	800

The results of the experiment - approach 1 (STEM Method)

According to the STEM Method the first step to multi-criterion Method of optimization is the separated realization of the calculations according to the single criteria.
First of all it is executed the optimization calculation according to the first objective function – which minimizes the maximum time of evacuation.
After the end of the calculation we reached the following results - the evacuation of the population will be realized according to the plan declared in table 3. Table 4 declares how to assign the evacuated communities to the evacuation centers.

Table 3: The results reached after minimizing of the first objective function.

	Capacity of the vehicles	Com. 1	Com. 2	Com. 3	Com. 4	Com. 5	Com. 6
Station 1	5/1	-/-	-/-	1/1	-/-	-/-	1/-
Station 2	3/1	-/-	2/-	-/-	1/-	-/1	-/-
Station 3	3/0	2/-	-/-	-/-	-/-	-/-	-/-
Station 4	4/1	-/-	-/-	-/-	1/-	1/1	-/-
	Capability of the vehicles	Com. 7	Com. 8	Com. 9	Com. 10	Com. 11	
Station 1	5/1	-/-	1/-	1/-	1/-	-/-	
Station 2	3/1	-/-	-/-	-/-	-/-	-/-	
Station 3	3/0	-/-	-/-	-/-	-/-	1/-	
Station 4	4/1	1/-	1/-	-/-	-/-	-/-	

Notice: The values are indicated by pairs X/Y in tabla 3 and represent data about the vehicles allocated for every single type before the declaration of the evacuation and their movement after the declaration of the evacuation among the decided places. The symbol X is related to the number of vehicles with capacity of 150 available places. The symbol Y is related to the number of vehicles with capacity of 250 available places

Table 4: The assignment of the evacuated communities to the evacuations centers after first step.

	Com. 1	Com. 2	Com. 3	Com. 4	Com. 5	Com. 6	Com. 7
EC 1			x	x		x	
EC 2		x					x
EC 3	x				x		
	Com. 8	Com. 9	Com. 10	Com. 11			
EC 1			x				
EC 2	x						
EC 3		x		x			

EC – center of evacuation

The value of the objective function is 70. This fact means that the evacuation will be realized by 70 minutes from its declaration.
In the further step it is calculated the optimum solution according to the second objective function which minimizes the number of unused places in the vehicles.

After the end of the calculation it was reached the following solution - the evacuation of the citizens will pass according to the plan declared in table 5. Table 6 declares how to assign the evacuated communities to the evacuation centers.

Table 5: The results reached after minimizing of the second objective function.

	Capacity of the vehicles	Com. 1	Com. 2	Com. 3	Com. 4	Com. 5	Com. 6
Station 1	5/1	2/-	-/-	2/-	-/-	-/1	-/-
Station 2	3/1	-/-	-/-	-/-	-/-	-/1	-/-
Station 3	3/0	-/-	2/-	-/-	-/-	-/-	1/-

Station 4	4/1	-/-	-/-	-/-	2/-	-/-	-/-
	Capacity of the vehicles	Com. 7	Com. 8	Com. 9	Com. 10	Com.11	
Station 1	5/1	1/-	-/-	-/-	-/-	-/-	
Station 2	3/1	-/-	-/-	1/-	1/-	1/-	
Station 3	3/0	-/-	-/-	-/-	-/-	-/-	
Station 4	4/1	-/-	-/1	-/-	-/-	-/-	

Table 6: The assignment of the evacuated communities to the evacuations centers after second step.

	Com. 1	Com. 2	Com. 3	Com. 4	Com. 5	Com. 6	Com. 7
EC 1		x	x				x
EC 2					x	x	
EC 3	x			x			
	Com. 8	Com. 9	Com. 10	Com. 11			
EC 1							
EC 2				x			
EC 3	x	x	x				

The value of the objective function is 648. This fact means that there are 648 free places not utilized in the vehicles.

According to theoretical proceeding [Fiala, 2008] the matrix of values of the objective functions was constructed, the weights were calculated and two-criterion model was constructed. After the end of the calculation it was reached the following solution - the evacuation of citizens will pass according to the plan declared in table 7.

Table 7: The reached plan of the evacuation.

	Capacity of the vehicles	Com. 1	Com. 2	Com. 3	Com. 4	Com. 5	Com. 6
Station1	5/1	1/-	-/-	2/-	-/-	-/1	1/-
Station 2	3/1	1/-	-/-	-/-	1/-	-/-	-/-
Station 3	3/0	-/-	2/-	-/-	-/-	-/-	-/-
Station 4	4/1	-/-	-/-	-/-	1/-	-/1	-/-
	Capacity of the vehicles	Com. 7	Com. 8	Com. 9	Com. 10	Com. 11	
Station1	5/1	-/-	-/-	-/-	1/-	-/-	
Station 2	3/1	1/-	-/1	-/-	-/-	-/-	
Station 3	3/0	-/-	-/-	-/-	-/-	1/-	
Station 4	4/1	-/-	-/-	1/-	-/-	-/-	

Table 8 declares how to assign the evacuated communities to evacuation centers.

Table 8: The final assignment of the evacuated communities to the evacuations centers.

	Com. 1	Com. 2	Com. 3	Com. 4	Com. 5	Com. 6	Com. 7
EC 1	x	x				x	
EC 2			x				x
EC 3				x	x		
	Com. 8	Com. 9	Com. 10	Com. 11			
EC 1			x				
EC 2	x						
EC 3		x		x			

The value of the objective function is 0. This fact means that the optimal solutions reached by the previous steps were not aggravated. The maximum time of evacuation is 70 min, and the number of unused places in the vehicles is 648.
The convergence process of calculation towards the optimal solution in the software Xpress-IVE is shown in figure 1.

Figure 1: The convergence process of calculation for the first approach.

The results of the experiment - approach 2 (aggregate objective function)

The second approach is from the point of view of the preparation of input data easier than the first one.

Until now we have seen that the first approach consists of 3 phases, for the second approach only one phase is needed in order to get the optimal solution. Only one model must be solved. Please notice that the value of prohibitive constant is $T = 1000$.
After the end of the calculation it was reached the following solution - the evacuation of citizens will pass according to the plan declared in the table 9. Table 10 declares how to assign the evacuated communities to the evacuation centers.

Table 9: The reached plan of the evacuation.

	Capacity of the vehicles	Com. 1	Com. 2	Com. 3	Com. 4	Com. 5	Com. 6
Station 1	5/1	-/-	2/-	2/-	-/-	-/-	-/-
Station 2	3/1	-/-	-/-	-/-	2/-	-/1	-/-
Station 3	3/0	-/-	-/-	-/-	-/-	-/-	-/-
Station 4	4/1	2/-	-/-	-/-	-/-	-/1	1/-

	Capacity of the vehicles	Com. 7	Com. 8	Com. 9	Com. 10	Com. 11
Station1	5/1	-/-	-/1	-/-	-/-	-/-
Station 2	3/1	-/-	-/-	-/-	-/-	-/-
Station 3	3/0	-/-	-/-	1/-	1/-	1/-
Station 4	4/1	1/-	-/-	-/-	-/-	-/-

Table 10: The assignment of the evacuated communities to the evacuations centers.

	Com. 1	Com. 2	Com. 3	Com. 4	Com. 5	Com. 6	Com. 7
EC 1	x				x		
EC 2			x				x
EC 3		x		x		x	

	Com. 8	Com. 9	Com. 10	Com. 11
EC 1				
EC 2	x			
EC 3		x	x	x

The objective function has the value 70648.
On the basis of its analysis, we can state that the maximizing of the evacuation time presents that the maximum time is 70 min and the number of unused places is 648.
If we compare the solution time of the model in the last step of the first approach with the solution time of the model with the aggregate criterion, we will see that the calculation time was longer in the experiment with the aggregate function.
The convergence process of calculation towards the optimal solution reached by the software Xpress-IVE is shown in figure 2.

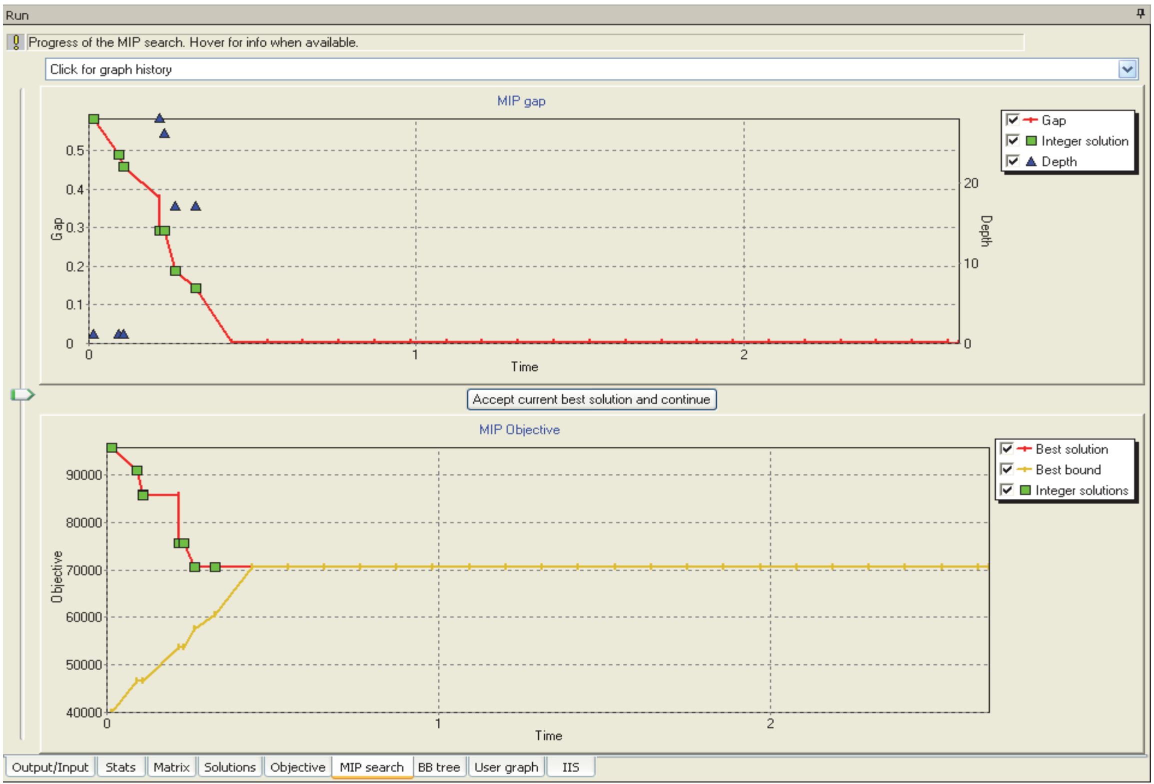


Figure 2: The convergence process of calculation for the second approach.

5. Conclusion

In the presented article we consider the issue of the optimization of the problem with two criteria regarding the area evacuation planning. In the paper two possible approaches for the solution of the model with two criteria are shown. In the final part the utility of the models is tested in the example in which the number of the evacuated communities is quite near to the real situation – of a smaller area. For the future times we plan other experiments with essentially more laborious problems to solve. According to the values of the objective function we can see that in both cases we reached the same quality of solution. In both experiments the optimal solution was reached.

References

[1] DAGANZO, C., F.: *The Cell Transmission Model, Part II: Network Traffic.* Transportation Research Part B: Methodological, Volume 29, Issue 2, April 1995, p. 79-93. ISSN 0191-2615.

[2] BISH, D., R.: *Demand Management in Evacuation Planning: Models, Algorithms, and Applications.* Dissertation. Virginia Polytechnic Institute and State University, 2006. 105 p.

[3] El-SBAYTI, H., H.: *Optimal Scheduling of Evacuation Operations with Contraflow.* Dissertation. University of Maryland, 2008. 210 p.

[4] FIALA, P.: *Modely a metody rozhodování.* Praha: Oeconomica, 2008. 292 p. ISBN 978-80-245-1345-4.

[5] JANÁČEK, J.: *Optimal Evacuation Plan Design with IP-solver*. Communications, 2009, Volume 11, Issue 3, p. 29-35. ISSN 1335-4205

[6] LIM, G., J., ZANGENEH, S., BAHARNEMATI, M., R., Assavapokee, T.: *A simple binary search algorithm for short notice evacuation scheduling and routing.* In Proceedings of the Industrial Engineering Research Conference*,* Miami, May, 2009.

[7] LU, Q.; GEORGIE, B.; SHEKHAR, S.. Capacity Constrained Routing Algorithms for Evacuation Planning: A Summary of Result. In Advances in Spatial and Temporal Databases Lecture Notes in Computer Science, 2005, Volume 3633/2005, pp. 291-307, DOI: 10.1007/11535331_17

[8] PEL, A., J.: Evacuation Plan Evaluation: Assessment of Vehicular Evacuation Schemes by Means of an Analytical Dynamic Traffic Model. Thesis. Delft University of Technology, 2007. 54 p.

[9] SIMONOVIC, S., P., Ahmad, S.: *Computer-based Model for Flood Evacuation Emergency Planning.* Natural Hazards, Volume 34, Number 1, January 2005 , pp. 25-51. DOI: 10.1007/s11069-004-0785-x

[10] SOUTHWORTH, F.. Regional Evacuation Modeling: A State of the Art Review. Oak Ridge: Oak Ridge National Laboratory, 1991, 65 p.

[11] TEICHMANN, D., PEŠKO, Š.: *Two-Criteia Optimalization of the Process Evacuation.* In World and Homeland Security, 6th International Konference. Brno, Jun 2010, Ed. Horák, R., Schwarz, R.. pp. 602 – 611. ISBN 978-80-7231-728-8

[12] ZENG, M., WANG, Ch.: *Evacuation Route Planning Algorithm : Longer Route Preferential.* In Advances in Neural Network ISNN 2009. 6th International Symposium on Neural Network, ISNN 2009 Wuhan, China, May 2009 Proceedings, Part I. Ed. Yu, He, Zhang pp. 1062 – 1071. ISSN 0302-9743

Current address

Teichmann, Dušan, Ing., Ph.D.
VŠB-Technical University of Ostrava, Faculty of Mechanical Engineering, 17. listopadu 15/2172, 708 33 Ostrava - Poruba, tel. +420597324575,
e-mail: dusan.teichmann@vsb.cz

Dorda, Michal, Ing., Ph.D.
VŠB-Technical University of Ostrava, Faculty of Mechanical Engineering, 17. listopadu 15/2172, 708 33 Ostrava - Poruba, tel. +420597325754,
e-mail: michal.dorda@vsb.cz

Grosso, Alessandra, Ing.
VŠB-Technical University of Ostrava, Faculty of Mechanical Engineering, 17. listopadu 15/2172, 708 33 Ostrava - Poruba, tel. +420597325754

Ivan, Martin, Ing.
VŠB-Technical University of Ostrava, Faculty of Mechanical Engineering, 17. listopadu 15/2172, 708 33 Ostrava - Poruba, tel. +420597324575

Journal of Applied Mathematics

volume IV (2011), number I

SIMULATION OF PATTERN RECOGNITION SYSTEM VIA MODULAR NEURAL NETWORKS

VOLNÁ Eva, (CZ), KOTYRBA Martin, (CZ), KOCIAN Václav, (CZ), JANOŠEK Michal, (CZ)

Abstract. It is the purpose of the present paper to suggest an approach to utilization of mathematical models of a classification system for pattern recognition. Pattern recognition has a long history but has recently become much more widespread as the automated capture of signals and images has become cheaper. Very many of the applications of neural networks are to classification, and so are within the field of pattern recognition.
This article describes a classification system for pattern recognition based on artificial neural networks with modular architecture. We use a three layer feedforward network model that is learned with the backpropagation algorithm for all experiments. Our experimental recognition objects were digits and their type fonts. We also propose outline further development on this topic in conclusion.

Key words. Simulation, pattern recognition, modular neural networks, neuro-classifier.

Mathematics Subject Classification: Primary 82C32, 68T10; Secondary 68T05.

1 Classification via Neural Networks

Classification is one of the most active research and application areas of neural networks. The advantage of neural networks lies in the following theoretical aspects. First, neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. Second, they are universal functional approximators in that neural networks can approximate any function with arbitrary accuracy [10]. Since any classification procedure seeks a functional relationship between the group membership and the attributes of the object, accurate identification of this underlying function is doubtlessly important. Third, neural networks are nonlinear models, which makes them flexible in modeling real world complex relationships. Finally, neural networks are able to estimate the posterior probability, which provides the basis for establishing classification rule and performing statistical analysis [15].

On the other hand, the effectiveness of neural network classification has been tested empirically. Neural networks have been successfully applied to a variety of real-world classification tasks in industry, business and science. Applications include bankruptcy prediction [7; 13], handwriting recognition [8], speech recognition [16], product inspection [3], fault detection [6], medical diagnosis [9], bond rating [11] etc. A number of performance comparisons between neural and conventional classifiers have been made by many studies, e.g. [1; 2; 4]. In addition, several computer experimental evaluations of neural networks for classification problems have been conducted under a variety of conditions, e.g. [12; 18].
This article describes a classification system for handwritten digit recognition based on artificial neural networks. We propose an study of a pattern recognition system using neural network technologies and outline further development on this topic.

2 Artificial neural networks

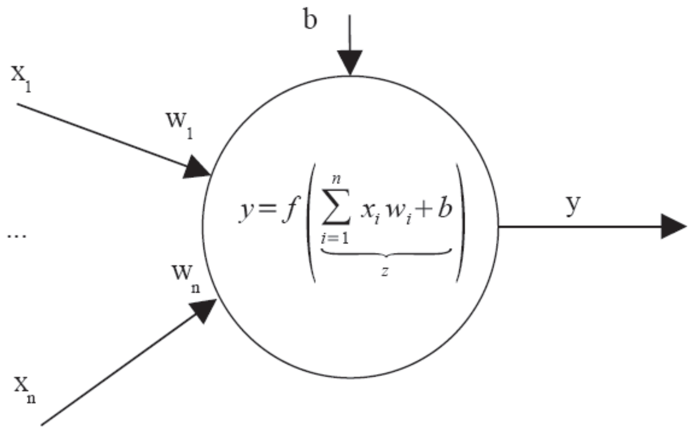


Figure 1. A simple artificial neuron [14].

An Artificial Neural Network (ANN) is a massively parallel system, inspired by the human neural system. Its units, neurons (one is shown in Figure 1), are interconnected by connections called synapses. A neuron obtains input signals ($x_1, ..., x_n$) and relevant weights of connections ($w_1, ..., w_n$), optionally a value called bias b is added in order to shift the sum relative to the origin. The weighted sum of inputs is computed and the bias is added so that we obtain a value called stimulus or inner potential (z) of the neuron. After that it is transformed by an activation function f into output value y, which may be propagated to other neurons as their input or be considered as an output of the network. A i-unit's state, y_i is computed as it is shown in equations (1), where b_i is its bias and the activation function is a sigmoid. The purpose of the activation function is to perform a threshold operation on the potential of the neuron.

$$\begin{aligned} z_i &= \sum_{j=1}^{n} w_{ij} x_j + b_i \\ y_i &= \left(1 + e^{-z_i}\right)^{-1} \end{aligned} \tag{1}$$

This article describes a classification system for pattern recognition based on artificial neural network with modular architecture. Multilayer feedforward neural networks belong to the most common ones in practical use. Its architecture contains units organized into three different types of layers. A subset of input units has no input connections from other units; their states are fixed by the problem. Another subset of units is designated as output units; their states are considered the result of the computation. Units that are neither input nor output are known as hidden units.
Learning algorithm of such neural network called backpropagation and belongs to a group called "gradient descent methods". When looking at Figure 2, it is obvious that the initial position on the weight landscape greatly influences both the length and the path made when seeking the global minimum.

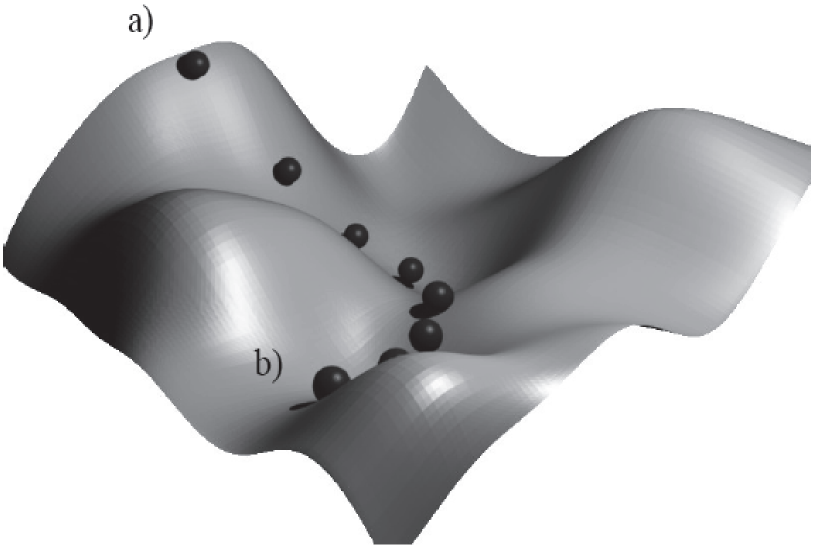


Figure 2. An intuitive approach to the gradient descent method, looking for the global minimum [14]. The starting point is a), the final one is b).

Now, a more formal definition of the backpropagation algorithm (for a three layer network) is presented [14].

1. The input vector is presented to the network.
2. The feedforward is performed, so that each neuron computes its output (y_i) following the formula (1) over neurons in previous layer.
3. The error on the output layer is computed for each neuron using the desired output (o_i) on the same neuron, see formula (2).

$$err_i^0 = y_i(1 - y_i)(o_i - y_i) \tag{2}$$

4. The error is propagated back to the hidden layer over all the hidden neurons (h_i) and weights between each of them and over all neurons in the output layer see formula (3).

$$err_i^h = h_i(1 - h_i)\sum_{j=1}^{r} err_j^0 w_{ij}^0 \tag{3}$$

5. Having values err_j^0 and err_i^h computed, the weights from the hidden to the output layer and from the input to the hidden layer can be adjusted following formulas (4), where α is the learning coefficient and x_i is the *i-th* neuron in the input layer.

$$\begin{aligned} w_{ij}^0(t+1) &= w_{ij}^0(t) + \alpha\, err_j^0 h_i \\ w_{ij}^h(t+1) &= w_{ij}^h(t) + \alpha\, err_j^h x_i \end{aligned} \tag{4}$$

6. All the preceding steeps are repeated until the total error of the network (5) over all training pairs does not fall under certain level, where *m* is number of units in the output layer.

$$E = \frac{1}{2}\sum_{i=1}^{m}(y_i - o_i)^2 \tag{5}$$

The formulas in step three and four are products of derivation of the error function on each node. A detailed explanation of this derivation as well as of the complete algorithm can be found in [5].

3 Modular neural networks

The modular network architecture has advantages in terms of learning speed [17]. Several characteristics of modular architectures suggest that they should learn faster than networks with complete sets of connections between adjacent layers. One such characteristic is that modular architectures can take advantage of function decomposition. If there is a natural way to decompose a complex function into a set of simpler functions, then a modular architecture should be able to learn the set of simpler functions faster than a monolithic multilayer network. In addition to their ability to take advantage of function decomposition, modular architectures can be designed to reduce the presence of conflicting training information that tends to retard learning. This occurs when the backpropagation algorithm is applied to a monolithic network containing a hidden unit that projects to two or more output units. Moreover, modular architectures generalize better because they perform local generalization in the sense that only learns patterns from a limited region of the input space. Modular architectures tend to develop representations that are more easily interpreted than the representations developed by single networks. As a result of learning, the hidden units of the system used in separate networks for the tasks' contributes to the solutions of these tasks in more understandable ways than the hidden units of the single network applied to both tasks.
In our experiments, we use modular neural networks that are derived from classical feedforward networks with some connections missing, usually a task decomposing network into some modules. These modules usually have only one input layer, by which the entire input vector is presented to "inner" modules, each of which is represented by its own units in the hidden and output layer.

4 Modular Neuroclassifier

In this article, for all experiments a three layer feedforward network model is used and it is learned with the backpropagation algorithm. Our experimental recognition objects were digits and their type fonts. Digits acquired as binary images. A digitized character image consists of pixels, usually black on a background of white. The three layer net was trained using 50 digits; that is one sample of each class 0 through 9 was presented, followed by another set of 0 – 9 and so on. The images were uniformly divided into 9 x 7 pixel grids. The whole training set is shown in the Figure 3. In our simulations the neural network performs a task that is decomposable into simple tasks. The input layer is fully interconnected with the hidden layer. In the split model the pattern of connectivity between the hidden and output layers is restricted by partitioning the hidden nodes into two subsets (e.g. modules). Nodes in each of the subsets are connected only to the output nodes that are associated with one of the modules. The split model thus functions as two distinct systems, overlapping only at the input level. We have used the modular neural network architecture that represents both shape and its font. The network architecture is the following: 63 – 40 –15. The shape recognition was represented by one system and we use the network architecture 63 – 20 –10, while the font recognition was represented by another system and we use the network architecture 63 – 20 – 5. The input layer (containing 63 units) was fully interconnected with the hidden layer containing 40 units, which were divided into two modules (20 units were associated with the first module and 20 units were associated with the second module). Each unit in the hidden layer was interconnected only with output units associated with its module (10 units were associated with the first module and 5 units were associated with the second module).

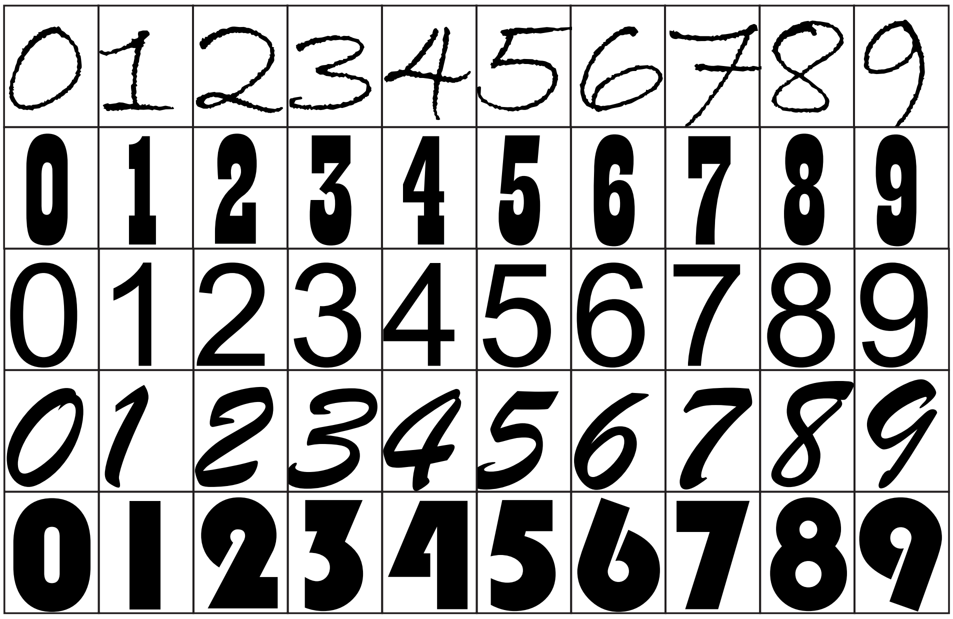


Figure 3. The training set.

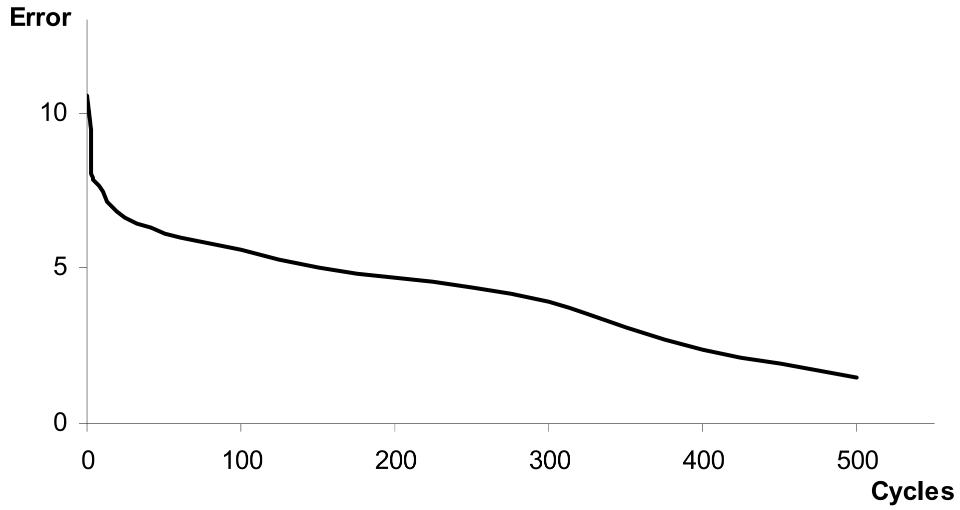


Figure 4. Error function history (average values).

The training algorithm was backpropagation, we used the following parameters: learning rate was 0.4 and momentum was 0.1. History of error functions during whole calculation is shown in the Figure 4. There is shown an average value of the error functions, because the adaptation with backpropagation algorithm was applied 10 times during 500 cycles on average.
After the training phase another set of 50 digits was used to test the recognition rate. The recognition rate using neural network technology is acceptable, averaging about 90% for training and 88% for testing, e.g. 91% for testing digits recognition and 85% testing font recognition . In detail, see Table 1. Other numerical simulations give very similar results.

character:	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
training	89%	91%	89%	88%	92%	89%	91%	92%	90%	89%
testing digit recognition	91%	91%	92%	91%	92%	90%	91%	92%	90%	90%
testing writer recognition	86%	89%	76%	80%	89%	86%	87%	88%	86%	83%

Table 1. Detail of the recognition rate using neural network technology for training and testing.

5 Conclusions and future works

Handwritten digit recognition is an important task in automated document analysis. Different methods including neural networks or statistical analysis, structured or syntactic approaches have been used to solve these problems. Applications have been developed to read postal addresses, bank checks, tax forms, and census forms, including reading aids for the visually impaired, among others. The handwritten digit recognition problem is a suitable task to explore new approaches in 2-D pattern recognition classifiers because it is a complex problem but restricted to only ten classes.
The use of artificial neural system for various recognition tasks is well founded in the literature [3; 6; 7; 8; 9; 11; 13; 16] etc. The advantage of using neural nets is not that the solution they provide is

especially elegant or even fast; it is that system 'learns' it own algorithm for the classification task, and does so on actual samples data.
In general, performance of the neural network seems quite good. The errors made by the network are only in a way understandable: the misclassified digits are often misshaped and carry features normally found in other digits. In this article we provided handwritten digit recognition for digits 0 to 9 and their fonts of using pattern recognition systems with neural network technology. This article also demonstrated that the effectiveness of such system is very good.
Our future work should address to signature verification that is an important research topic in the area of biometric verification. In bank and public offices, signature verification remains one of the most acceptable means of verifying cheque/document legitimacy. Since manual verification for a large amount of checks or documents is tedious and easily influenced by physical and psychological factors, automatic processing by computers is advocated.

Acknowledgement

The paper was supported University of Ostrava grant SGS/PřF/2011. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1.] ABU-NIMEH, S., NAPPA, D., WANG, X., and NAIR, S. A comparison of machine learning techniques for phishing detection. In *Proceedings of the eCrime Researchers Summit*, 2007.
[2.] Al-ASSAF, Y., El KADI, H. "Fatigue life prediction of composite materials using polynomial classifiers and recurrent neural networks," Composite Structures, vol. 77, no. 4, pp. 561–569, 2007.
[3.] CASASENT D., CHEN, X. "New training strategies for RBF neural networks for X-ray agricultural product inspection," *Pattern Recognition*, vol. 36(2), pp. 535-547, 2003.
[4.] CORREA, M., BIELZA, C., PAMIES-TEIXEIRA, J. "Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process". *Expert Systems with Applications: An International Journal*, Vol. 36 No 3, pp. 7270-7279, 2009.
[5.] FAUSETT, L.V. Fundamentals of Neural Networks. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1994.
[6.] FEKIH, A., H. XU and F. N. CHOWDHURY, Neural Networks Based System Identification Techniques for Model Based Fault Detection of Nonlinear Systems, *International Journal of Innovative Computing, In-formation and Control*, vol.3, no.5, pp.1073-1085, 2007.
[7.] GHIASSI, M., SAIDANE, H., & ZIMBRA, D.K. "A dynamic artificial neural network model for forecasting series events." *International Journal of Forecasting*, 21, (2005) 341–362.
[8.] GRAVES, A., FERNANDEZ, S. LIWICKI, M., BUNKE, H. and SCHMIDHUBER, J. "Unconstrained Online Handwriting Recognition with Recurrent Neural Networks," J. Platt, In Koller, D., Singer, Y. and Roweis, S. (eds.) *Advances in Neural Information Processing Systems, 20,* 2008.
[9.] HABAS, P.A., ZURADA, J.M., ELMAGHRABY, A.S., TOURASSI, G.D. "Particle swarm optimization of neural network CAD systems with clinically relevant objectives", *Proceedings of Medical Imaging 2007*: Computer-Aided Diagnosis; 2007.
[10.] HORNIK, K. "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, 4, (1991) 251–257.

[11.] HUANG, Z., CHEN, H., HSU, C.-J., CHEN, W.-H., & WU, S. "Credit rating analysis with support vector machine and neural networks: A market comparative study." *Decision Support Systems*, 37, (2004)543–558.
[12.] OU, G., MURPHEY, Y.L., and FELDKAMP, L. "Multiclass pattern classification using neural networks." In Proceeding of the *International conference on Pattern Recognition* (ICPR 2004), volume IV, 2004, pp. 585-588.
[13.] PENDHARKAR, P. C. "A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem." *Computers and Ops Res.*, 2005, **32**(10), 2561–2582.
[14.] RAMIK, D. Modular neural networks. *Bachelor thesis.* University of Ostrava, 2006.
[15.] RICHARD, M. D. and LIPPMANN, R. "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Comput*., 3, (1991) 461–483.
[16.] SKOWRONSKI, M. D., HARRIS, J. G. "Automatic speech recognition using a predictive echo state network classifier." *Neural Netw.* 20, (2007) 414–423.
[17.] VOLNA, E. Emergence of modularity in evolved neural networks. Neural Network World. Int. Journal on Neural & Mass-Parallel Comp. and Inf. Systems. Vol.13, No.6, pp. 617-628, 2003.
[18.] ZHANG, G. P. "Neural networks for classification: A survey." *IEEE Transactions Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 30(4), (2000) 451–462.

Current address

Eva Volná, doc. RNDr. PaedDr. PhD.
University of Ostrava, Faculty of Science
Dept. of Department of Informatics and Computers
30th dubna st. 22, 70103 Ostrava, Czech Republic
email: eva.volna@osu.cz

Martin Kotyrba, Mgr.
University of Ostrava, Faculty of Science
Dept. of Department of Informatics and Computers
30th dubna st. 22, 70103 Ostrava, Czech Republic
email: martin.kotyrba@student.osu.cz

Václav Kocian, Mgr.
University of Ostrava, Faculty of Science
Dept. of Department of Informatics and Computers
30th dubna st. 22, 70103 Ostrava, Czech Republic
email: vaclav.kocian@ student.osu.cz

Michal Janošek, Mgr.
University of Ostrava, Faculty of Science
Dept. of Department of Informatics and Computers
30th dubna st. 22, 70103 Ostrava, Czech Republic
email: michal.janosek@ student.osu.cz

Journal of Applied Mathematics

volume IV (2011), number I

APPLICATION OF SOFTWARE MATHEMATICA IN MATHEMATICS

BĚLAŠKOVÁ Silvie, (CZ), OPLATKOVA Zuzana, (CZ)

Abstract. This paper is focused on usage of software *Mathematica* as a visualization tool in study course Mathematics. To show the advantages of usage *Mathematica* software a few examples from the field of derivation and integration of a function with two variables were selected and described.
Key words. Mathematica, education, mathematics

Mathematics Subject Classification: 97A80.

1 Introduction and motivation

The huge expansion of information technologies, which take part into many fields currently, is one of the main initiations to connect mathematics with computer technology possibilities. The development of computer technologies and software leads to use it also in the education for visualization and explanation of the taught problems. This method has a lot of sympathizers but also a lot of opponents. Through indisputable advantages of IT in this field we put emphasis mainly on classical approach and we look also for a harmony between approved methods and news offered by IT.
The main aim of this paper is to show that usage of information technologies in education of mathematics is a suitable way. It will be shown what the steps to this conclusion led.
During the lessons teachers had a lot of problems to motivate students to study mathematics. The most often argumentation of our students was that they came to study information technologies to Faculty of Applied Informatics and not mathematics. Even though our teachers tried to explain that mathematics is very important also in the area of information technologies; students were still on the opposite side. Before the final decision to innovate style of education number of hours was discussed too. The experience of Slovak Technical University in Bratislava was taken into mind and a similar model was used – two hours of lectures and two hours of classical approach seminar that alternates with one hour of classical and one hour of computer computation. Exercises of solving different problems are the same for classical approach and programming in Mathematica software to see the difference and to visualize the problems. A choice of the software was obvious –

Mathematica (www.wolfram.com) [1.], [2.]. This software has been bought for our faculty in the Unlimited version and therefore students are allowed to ask for one home user license. On that account, they can work also at home legally. Software *Mathematica* is used not only in Mathematics but also in a lot of other study courses like Cryptology, Theory of Programs, Basics of Informatics, Methods and Applications of Artificial Intelligence and others. Student's responds to questions about *Mathematica* software were positive, similarly as from our colleagues. Students get used to use *Mathematica* software at home and send seminar works from the field of mathematics programmed in the *Mathematica*. Examples of some works will be introduced in following sections. First part of this paper is focused on exercises prepared for students. The advantages of *Mathematica,* which bring to teachers, will be described briefly. Also a discovered problem will be mentioned. Next part will be focused on student seminar works with their permission. As a conclusion contributions for the tuition will be evaluated. A prepared project connected with studying materials for our students will be introduced. This project is a cooperation of teachers from two departments – dept. of mathematics and dept. of informatics and artificial intelligence.

2 Tutorials with *Mathematica* software

The gold principle according to J. A. Komensky is illustrative nature during the tutorials. In many cases, the educational process needs to replace the immediate contact by graphics, model or a simulation. Visualization is one of the most useful and offered tools of software concerned to mathematics. *Mathematica* software offers a command Manipulate that is very powerful and can bring closer many effects to students.

2.1 Derivation of a function with two variables

Examples of exercises were selected from the study course Mathematics 2 in the summer term of the first year at our faculty. First illustrative case comes from the area of derivation of a function with two variables [3.].

Figure 1 was developed in the software *Mathematica* and was edited in the way required for presentation on the www pages too. These web pages for the e-learning support are under construction. Departments of Mathematics and Department of Informatics and Artificial Intelligence, students as scientific auxiliary cooperate on these web pages.

Notebook (native format of *Mathematica* software) on

Figure 1 describes a derivation and its usage with description and commands which *Mathematica* offer for derivation computation.

Figure 1 shows the derivation of a function according the variable x. Students were not able to understand the fact that a variable y is understood as a constant. This simulation on

Figure 1 helps to explain this condition. A variable y is depicted as a section plane and also a curve – an intersection of a plane and surface - is marked. A section plane is possible to setup differently and follow the changes of derivation in the dependence of point coordinate changes.

Figure 2 shows the final state of derivation according both variables.

Figure 3 depicts a webpage in webMathematica where the notebook from the

Figure 2 is used. As webMathematica 3 works with a command Manipulate, the same behaviour as in notebook can be found also in webpages. The only difference is that Manipulate in webpages does not support editable fields.

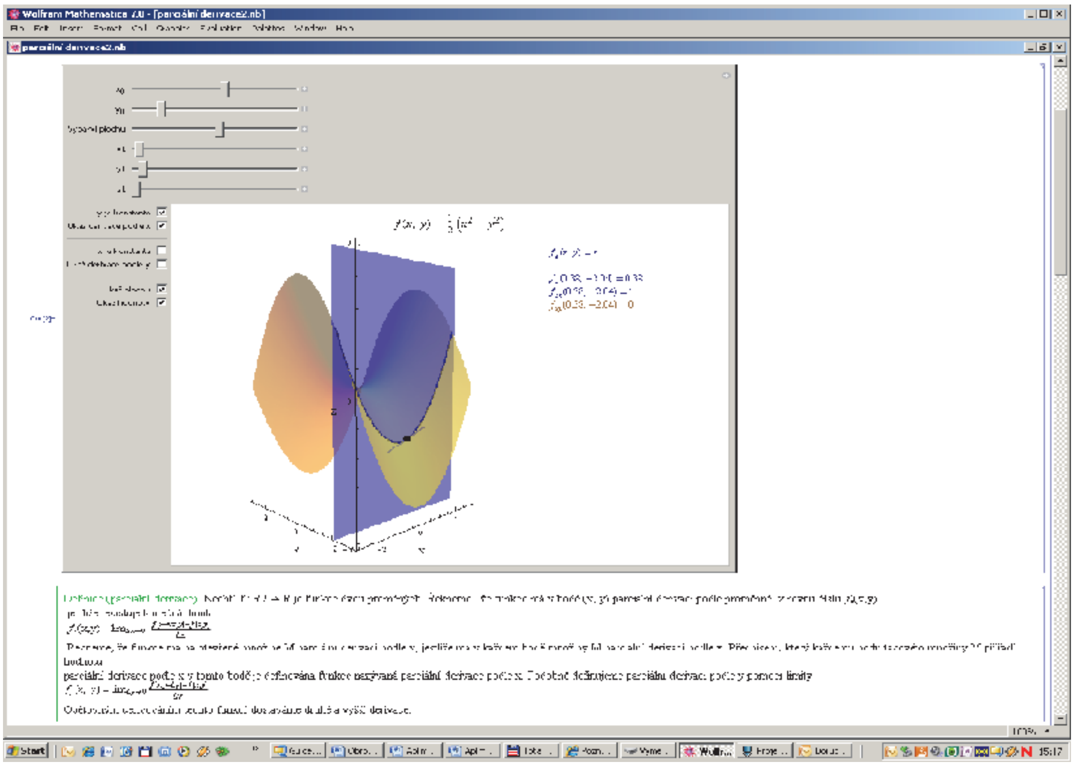


Figure 1: A notebook of Mathematica with a derivation of a function with two variables

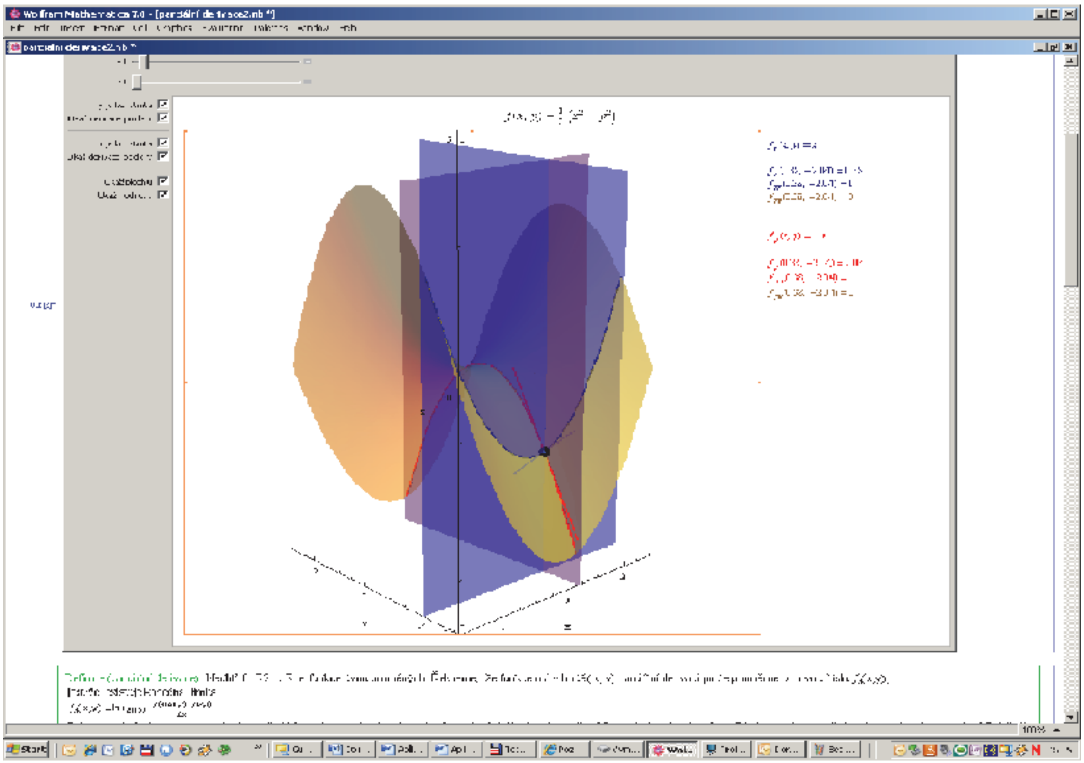


Figure 2: Derivation of a function with two variables – final state

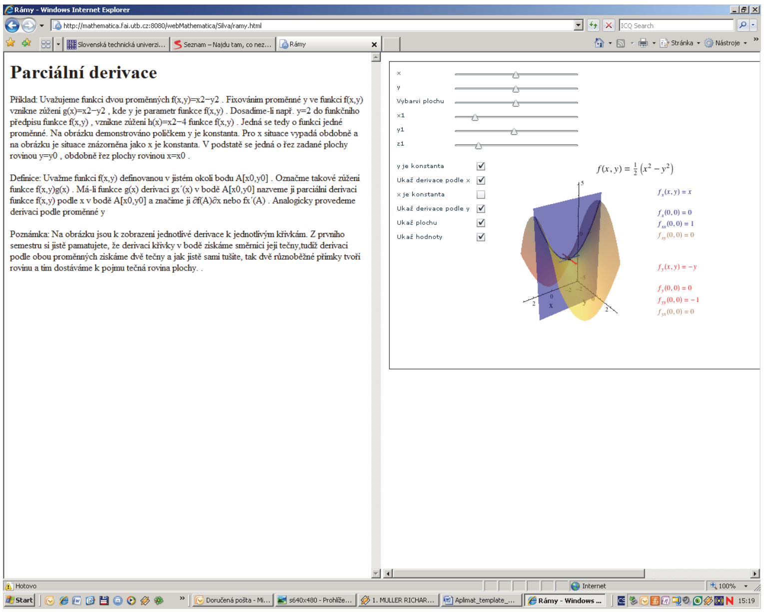


Figure 3: webMathematica version of the notebook in Figure 2

2.2 Seminar works

An advantage of home use license for students is to give them topics for seminar works. They can work on them at home. Students also come with their own ideas how to develop notebooks in a better way. Examples of student seminar works are in
Figure 4 -
Figure 5.

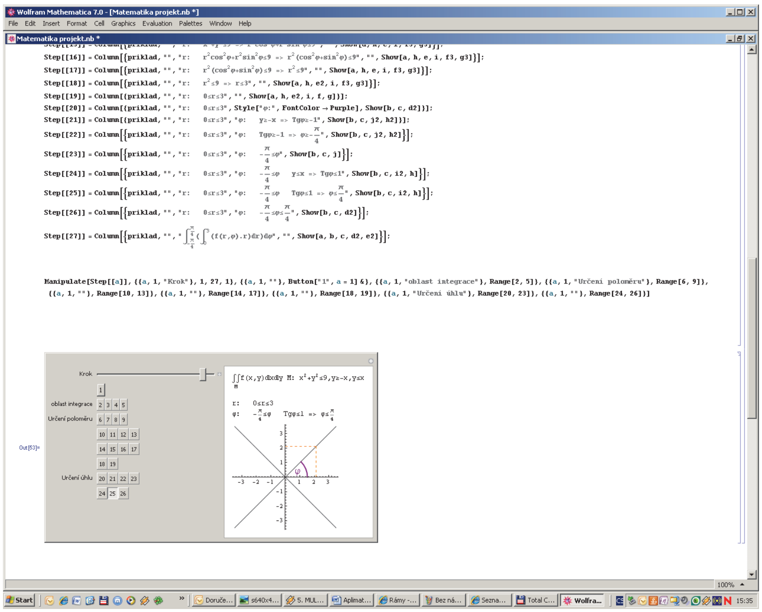


Figure 4: Integration of a function with two variables and conversion into polar coordinates

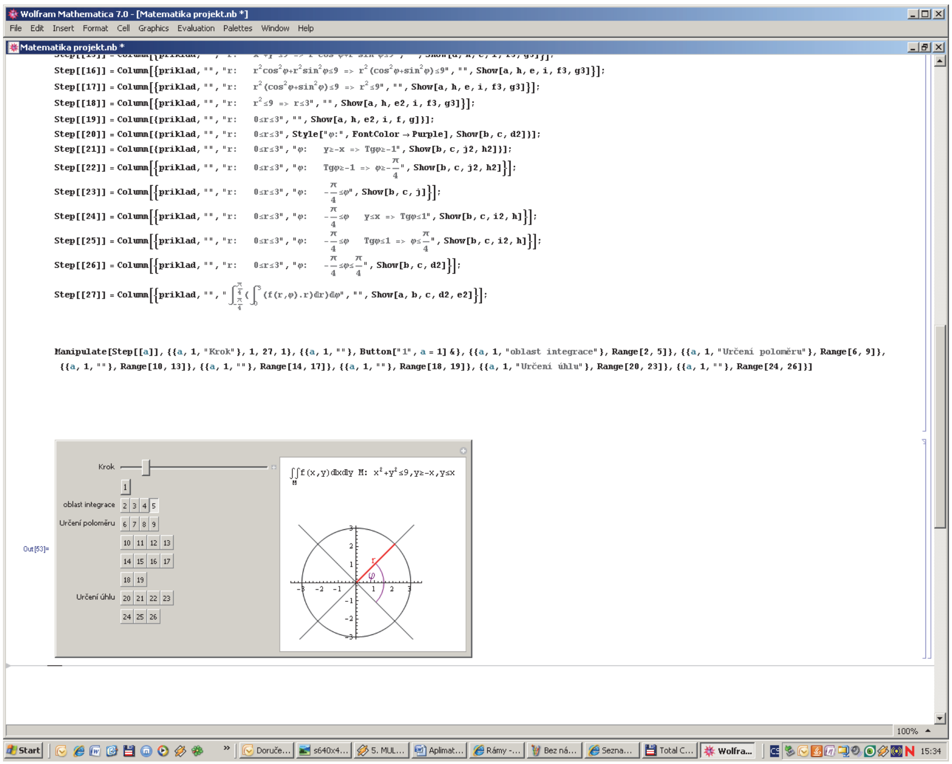


Figure 5: Integration of a function with two variables and conversion into polar coordinates II.

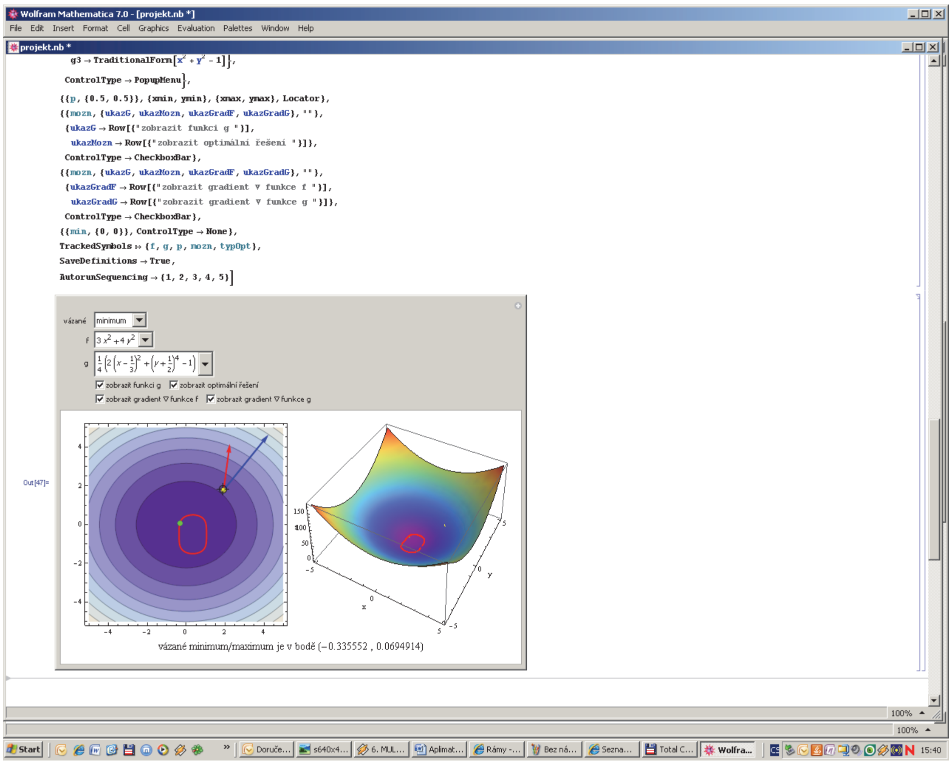


Figure 6: Searching of a weighed minimum and maximum

Seminar works are prepared in software *Mathematica* as a required part for obtaining a credit. Each student has his own assignment. Figure 4 and Figure 5 show a seminar work of a first year student who developed an application for integration of a function with two variables and visualization of each step during the process. This helps also to other students to understand better the process of integration. A part of this application is a conversion into polar coordinates.
Next seminar work dealt with a searching of a weighed minimum and maximum. This went hand by hand with a gradient and student depicted it within the dependence on a movement over the surface (Figure 6).

3 Conclusion

This article deals with an explanation why software for visualization like *Mathematica* is suitable to involve for education in the field of mathematics. Students understand better if they see graphical examples than they only hear about the taught problem. The only discovered problem is that some students try to cheat with the usage of such software like *Mathematica* or www.wolframalpha.com (software *Mathematica* on web pages) during the tests and exams. The paper shows also examples of seminar works which have been made by our students and which helps students to study much more effectively. Classical approaches are not forgotten at all. On the contrary, these are used within new approaches together and advantages from both approaches are tried to apply during tutorials and lectures.

Acknowledgement
We thank to student Martin Weiser for permission to use the printscreens of their work.

References

[1.] DOBRÁKOVÁ J., KOVÁČOVÁ M., ZÁHONOVÁ V.: *Mathematica*, Slovenská technická univerzita v Bratislavě, 2008, ISBN:978-80-89313-19-8
[2.] TROTT M., *The Mathematica Books for Programming*, Springer, 2004, ISBN: 978-0-387-94282-7
[3.] KRENEK J, OSTRAVSKY J.: *Diferenciální a integrální počet funkce jedné proměnné s aplikacemi v ekonomii*, učební texty vysokých škol, UTB Zlín, 2006, ISBN 80-7318-354-4.

Current address

Silvie Bělašková, Bc. Mgr.
Tomas Bata University in Zlín, Faculty of Applied Informatics, Department of Informatics and Artificial Intelligence, Nad Stranemi 4511, 760 05 Zlin, Czech Republic, tel: +420576035108, email: belaskova@fai.utb.cz

Zuzana Oplatková, Ing. Ph.D.
Tomas Bata University in Zlín, Faculty of Applied Informatics, Department of Informatics and Artificial Intelligence, Nad Stranemi 4511, 760 05 Zlin, Czech Republic, tel: +420576035190, email: oplatkova@fai.utb.cz

Journal of Applied Mathematics

volume IV (2011), number I

MATURITA Z MATEMATIKY NANEČISTO NA VŠB - TU OSTRAVA II

BOHÁČ Zdeněk, (CZ), DOLEŽALOVÁ Jarmila, (CZ), KREML Pavel, (CZ)

Abstrakt. Nedostatečné znalosti z matematiky u studentů nastupujících do prvních ročníků technických vysokých škol jsou častou příčinou neúspěchu ve studiu. Je to jen utkvělá představa vysokoškolských pedagogů nebo holá skutečnost? Autoři předkládaného článku využili příležitosti zavedení tzv. státní maturity z matematiky a s použitím typového zadání zveřejněného v [7] a úloh testovacího kola použitých v akademickém roce 2010/2011 prověřili znalosti studentů 1. ročníku pěti technických fakult VŠB – TU Ostrava. Výsledky jsou spíše alarmující než povzbudivé. Řada studentů by u maturity neuspěla. Přitom znalosti požadované v testu příliš nepřesáhly požadavky kladené na absolventy základních škol. V textu je proveden rozbor výsledků z pohledu fakult, jejichž studenti se testu podrobili a z pohledu typu středních škol, které studenti absolvovali.

Klíčová slova. Státní maturita, matematika, výsledky.

SCHOOL-LEAVING EXAM FROM MATHEMATICS EXPERIMENTALLY AT VŠB - TU OSTRAVA II

Abstract. Lack of knowledge of mathematics is a common cause of failure in a study for the students entering the first year at technical universities. Is it just a fixed idea of university teachers, or is it a sober reality? The authors of the presented article used the occasion of the introduction of a so called national graduation (secondary-school-leaving examination) in mathematics. Using a type assignment published in [7] and the exercises of the test run in the academic year 2010/2011, we examined the knowledge of students of the first grade at five technical faculties of VSB – TU Ostrava. The results are more alarming than encouraging. Many students would not succeed in the school-leaving examination, although the knowledge required in the test did not much exceed the requirements for graduates of primary schools. The text includes an analysis of the results in terms of faculties whose students undergo the test and in terms of types of secondary schools whose students graduated.

Key words. School-leaving Exam, Mathematics, Results.

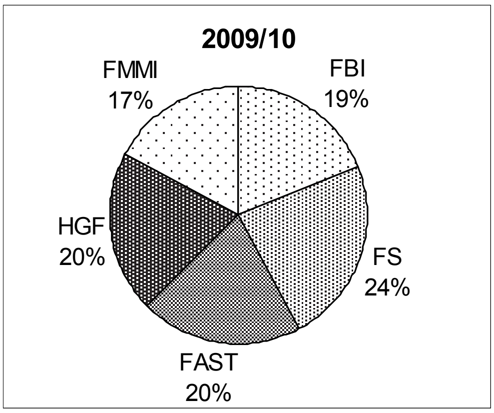
Mathematics Subject Classification: Primary 97U70

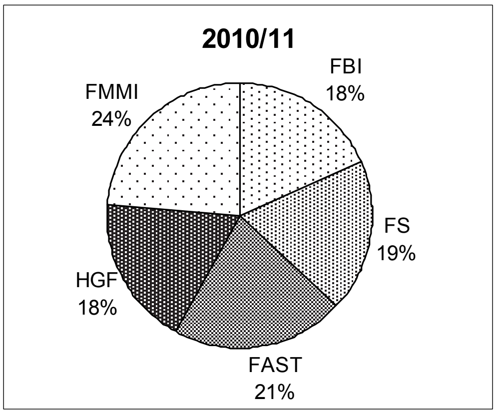
1 Úvod

Mnoho vysokoškolských pedagogů (a to nejen vyučujících matematiky) zastává názor, že úroveň matematických znalostí a dovedností studentů je dlouhodobě nízká [1-5]. Úroveň matematických znalostí současných studentů jsme zjišťovali na základě dvou maturitních testů z matematiky.

2 Testování ve školním roce 2009/10

Na internetu jsme vybrali vzorový ilustrační maturitní test státní maturitní zkoušky z matematiky z roku 2008 (jediný, který byl k dispozici), [7]. Zvolili jsme základní úroveň obtížnosti. Test obsahoval celkem 18 uzavřených úloh, přičemž některé byly rozděleny na podúlohy. Maximálně bylo možno získat 50 bodů. Dodrželi jsme čas 90 minut určený k řešení testu podle pokynů na internetu a umožnili použití kalkulačky. Test absolvovalo bez předchozího upozornění 242 studentů pěti technických fakult VŠB – TUO na přelomu dubna a května 2010 ve cvičeních z předmětu Matematika II (Fakulta stavební – FAST, Fakulta bezpečnostního inženýrství – FBI, Fakulta strojní – FS, Fakulta metalurgie a materiálového inženýrství – FMMI), případně Matematika I (Hornicko-geologická fakulta – HGF). Tento způsob zadání měl dvě negativní stránky. Studenti se na test na rozdíl od maturitní zkoušky nemohli cíleně připravit a také jejich motivace byla nulová (test nemohl ovlivnit známku z daného předmětu).
Snažili jsme se, aby zastoupení studentů jednotlivých fakult v základním souboru bylo rovnoměrné, viz obr. 1a. Výsledný průměrný bodový zisk na jednotlivých fakultách ukazuje graf na obr. 2, průměrný bodový zisk všech účastníků z VŠB-TUO byl 20,1. Podrobnější rozbor výsledků lze nalézt v [6].

a) 

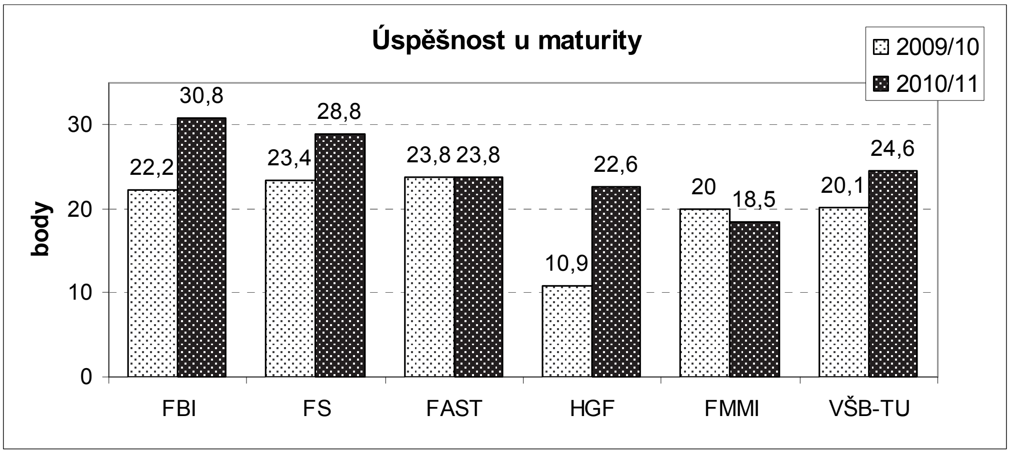
b) 

Obr. 1: Zastoupení jednotlivých fakult

3 Testování ve školním roce 2010/11

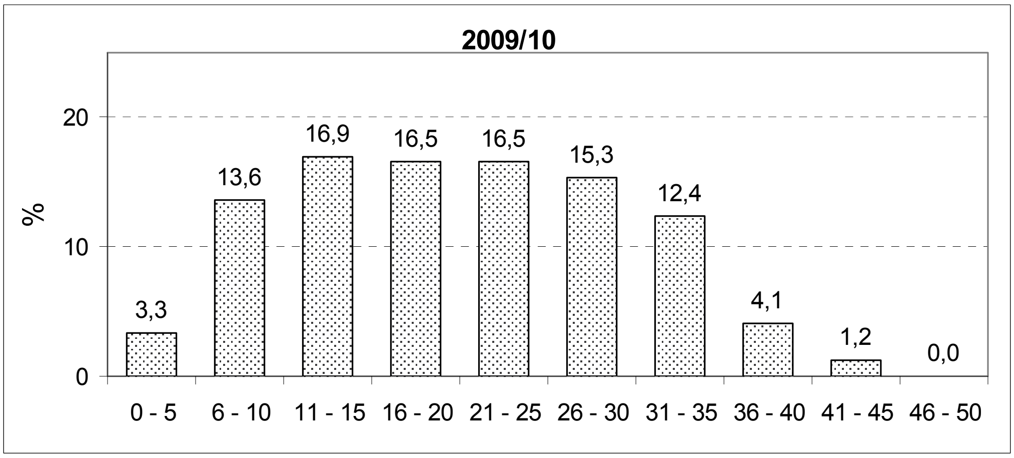
Před definitivním spuštěním státních maturitních zkoušek v České republice proběhlo na podzim školního roku 2009/10 zkušební kolo, tzv. maturita nanečisto. Využili jsme zadání této akce a znovu na náhodně vybraném vzorku studentů zjišťovali aktuální stav jejich matematických znalostí. Zvolili jsme opět základní úroveň obtížnosti. Test tentokrát obsahoval 20 úloh, přičemž některé byly opět rozděleny na podúlohy. Prvních deset úloh bylo otevřených, dalších deset uzavřených.

Maximální možný bodový zisk činil 50 bodů. Podmínky pro zadávání se nezměnily. Test absolvovalo bez předchozího upozornění 251 studentů výše uvedených pěti technických fakult VŠB – TUO v listopadu 2010. Zastoupení jednotlivých fakult v základním souboru bylo skoro rovnoměrné, viz obr. 1b. Průměrný bodový zisk na jednotlivých fakultách je uveden v grafu na obr. 2, průměrný bodový zisk všech účastníků z VŠB-TUO je 24,55. Je zřejmé, že výsledky v letošním roce byly s výjimkou FMMI a FAST o něco lepší. Přesto nemůžeme být spokojeni, protože v průměru nedosahují ani poloviny možného bodového zisku.

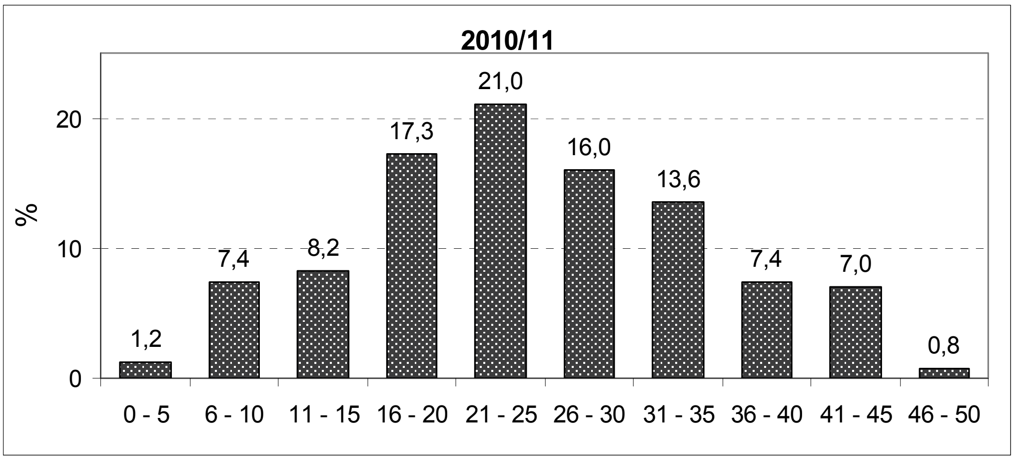


Obr. 2: Průměrný bodový zisk maturitních testů podle fakult a celkový průměr na VŠB-TUO

Grafy na obr. 3 znázorňují souhrnné výsledky testu v akademickém roce 2009/10 (obr. 3a) a v roce 2010/11 (obr. 3b). Z četností v krajních třídách je vidět, že u druhého testu ubylo studentů s nejhoršími výsledky (0-5) a přibyli 2 studenti ve třídě (46-50). Současně je zřejmé, že druhý test vykazuje výraznější maximum.



a)



b)

Obr. 3: Souhrnné výsledky (pro všechny fakulty)

4 Porovnání výsledků na jednotlivých fakultách

4.1 Testování ve školním roce 2009/10

Z grafu na obr. 2 je vidět, že s výjimkou HGF byly výsledky vcelku vyrovnané. Mimořádně nízký bodový zisk HGF můžeme zdůvodnit skutečností, že studenti HGF na rozdíl od ostatních fakult v době psaní testu ještě neabsolvovali žádnou zkoušku z matematiky (zkouška z předmětu Matematika I je až v letním semestru) a nejslabší z nich tedy nebyli ještě vytříděni. Špatný výsledek je i přesto zarážející vzhledem k tomu, že studenti HGF v zimním semestru absolvovali předmět Základy matematiky, jehož obsahem je systematické opakování středoškolské matematiky.
Nejlepších výsledků dosáhli studenti FAST, těsně následováni FS (obr. 2). Na FAST počet zájemců o studium převyšuje počet přijímaných a také studenti v zimním semestru prošli tříděním u zkoušky z předmětu Matematika I a na rozdíl od ostatních fakult také z náročného předmětu Deskriptivní geometrie. Relativně lepší výsledky na FS zřejmě ovlivnil fakt, že v zimním semestru museli všichni studenti povinně absolvovat předmět Základy matematiky (rozsah 0+2, opakování středoškolské matematiky). Předmět je ukončen klasifikovaným zápočtem, což je pro studenty motivující.

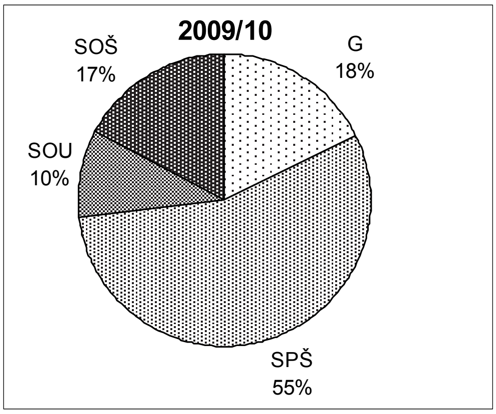
4.2 Testování ve školním roce 2010/11

Druhý test absolvovali studenti všech fakult za stejných podmínek, protože proběhl během prvního semestru jejich studia na VŠB-TU. Jistou výhodu měli studenti HGF a FS, kteří během prvního semestru musí absolvovat předmět Základy matematiky. Porovnáním grafů na obr. 2 vidíme, že k jistému zlepšení došlo na FBI, FS a HGF. Naopak horší výsledky jsou na FMMI. Tyto změny se však neprokázaly jako statisticky významné.

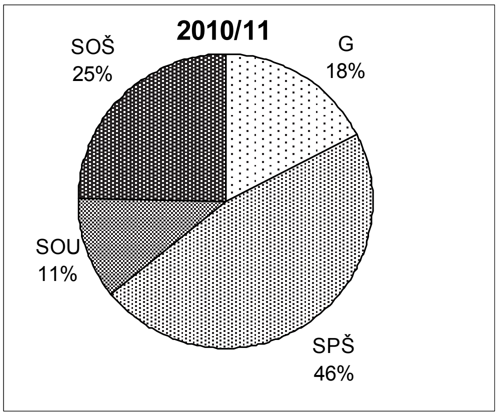
5 Výsledky na různých typech středních škol

Zkoumaný soubor jsme rozdělili podle typu střední školy, kterou studenti absolvovali, do čtyř tříd – Gymnázia (G), Střední průmyslové školy (SPŠ), Střední odborná učiliště s maturitou (SOU) a

Střední odborné školy (absolventi Obchodních akademií, Středních zdravotních škol a dalších). Podle očekávání mezi respondenty převládali na technických fakultách VŠB-TUO absolventi SPŠ [1, 3, 5]. Počet gymnazistů zůstal v posledních dvou letech konstantní, narostl počet absolventů odborných škol a učilišť na úkor absolventů průmyslovek.



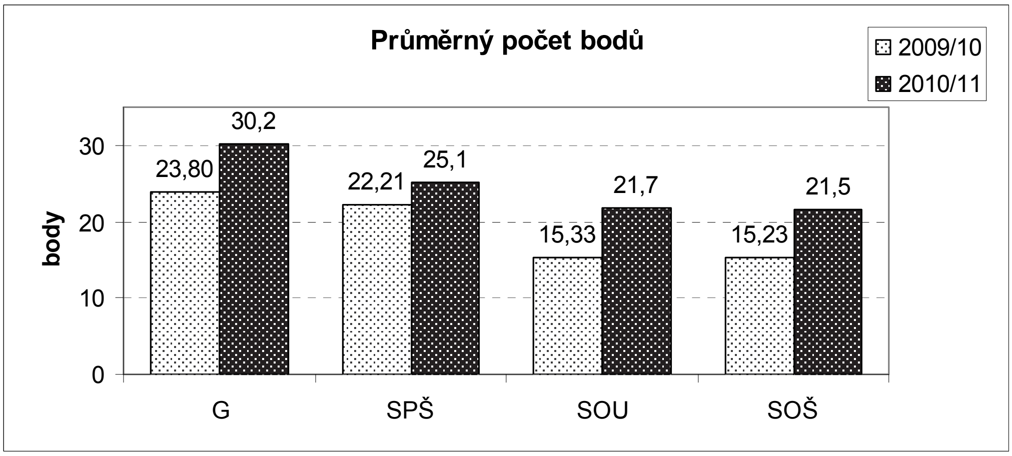
a)



b)

Obr. 4: Rozdělení základního souboru podle absolvované střední školy

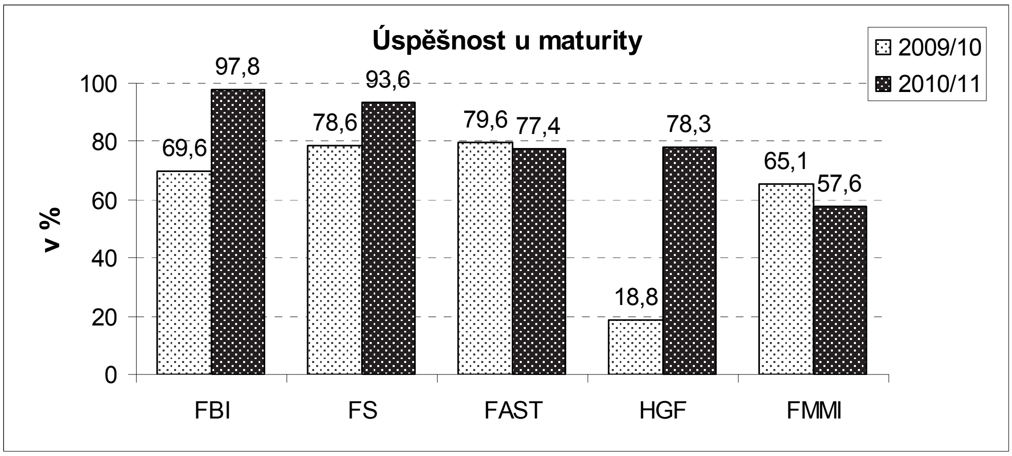
Výsledky testu na jednotlivých typech středních škol naplnily očekávání (obr. 5), [2, 4, 5]. Je zajímavé, že ve všech kategoriích je letos průměrný bodový zisk vyšší. Rozdíl ve výsledcích absolventů SOU a SOŠ v obou testech byl zanedbatelný.



Obr. 5: Průměrný počet bodů pro jednotlivé typy středních škol

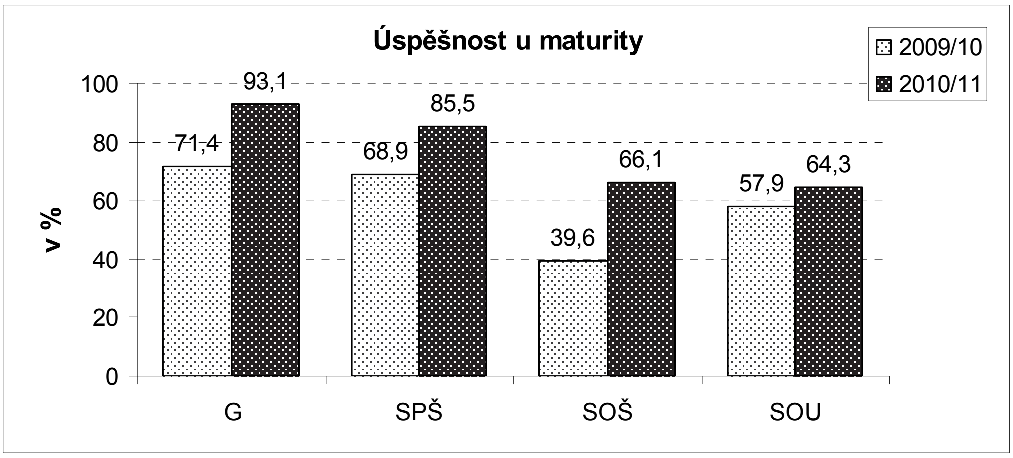
6 Úspěšnost u státní maturitní zkoušky

Pro letošní test státní maturitní zkoušky z matematiky byla jako hranice pro úspěšné složení zkoušky stanovena minimální hodnota 17 bodů. V loňském testu by podle tohoto kritéria celkem uspělo 62,66% respondentů, v letošním roce 79,68%. Úspěšnost na jednotlivých fakultách ukazuje stabilní výsledky na FAST, zlepšení nastalo na FBI, FS a zejména HGF, nevýrazné zhoršení na FMMI (obr. 6).



Obr. 6: Úspěšnost u maturity podle fakult

Úspěšnost u maturity podle typu střední školy potvrzuje známou skutečnost, že lépe jsou připraveni studenti gymnázií a středních průmyslových škol než studenti středních odborných škol a středních odborných učilišť s maturitou (obr. 7). Všechny kategorie přitom letos dosáhly lepších výsledků než v loňském roce.



Obr. 7: Úspěšnost u maturity podle typu střední školy

7 Připomínky a názory

Volba příkladů je podle nás diskutabilní. Především by se v testu v žádném případě neměly vyskytovat takové příklady, v nichž výsledek jedné úlohy ovlivní úlohy následující. Pak student zbytečně ztrácí body, protože dosadil chybné vstupní hodnoty z předchozího úkolu, i když má zcela správný postup řešení.

Problematické jsou také jednoduché příklady, které odpovídají učivu základní školy. Je jasné, že pokud má maturitní zkouška otestovat široké spektrum studentů od středních odborných učilišť po specializovaná gymnázia, je těžké najít úroveň úloh vyhovující všem. Proto by stálo za úvahu, zda nerozdělit testy podle náročnosti do více úrovní.

Dalším problémem je zařazení příkladů, jejichž výsledek nevyžaduje nic jiného než výpočet na kalkulačce, případně odečtení údajů z grafu nebo dosazení do vzorce a následující výpočet na kalkulačce. Obsahem maturitní zkoušky by měly být také úlohy, které ověří nejen základní znalosti středoškolské matematiky, ale i to, jak student umí logicky myslet.

Testy s volitelnou odpovědí (všechny úlohy u loňského testu, úlohy 11-20 u letošní maturity nanečisto) nejsou pro matematiku vhodné. Řada studentů úlohy neřešila, ale jen tipovala výsledky. Daleko lepším způsobem ověřování vědomostí z matematiky jsou otevřené úlohy, přestože jejich hodnocení je náročnější.

Pokyny pro opravující pedagogy v testu 2009/10 uvedené na internetu byly nedostatečné. Hodnocení různými pedagogy se lišilo až o 6 bodů. Abychom dosáhli co největší objektivity při zpracování výsledků, byly nakonec všechny testy hodnoceny jen jedním pedagogem. Po této zkušenosti opravoval i v testu 2010/11 danou úlohu vždy jen jeden pedagog. Pro státní maturitní zkoušku by v zájmu objektivity výsledků měly být pokyny detailně rozpracovány, aby se zcela vyloučil subjektivní názor opravujícího zejména u otevřených úloh.

Hodnocení úloh bylo podle našeho názoru rovněž problematické. Především nebyl vůbec brán zřetel na správný postup řešení, což je v matematice podstatné. Rozhodující byl jen numerický výsledek. Některé příklady byly hodnoceny jen 0 nebo 3 body ve vylučovacím smyslu a nebylo možno přidělit body za částečné vyřešení. V jiných příkladech se 3 podúlohy stejné obtížnosti hodnotily 4 body (rok 2009/10) nebo naopak 4 podúlohy stejné obtížnosti se hodnotily 3 body (všechny 4 správně za 3 body, 3 správně za 1 bod, zbytek 0 bodů), což je dost nelogické.

8 Závěr

Znalosti středoškolské matematiky, které naši studenti předvedli v maturitním testu z matematiky, jsou slabé. Mnozí studenti nebyli schopni ani dosadit do vzorce. Řada studentů jen zaškrtávala odpovědi, aniž se snažila úlohy řešit.

Na vysoké škole technického typu je nutné rutinní zvládnutí základních matematických dovedností (úprava algebraických výrazů, počítání se zlomky, mocninami, odmocninami, logaritmy, elementární funkce, analytická geometrie přímky a kuželoseček v rovině).

Přísloví „méně někdy znamená více“ je v úvahách o osnovách matematiky na středních školách zcela na místě. Nepotřebujeme, aby se na běžných středních školách vyučoval diferenciální a integrální počet, případně počet pravděpodobnosti. Bylo by vhodnější disponibilní časový prostor věnovat důkladnému procvičení výše uvedených témat.

Literatura

[1] DOLEŽALOVÁ, J. - KREML, P.: *Souvislost studijních výsledků na střední škole s úspěšností studia na vysoké škole technické.* In Sborník 26. konference VŠTEZ, s. 47-52, Bohdaneč 2000, ISBN 80-7015-732-1.

[2] DOLEŽALOVÁ, J. - KREML, P.: *O některých problémech výuky matematických předmětů na VŠB – TU Ostrava.* Proceedings of 3rd Scientific Colloqium, s. 192-196, Praha 2001, ISBN 80-7080-423-8.

[3] DOLEŽALOVÁ, J. - KREML, P.: *Jsou studenti připraveni ke studiu na technických fakultách VŠB-TU Ostrava?* Sborník 28. mezinárodní konference VŠTEZ, s. 73-80, EDIS - vydavatelstvo ŽU, Žilina 2004, ISBN 80-8070-287-X.

[4] DOLEŽALOVÁ, J. - KREML, P.: *Analýza problémů způsobujících předčasné ukončení studia na FS a FAST VŠB – TU Ostrava.* 4th International Conference APLIMAT, Part I, str. 99-108, Bratislava 2005, ISBN 80-969264-1-1.

[5] DOLEŽALOVÁ, J. - KREML, P.: *Úroveň uchazečů o studium na technických fakultách VŠB-TU Ostrava.* Sborník 29. mezinárodní konference VŠTEZ, s. 43-48, UTB – Academia centrum Zlín, Zlín 2006, ISBN 80-7318-450-8.

[6] BOHÁČ, Z. – DOLEŽALOVÁ, J. – KREML, P.: *Maturita z matematiky nanečisto na VŠB-TU Ostrava*. In Sborník 19. mezinárodní konference Moderní matematické metody v inženýrství. Dolní Lomná, s. 9-13, ISBN 978-80-248-2342-3.

[7] Ilustrační testy a zadání předmětů společné části maturitní zkoušky: http://www.novamaturita.cz/index.php?id_document=1404033125

Kontaktní adresa

Zdeněk Boháč, doc. RNDr. CSc.,
VŠB – TU Ostrava, Katedra matematiky a deskriptivní geometrie, 17. listopadu 15,
708 33 Ostrava – Poruba, +420 597 324 182,
e-mail: zdenek.bohac@vsb.cz

Jarmila Doležalová, doc. RNDr. CSc.,
VŠB – TU Ostrava, Katedra matematiky a deskriptivní geometrie, 17. listopadu 15,
708 33 Ostrava – Poruba, +420 597 324 185,
e-mail: jarmila.dolezalova@vsb.cz

Pavel Kreml, doc. RNDr. CSc.,
VŠB – TU Ostrava, Katedra matematiky a deskriptivní geometrie, 17. listopadu 15,
708 33 Ostrava – Poruba, +420 597 324 175,
e-mail: pavel.kreml@vsb.cz

Journal of Applied Mathematics

volume IV (2011), number I

PERIODIC FUNCTIONS IN SYSTEM *MATHEMATICA*

DOBRAKOVOVÁ Jana, (SK), ZÁHONOVÁ Viera, (SK)

Abstract. In this contribution we want to show several methods of defining, drawing graphs and other abilities of manipulating with periodic functions, provided by the programming system *MATHEMATICA***.** We also compare advantages and disadvantages of individual methods.

Key words. periodic functions, *MATHEMATICA*, Piecewise, Round.

1 Introduction

The main aim of this article is to present various possibilities and ways how to work with periodic functions, mainly what about their definitions, graphs, as well as their differentiation, integration and limits. We dealt with this issue namely in lectures and seminars about the Fourier series. In the education we use the programming system *MATHEMATICA*. We solved the problem, how to define and handle in this system a real periodic function, with given period, and defined usually by a formula (an analytic expression), mostly an elementary function, just only on an interval of the length of one period.

We have used this programming system at our faculty for several years not only in some subjects in basic bachelor programs, but also in several special subjects of advanced mathematics in master graduate programs. Students, who learned to work with this system at the begining of their study, use it for the elaboration of their individual works, presented in various competitions, bachelor works, diploma projects and others. We have many positive experiences ([1] - [7]) and reactions, not only from students.

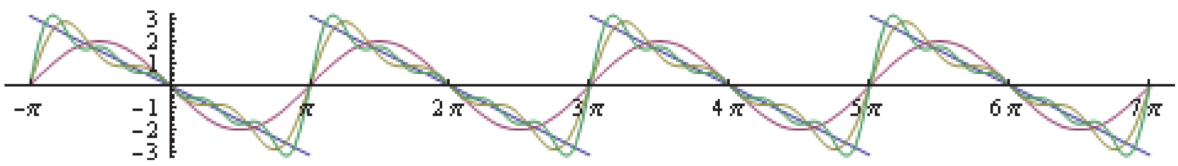
Initially we started to work with the version *MATHEMATICA* 2, then we gradually introduced to education new upgraded versions and we utilized all available innovations appropriate and applicable to our study. At present time we use the version *MATHEMATICA* 7.

2 Periodic functions and Fourier series

In paper ([2]) we discussed utilization of the programming system *MATHEMATICA* in teaching Fourier series. This subject is especially appropriate for using the system, not only because of very frequent lengthy numerical calculation of Fourier coefficients, but also, and mainly, for graphical interpretation and visualization of the fact, that approximation of the original function is improving with increasing degree of Fourier polynomials.
Consider, for example, a simple periodic function f_1, defined as follows:
$f_1 : y = -x,$ for $x \in (-\pi, \pi\rangle$, and f_1 is periodic, with the period $T = 2\pi$.
In such a case it is no problem, neither for students, to find its Fourier series. Since f_1 is an odd function, it is sinus series

$$\sum_{n=1}^{\infty} \frac{2(-1)^n}{n} \sin nx$$

Now one can easily express, for example, the first, third, or fifth Fourier polynomial, but to draw their graphs is not so trivial, excepting graph of the first polynomial. Drawing these graphs, in order to compare them with graph of original function, without any support of computer, is very labourious and too lengthy. But with the aid of the programming system it is a trivial task. In the following picture there are graphs of those three polynomials together with graph of the function f_1.



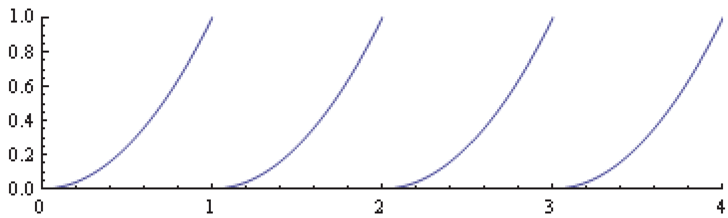
It is appropriate, that individual graphs in the picture are colored differently, what gives better visualization of the fact, that approximation of original function gradually improves, with increasing degree of Fourier polynomials. We have an annual repetitive experience, that the students reaction on these graphs is unambiguously positive. It is likely that some of students fully understand the concept of Fourier series convergence and approximation of given function just thanks these graphs. But in drawing them we met one, at the first sight trivial problem.
In order to achieve the objectives, it is appropriate, if graphs are drawn on an interval of a length equal to a multiple of the period.
And just that reverts to the original question and also to the main target of this article. It is an answer to the question in what way, as simply as possible, to define in the system *MATHEMATICA* a periodic function, given in a similar way as the above mentioned function f_1. It means by a formula on the interval of the length of one period. Further, we want such a definition, which enables to draw graph of the function on an interval of any length (similarly, as in previous figure is graph of the function f_1 on four periods length interval), to to calculate its values at arbitrary points, to differentiate and integrate it, eventually to calculate its limits at points of discontinuity and overall, to manipulate it as any else elementary function.

Obviously, it does not concern functions, defined on their domains of definition (usually the set of all real numbers) with the aid of periodic functions, like, for example, basic trigonometric functions sin, cos, tan, cot. Their periodicity than follows from the periodicity of these trigonometric functions. For instance $f : y = |sin\, x|$, $g : y = sin\, x \cos x$, and likewise. For such functions problems of this kind do not arise.

3 How to draw graph of a periodic function

If we are only interested in drawing graphs of periodic functions, then we need no special commands. It is enough if we „shift“ the given function along x-axis so many times, as required.
We illustrate that with the following example.
Let us draw graph of the function f_2, defined as follows, on the interval $\langle 0,4)$. $f_2 : y = x^2$, for $x \in \langle 0,1)$, and f_2 is periodic, with the period $T = 1$.
We can use the following simple procedure:

```
Show[Table[Plot[(x - a)^2, {x, a, a + 1}, PlotRange -> {{0, 4}, {0, 1}},
  AspectRatio → Automatic], {a, 0, 3}]]
```



Of course, in this way we have not defined the function f_2, we have only drawn its graph on the given inteval.

4 The command Piecewise

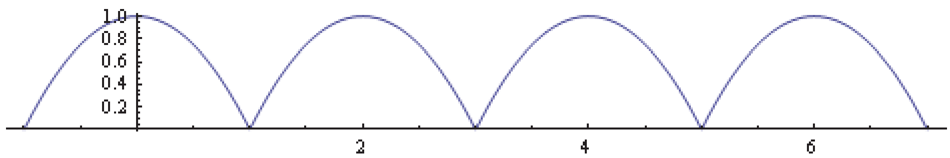
If we want not only to draw graph of a similar function as above, but also to define it, we can use the command **Piecewise**. As seen already from the title of this command, by means of the command **Piecewise** are functions defined by parts.
Consider a function
$f_3 : y = 1 - x^2$, for $x \in \langle -1,1)$, and f_3 is periodic, with the period $T = 2$
We apply the command **Piecewise** to define the function f_3 on an interval of four periods length, say on $\langle -1,7)$.

```
f3[x_] = Piecewise[{{1 - x^2, -1 ≤ x ≤ 1}, {1 - (x - 2)^2, 1 ≤ x < 3},
   {1 - (x - 4)^2, 3 ≤ x < 5}, {1 - (x - 6)^2, 5 ≤ x < 7}}];
```

Deficiency of this definition consists in the fact, that the function f_3 is correctly defined just only on the interval $\langle -1,7)$. Outside this interval of definition by the command **Piecewise** is the given function considered as zero. In general, by means of this command, any non-zero function can be defined only on an interval with a finite length.

Now, since the function f_3 is defined (despite the fact that correctly only on interval $\langle -1,7)$), we are able to draw its graph on the interval $\langle -1,7)$, without any problems, by means of the usual command **Plot**.

```
Plot[f3[x], {x, -1, 7}, AspectRatio → Automatic]
```



Further, we can perform all usual mathematical operations, as differentiation, integration, calculating limits. Let us start with differentiation.

```
D[f3[x], x]
```

$$\begin{cases} 0 & x < -1 \\ -2x & -1 < x < 1 \\ 4-2x & 1 < x < 3 \\ 8-2x & 3 < x < 5 \\ 12-2x & 5 < x < 7 \\ 0 & x > 7 \\ \text{Indeterminate} & \text{True} \end{cases}$$

It means that function is differentiated by parts and its derivative does not exist at points $-1, 1, 3, 5, 7$. It is clear that outside the interval $\langle -1,7)$, as derivative of a zero function, it is equal to zero. What about integrating, the situation is simile.

```
Integrate[f3[x], x]
```

$$\begin{cases} 0 & x \le -1 \\ \frac{2}{3} + x - \frac{x^3}{3} & -1 < x \le 1 \\ \frac{8}{3} - 3x + 2x^2 - \frac{x^3}{3} & 1 < x \le 3 \\ \frac{62}{3} - 15x + 4x^2 - \frac{x^3}{3} & 3 < x \le 5 \\ \frac{212}{3} - 35x + 6x^2 - \frac{x^3}{3} & 5 < x \le 7 \\ \frac{16}{3} & \text{True} \end{cases}$$

The result we achieved is also calculated by parts and represents one of infinitely many antiderivatives of the given function, The integration constants in each interval are chosen so that this antiderivative is continuous.
And finally, let us calculate definite integral of f_3 on any interval of two periods length, say on $\langle -1,3\rangle$.

```
Integrate[f3[x], {x, -1, 3}]
8
—
3
```

It can be easily verified corretness of the result.
Therefore we see that a function, defined by means of the command **Piecewise** is in a sence „complete". The only disadvantage of this definition consists in the fact that any non-zero function can be defined only on an interval with a finite length.

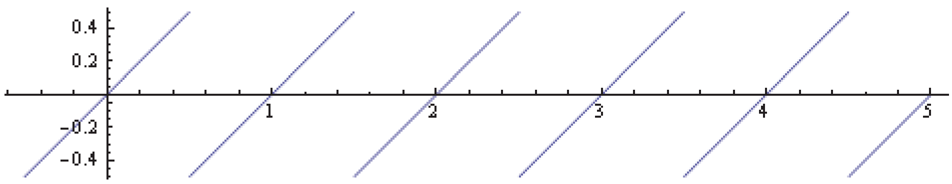
5 The command Round

The basic meaning of the command **Round** is to round off given number to the closest integer. There is a problem if the closest integer does not exist, in other words if there are two integers with the same least distance from the given number. In this case as the closest integer is always chosen the even number, irrespective whether greater or less than the given number. For instance:

```
Round[{-3.7, Pi, 1.5, 2.5}]
{-4, 3, 2, 2}
```

Now we will use this command to define a periodic function. First we show its application in a simple example:

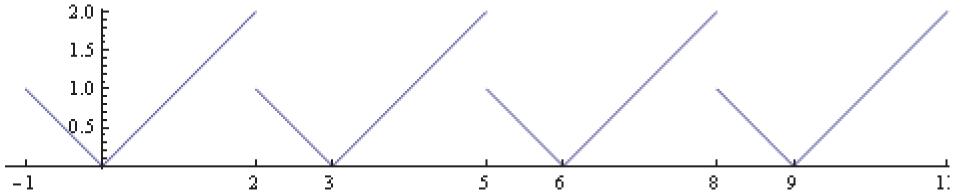
```
Plot[x - Round[x], {x, -0.5, 5}, AspectRatio → Automatic]
```



We have obtained graph of a function f_4, identical on the interval $\langle -0{,}5,\ 0{,}5\rangle$ with the function $y = x$ and else, excepting points of „jumps" (for example, because of the operation **Round** , it is defined : $f_4(-0{,}5) = -0{,}5 \neq 0{,}5 = f_4(0{,}5)$), is periodic, with the period $T = 1$.
In the similar way we can get periodic continuation for arbitrary elementary function, defined on an interval of a finite length.
Let us define and draw graph of the function
$f_5 : y = |x|$, for $x \in \langle -1,2)$, and f_5 is periodic, with the period $T = 3$,

with the aid of the command **Round** :

```
f5[x_] = Abs[x - 3 * Round[(x - 3 / 2 + 1) / 3]];
Plot[f5[x], {x, -1, 11}, AspectRatio → Automatic,
 Ticks → {{-1, 0, 2, 3, 5, 6, 8, 9, 11}, Automatic}]
```



This definition enables to calculate value of the function at any real number and to draw its graph on any bounded interval, which can be easily verified. For example:

```
f5[104]
```

2

Unfortunately, in the programming system it is not possible to differentiate and to integrate (to find its antiderivative) a function, defined in such a way. On the other hand, it is no problem to calculate its definite integral and limits, or one-sided limits at any point.

```
Integrate[f5[x], {x, -1, 5}]
```

5

```
Limit[f5[x], x → 2, Direction → -1]
```

1

```
Limit[f5[x], x → 2, Direction → +1]
```

2

6 Other ways

This method of defining periodic functions, based on the command **Round,** can be generalized. If *a* is an arbitrary real number, *T* is an arbitery positive number and g is an arbitrary elementary function defined on the interval $\langle a, a+T)$, then the function $f : y = g(x)$, for $x \in \langle a, a+T)$, and f is periodic, with the period T,
can be defined as follows. First we predefine

```
f[x_] = g[x - T * Round[(x - T / 2 - a) / T]];
```

Then it suffices to define *a*, *T* and *g*. In the case of foregoing function f_5 it has the form

```
g[x_] = Abs[x]; T = 3; a = -1;
```

In this way we could define and draw graphs of all previously mentioned periodic functions.

Another type of periodic functions which may occur is a periodic function defined on an interval of one period length in several ways. Let us consider, for instance, a periodic function f with the period $T = 4$, on the interval $\langle 0, 4)$ defined by parts, as follows:

$$f : y = \begin{cases} x, & 0 \le x < 1, \\ 1, & 1 \le x < 3, \\ 4 - x, & 3 \le x < 4 \end{cases}$$

We will define it combining both commands, **Piecewise** and **Round.** First we define by means of the command **Piecewise** the „auxiliary" function *g* and then also constants *a* and *T:*

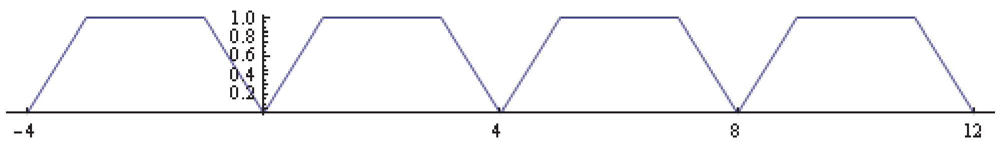
```
g[x_] = Piecewise[{{x, 0 ≤ x < 1}, {1, 1 ≤ x < 3}, {4 - x, 3 ≤ x < 4}}];
T = 4;
a = 0;
```

Further we apply the generalized definition for f:

```
f[x_] = g[x - T * Round[(x - T / 2 - a) / T]];
```

In this way is the function f defined for all real numbers, we are able to draw its graph, to calculate its definite integral on any interval and also, by means of the function *g,* to find its derivative and antiderivative, similarly as we found out above for the function f_3.

```
Plot[f[x], {x, -4, 12}, AspectRatio → 1 / 10,
 Ticks → {{-4, 0, 4, 8, 12}, Automatic}]
```



References

[1] DOBRAKOVOVÁ. J.,KOVÁČOVÁ M., ZÁHONOVÁ V.:*MATHEMATICA 5.2 pre stredoškolských učiteľov, študijné materiály* , STU Bratislava, 2006.

[2] DOBRAKOVOVÁ., J.,ZÁHONOVÁ V.: *Fourierove rady a systém MATHEMATICA*, XXVI. Mezinárodní kolokvium o řízení osvojovacího procesu, Brno, 2007, CD-ROM.

[3] DOBRAKOVOVÁ J.: *Ako definovať integrál*, XX.Mezinárodní kolokvium o řízení osvojovacího procesu, Vyškov (2002), 81-84.

[4] DOBRAKOVOVÁ J., ZÁHONOVÁ V.: *Line integrals and MATHEMATICA*. In XXIV International Colloquium on the Acquisition Process Management, University of Defence, Brno 2006, ČR, ISBN 80-7231-139-5, str. 34, CD-ROM

[5] KOVÁČOVÁ M.: *MATHEMATICA a CALCCENTER 3: porovnanie dvoch programových systémov*, Department of Mathematics report series, Volume 13, 2005, University of Doth Bohemia,České Budějovice, ČR, ISSN 1214-4681, str. 73-78.

[6] ZÁHONOVÁ V.: *Možnosť využitia programového systému Mathematica pri výučbe matematiky v novokoncipovanom štúdiu na SjF STU v Bratislave,* Aplimat, Bratislava (2002), 449-452.

[7] ZÁHONOVÁ V.: *Niektoré grafické možnosti programového systému MATHEMATICA a možnosti jeho využitia na stredných školách*. In Department of Mathematics report series, Volume 13, 2005, University of Doth Bohemia,České Budějovice, ČR, ISSN 1214-4681, str. 229-232.

Current address

Dobrakovová Jana, doc., RNDr., CSc.
Institute of Natural Sciences, Humanities and Social Sciences
Mechanical Engineering, Slovak University of Technology,
Námestie slobody 17
812 31 Bratislava
e-mail: jana.dobrakovova@stuba.sk

Záhonová Viera, RNDr., CSc.
Institute of Natural Sciences, Humanities and Social Sciences
Mechanical Engineering, Slovak University of Technology,
Námestie slobody 17
812 31 Bratislava
e-mail: viera.zahonova@stuba.sk

Journal of Applied Mathematics

volume IV (2011), number I

A BUG IN MY HAND-HELD CAS CALCULATOR (AND A BAG OF SOLVERS)[1]

GUERRERO-GARCíA Pablo, (E)

Abstract. Learning theoreticians tell us that those examples in which unexpected results are obtained have a high degree of motivational power for the students. In this work a differential problem with periodic boundary conditions is presented. Apart from getting some symbolic solvers into trouble, the application of finite difference techniques leads to a system of linear equations whose coefficient matrix is symmetric positive definite and strictly diagonally dominant by rows, but using the reduced row echelon form of both the physical hand-held CAS/graphics calculator and its emulator software yields catastrophic results in spite of the problem being numerically well conditioned. Thus, the existence of a (currently unfixed) bug in such a routine has been revealed, and some related exercises and a description of the way it was discovered have also been included.

Key words and phrases. Calculators, Bugs, Symbolic solvers, Periodic boundary conditions, Linear algebra, Direct numerical methods, Classroom techniques.

Mathematics Subject Classification. Primary 97U70; Secondary 97D40, 65C20.

1 Two approaches from within Matlab

In order to shed some light on the second and third additional questions of my A052 TIME 2010 lecture [5]:

> Are you still thinking that symbolic computations are nowadays powerful enough to dispense with numerical computations of any kind?
>
> How large are the systems of linear equations you have ever faced with? Can they be solved with your favorite CAS/graphics calculator?

[1]Technical Report MA-10/02, 30 September 2010, http://www.matap.uma.es/investigacion/tr.html. Talk to be presented at the 10th International Conference on Applied Mathematics, section “New trends in mathematics education”, Bratislava (Slovakia), 1–4 Febr 2011.

let us consider the following problem:

> Determine/approximate a function such that the opposite of its second derivative added to 9 times the function yields 12 times the sine of the independent variable; it must also satisfy that both the function and its derivative yield the same at 0 and at 2π.

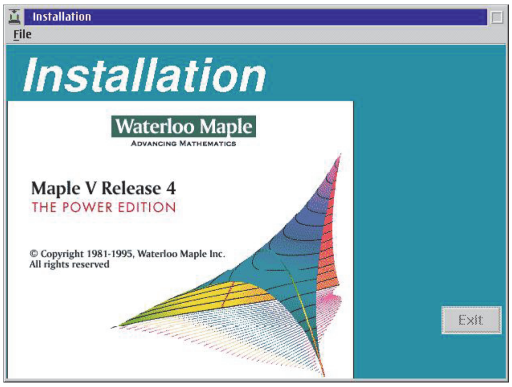
This problem can also be posed with a more technical wording as:

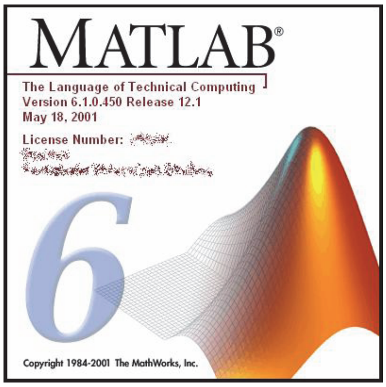
> If the symbolic solver does not give an analytic expression in terms of elementary functions, then apply finite difference techniques to approximately solve the two-point boundary-value linear problem
>
> $$-y''(t) + 9y(t) = 12\sin(t), \quad y(0) = y(2\pi), \quad y'(0) = y'(2\pi)$$
>
> with periodic Dirichlet and Neumann boundary conditions with period 2π and $0 \leq t \leq 2\pi$.

Since kernel of CAS Maple V Release 4 is available through Matlab v6.1 Symbolic Toolbox, we can try both a symbolic approach and a numerical approach from within the Matlab environment. But the symbolic approach in Matlab yields:

```
>> dsolve('-D2y+9*y=12*sin(t)','y(0)=y(2*pi)','Dy(0)=Dy(2*pi)')
Warning: Explicit solution could not be found.
```

On the other hand, the numerical approach in Matlab could be as follows:






```
>> a=0; b=2*pi; n=20;
>> h=(b-a)/n; t=a+h:h:b; % n=length(t)
>> A=diag(-ones(1,n-1),-1)+diag((2+9*h^2)*ones(1,n))+diag(-ones(1,n-1),1);
>> A(1,n)=-1; A(n,1)=-1; c=12*h^2*sin(t');
>> y=A\c;
>> plot([a t b+h],[y(end) y' y(1)]);
```

It is important to note that the very same results are obtained by using as fifth command above

```
>> y=rref([A c]); y=y(:,end);
```

To simplify further references, let us define a MATLAB function named `Thomas` to return both `A` and `c` in terms of `n`:

```
function [A,c]=Thomas(n)
a=0; b=2*pi; h=(b-a)/n; t=a+h:h:b; % n=length(t)
A=diag(-ones(1,n-1),-1)+diag((2+9*h^2)*ones(1,n))+diag(-ones(1,n-1),1);
A(1,n)=-1; A(n,1)=-1; c=12*h^2*sin(t');
```

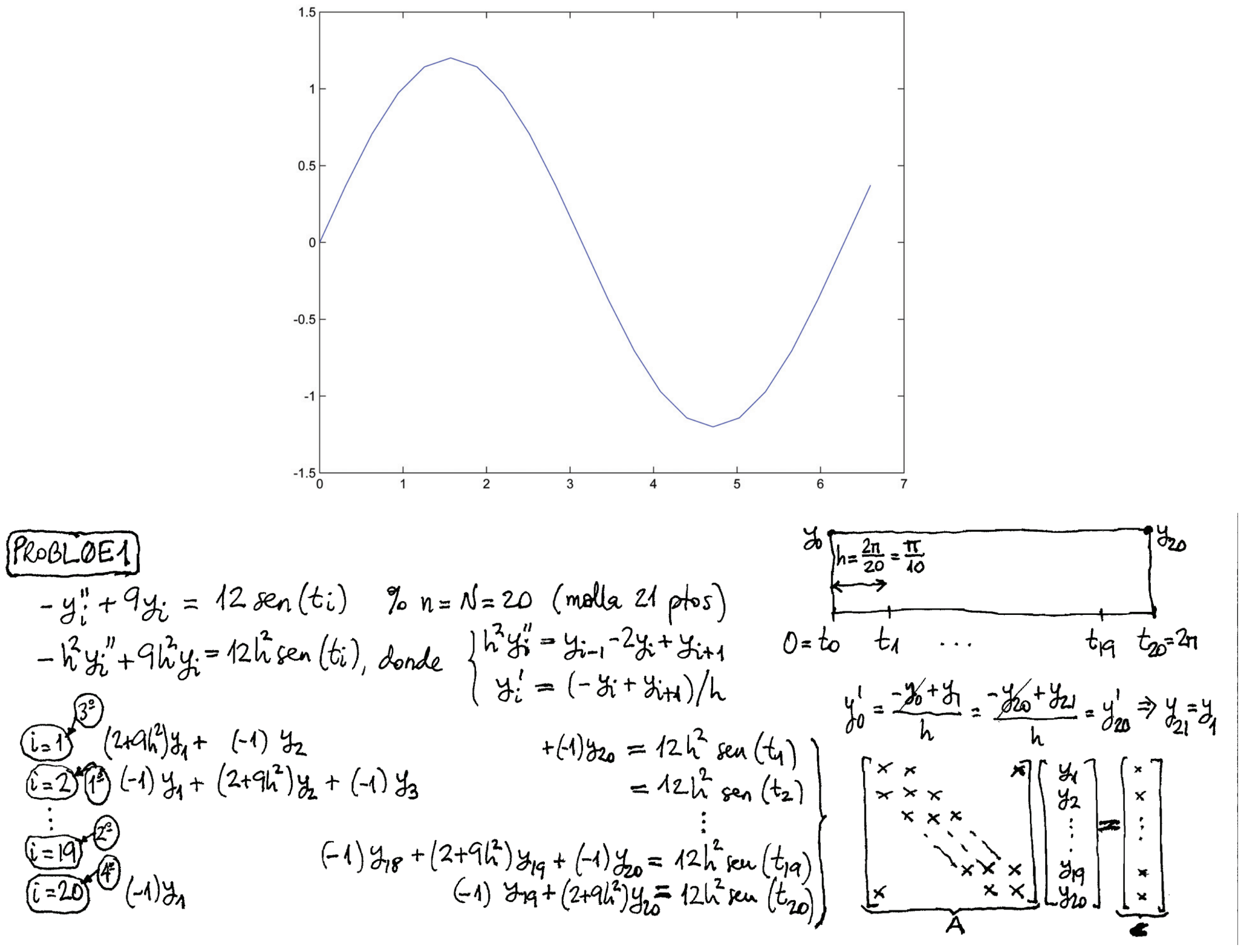


Figure 1: Numerical approximation and finite difference schema for $n = 20$

2 Explaining the numerical approach

A plot of the numerical approximation (cf. top of Figure 1) obtained leads us to think that the analytical expression of the solution must be quite simple. The numerical approach given above is actually a classical finite difference technique; if we consider the following numerical differentiation formulae

$$y(t_i) \approx y_i, \qquad y'(t_i) \approx \frac{y_{i+1} - y_i}{h}, \qquad y''(t_i) \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}$$

we obtain a system of linear equations $Ay = c$ as described in the schema of the bottom of Figure 1.

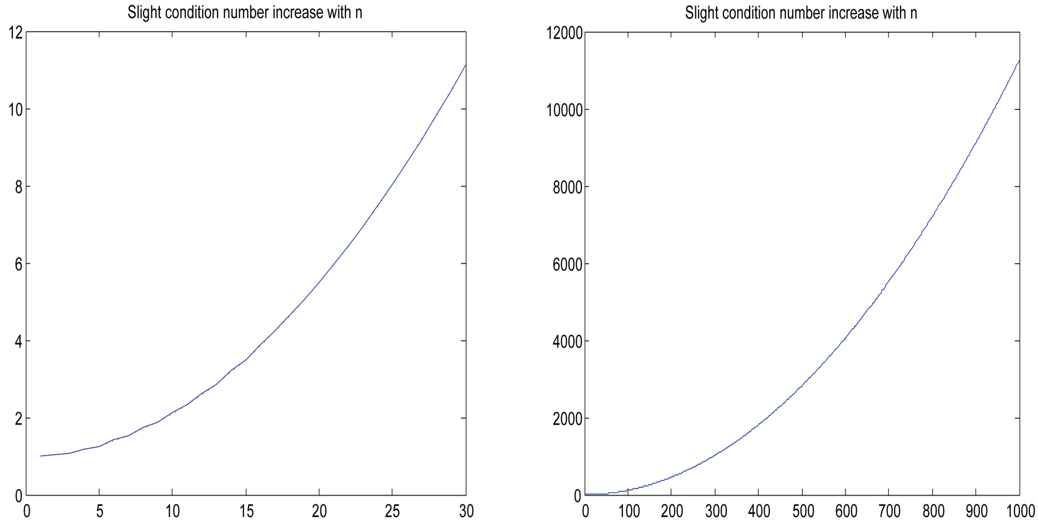
Hence this "numerical-friendly" approach amounts to solving an $n \times n$ system of linear equations whose coefficient matrix A turns out to be both symmetric positive definite and strictly diagonally dominant by rows! It is a suitable test problem since n can be as large as you want (e.g., $n = 1000$).

In order to check the features of the coefficient matrix we have designed the following Matlab procedure:

```
function ThomasNC(nmax,pnorm)
if nargin < 2, pnorm = 2; end;
for n = 2:nmax
  [A,c] = Thomas(n); [R,p] = chol(A); IsPDmatA = ~p; IsDDmatA = 1;
  for k = 1:n
    IsDDmatA=IsDDmatA-(abs(A(k,k))<=sum(abs(A(k,setdiff(1:n,k)))));
  end;
  if ~IsPDmatA | ~IsDDmatA
    error(['Symmetric matrix but not positive definite ', ...
           'or strictly diagonal dominant by rows!']);
  end;
  condnum(n) = cond(A,pnorm);
end; condnum(1) = 1;
fprintf('Highest condition number was %10.2f',max(condnum));
phideg2 = polyfit([1 nmax/2 nmax],condnum([1 nmax/2 nmax]),2)
plot(polyval(phideg2,1:nmax),':r');
hold on; plot(condnum,'b'); hold off;
title('Slight condition number increase with n');
```

By plotting `ThomasNC(30)` and `ThomasNC(1000)` above, namely the increase of the condition number $\|A\|_2\|A^{-1}\|_2$ for the linear equation system $Ay = c$ with coefficient matrix $A \in \mathbb{R}^{n\times n}$, we have checked that no heavy worsening appears as n increases (asymptotically, it seems to be $\mathcal{O}(n^2/100)$, or more specifically $1.13n^2/100 + 0.9887$) so problem is not ill-conditioned at all!

```
>> ThomasNC(30)
Highest condition number was      11.13
phideg2 =    0.0114   -0.0028    0.9915

>> ThomasNC(300)
Highest condition number was    1014.21
phideg2 =    0.0113    0.0001    0.9886
```




```
>> ThomasNC(500)
Highest condition number was    2815.48
phideg2 =    0.0113    0.0001    0.9887

>> ThomasNC(1000)
Highest condition number was   11258.91
phideg2 =    0.0113    0.0000    0.9887
```

3 Numerical approach with ClassPad 330

The following ClassPad 330 OS Version 3.04.5 program (May 2010) can be used to generate the table $\{t_i, y_i\}$ in terms of the number n of nodes; after running the program, both t and y will be available in the Main window for a subsequent use:

```
SetDecimal
2pi/n => h
fill(0,n,n) => A
listToMat(seq(12*h^2*sin(i*h),
          i,1,n,1)) => c
For 1 => i To n
  For 1 => j To n
    If i=j :Then
      2+9*h^2 => A[i,j]
    ElseIf abs(i-j)=1 :Then
      -1 => A[i,j]
    IfEnd
  Next
Next
-1 => A[n,1]
-1 => A[1,n]
seq(i*h,i,1,n,1) => t
matToList(A^(-1)*c,1) => y
For 1 => i To n
  Print {t[i],y[i]}
Next
```

It is well-known [3, 6, 1] that the solution of the system of linear equations can be obtained in a better numerical way than doing `matToList(A^(-1)*c,1)` with the following command

```
matToList(subMat(rref(augment(A,c)),1,n+1,n,n+1),1)
```

Surprisingly, it should be pointed out that after performing this change:

- The emulator ClassPad Manager Professional Version 3.04.5 (May, 2010) only works properly for $n \leq 7$ and it overflows for $n \geq 8$.
- The physical calculator works properly for $n \leq 19$ but it yields numerically incoherent results for $20 \leq n \leq 26$ and stops abruptly after 150 seconds with the message “Insufficient Memory” for $n \geq 27$.

Computing times (in seconds) in terms of n for the physical calculator ClassPad 330 OS v3.02 (those for OS v3.04.5 are similar) were:

n	4	5	6	7	8	9	10	11
seconds	5.27	7.38	10.81	16.88	24.47	40.35	60.69	66.81
coherence	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

n	12	13	14	15	16	17	18	19
seconds	83.16	99.87	105.88	112.69	128.47	152.10	154.97	180.34
coherence	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

n	20	21	22	23	24	25	26	27
seconds	177.44	181.25	190.68	207.75	233.56	233.31	266.47	229.78
coherence	No	No	No	No	No	No	No	?

Task 3.1 *You can perform the very same numerical computations with another CAS/graphics calculator like* TI-Nspire CAS *or* TI-Nspire.

As regards the numerical inaccuracies, let us show a table for $n = 20$ with nodes t_i at left-most column, approximations y_i obtained by inverse premultiplying at central column, and numerically incoherent approximations $\tilde{y}_i$ obtained with `rref` at right-most column:

```
0.31415926535898    0.37112462859811     7.10611516E-14
0.62831853071796    0.70592099277170    -0.3659850585
0.94247779607694    0.97161689173203    -1.753206564
1.25663706143592    1.14220415967668    -5.655900361
1.57079632679490    1.20098452626784   -15.70891527
1.88495559215388    1.14220415967668   -40.89995283
2.19911485751286    0.97161689173203  -103.5473485
2.51327412287183    0.70592099277170  -259.1303284
2.82743338823081    0.37112462859811  -645.585698
3.14159265358979    0.00000000000000 -1605.857843
3.45575191894877   -0.37112462859811 -3992.556335
3.76991118430775   -0.70592099277170 -9925.334483
4.08407044966673   -0.97161689173203 -24673.73773
4.39822971502571   -1.14220415967668 -61337.98555
4.71238898038469   -1.20098452626784 -152485.4557
5.02654824574367   -1.14220415967668 -379079.1427
5.34070751110265   -0.97161689173203 -942394.209
5.65486677646163   -0.70592099277170 -2342803.54
5.96902604182061   -0.37112462859811        0
6.28318530717959   -0.00000000000000        0
```

Hence we can conclude that it seems to be a bug on the `rref` implementation of the ClassPad 330! From learning theory [2, 8], we know that a bug example is a good motivational source [9] for our students.

4 Symbolic solvers

The symbolic approach with the ClassPad 330 (by using Main/Action/Advanced/dSolve)

```
> dSolve(-y''+9y=12sin(t),t,y)
{y=e^(3*t)*const(2)+e^(-3*t)*const(1)+6*sin(t)/5}
```

or with other CAS such as Maxima v5.21.1:

```
(%i1) ode2(-'diff(y,t,2)+9*y=12*sin(t),y,t)
(%o1) y = (6/5)*sin(t) + %k1 %e^(3t) + %k2 %e^(-3t)
```

does not allow us to specify periodic boundary conditions, but they can be imposed "manually" to obtain the constants by solving a system of linear equations with 2 unknowns with regular coefficient matrix and zero rhs. Both deliver a simpler expression than that obtained from within Matlab with Maple V Release 4 without the boundary conditions:

```
>> pretty(simple(dsolve('-D2y+9*y=12*sin(t)')))
(- 1/5 exp( 3t) cos(t) + 3/5 exp( 3t) sin(t)
 + 1/5 exp(-3t) cos(t) + 3/5 exp(-3t) sin(t)) cosh(3t) +
(  1/5 exp( 3t) cos(t) - 3/5 exp( 3t) sin(t)
 + 1/5 exp(-3t) cos(t) + 3/5 exp(-3t) sin(t)) sinh(3t) +
C1 cosh(3t) + C2 sinh(3t)
```

The "manual" imposition of the periodic boundary conditions from within Matlab with Maple V Release 4 is as follows:

```
>> syms C1 C2 t real;
>> y = C1*exp(3*t)+C2*exp(-3*t)+6*sin(t)/5;
>> dy = diff(y);
>> bc1 = subs(y,t,0)-subs(y,t,2*pi);
>> bc2 = subs(dy,t,0)-subs(dy,t,2*pi);
>> sol = solve(bc1,bc2,'C1,C2'); [sol.C1,sol.C2]
[ 0, 0]
>> subs(y,{C1,C2},{sol.C1,sol.C2})
6/5*sin(t)
```

Note that you can get non-zero constants by changing 2π into $\pi/2$ in problem statement. The first four lines of the output shown at the bottom of previous slide are obtained by using as second command above

```
>> y = dsolve('-D2y+9*y=12*sin(t)');
```

A quite similar "manual" imposition procedure can be followed with other CAS like that of the ClassPad 330 or Maxima v5.21.1—we are aware of no direct way to specify periodic boundary conditions in any of them. For example, in the latter CAS we can either use `ode2` and proceed as

```
(%i1) eq: -'diff(y,t,2)+9*y=12*sin(t);
(%i2) gen_sol: ode2(eq,y,t);
(%i3) define(gs(t),ev(y,gen_sol));
(%i4) define(dgs(t),diff(gs(t),t));
(%i5) par_sol: solve([gs(0)=gs(2*%pi),dgs(0)=dgs(2*%pi)],[%k1,%k2]);
(%i6) define(g(t),ev(gs(t),par_sol));
```

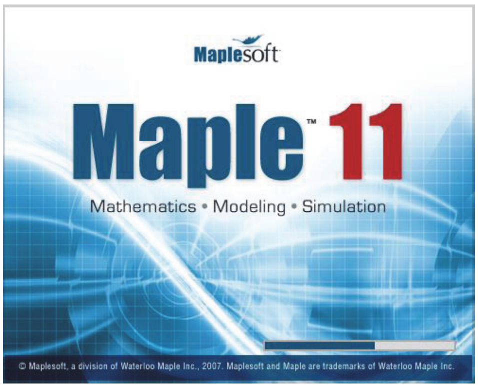
or else we can start with `load('contrib_ode)` and then

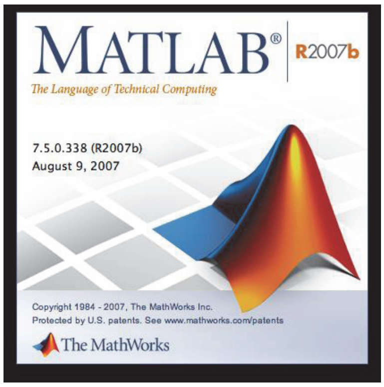
```
(%i1) eq: -'diff(y,t,2)+9*y=12*sin(t);
(%i2) contrib_ode(eq,y,t);
```

```
(%i3) cond1: at(rhs((%o2)[1]),t=0)=at(rhs((%o2)[1]),t=2*%pi);
(%i4) cond2: at(diff(rhs((%o2)[1]),t),t=0)=at(diff(rhs((%o2)[1]),t),t=2*%pi);
(%i5) solve([cond1,cond2],[%k1,%k2]);
(%i6) ratsimp(at(%o2,(%o5)[1]));
```

Task 4.1 *Have a shot at getting analytic solution directly with your favourite CAS solver. Get a differential problem with periodic boundary conditions whose solution cannot be given in terms of elementary functions.*

Maple 11 (Feb 2007), whose kernel is available since Aug 2007 through Matlab v7.5.0.338 Symbolic Toolbox v3.2.2 (R2007b), allow us to specify the periodic boundary conditions to find the explicit solution:






```
>> dsolve('-D2y+9*y=12*sin(t)','y(0)=y(2*pi)','Dy(0)=Dy(2*pi)')
6/5*sin(t)
```

On Sep 2007, the enterprise *SciFace* was purchased by *Mathworks* and its MuPad engine was included in Matlab Symbolic Math Toolbox to replace Maple engine from v7.6 (R2008a) on—in fact, it was already available through Matlab v7.5 Symbolic Toolbox v4.9 (R2007b+). MuPad was withdrawn from market as software in its own right on Sep 2008, and access to Maple kernel from the Matlab environment in recent releases must be done through so-called Maple Toolbox for Matlab by *Maplesoft*.

5 Some still unanswered questions

Let me briefly explain the name I have chosen to define the problem in §2. Saint Thomas is infamous for his declaration after Jesus' resurrection, and when he was finally confronted by the Lord several days later, Christ's response to him has always been taken as a rebuke (John 20:24–29):

> So the other disciples told him, 'We have seen the Lord!' But Thomas said to them, 'Unless I see the nail marks in his hands and put my finger where the nails were, and put my hand into his side, I will not believe it' (...) 'Because you have seen me, you have believed; blessed are those who haven't seen and yet have believed.'

Obviously, Saint Thomas was a person who needed a clear understanding, who wanted all the facts and all of the evidence up front; perhaps he might be considered a good researcher nowadays, and I am pretty sure that all of us might also need our own verification checks!

I was aware of this bug while I was in a meeting with teachers from secondary school to discuss the role of the calculators (Jun 2010) in current mathematical education, and one of them dared me to put forward an application problem in which the solution of a linear equation system larger than 3×3 was required. After undusting and posing this differential problem, I couldn't believe that my brand new CAS calculator was not able to solve those larger than 19×19! Hence the problem name.

As regards the title of the paper, I borrowed it (with minor adaptations) from a 2006 comedy film entitled *A Bug and a Bag of Weed*, but the only relationship with its plot is that I already reported this bug to ClassPad manufacturers ("I forgot to purchase the extended warranty"): they assumed not having a precise/sound explanation and they promised to fix the problem within a forthcoming update. The unedited answer letter from calculator manufacturer was:

> Regarding the ClassPad bug Prof. Guerrero-Garcia send in we finally got answer from R&D. Frankly speaking there is no solution for the problem yet, but R&D will try to solve the problem within one of the next updates. In fact the reason for the different answers of handheld and Manager is quite easy to understand: The handheld utilizes CASIO's own algorism with Decimal system while the Manager uses binary system. For this reason calculation mistakes due to many decimal places occure at different problems. We hope to improve both the ClassPad and the ClassPad Manager. At the moment we can say that calculation capacity of ClassPad is limited due to the small device - even if it comes to big matrices and large numbers of decimal places. (Which is often a problem in Computer calculation!) For more details please see Yagi San's translation of R&D's answer below:
>
> 1. Different answer from CP330 and CP manager. This is software alcoholism issue: CP330 utilize CASIO own algorism w/ Decimal system and PC software utilize binary system. If you use "ref(approx(" function instead of "ref", you may have closer result.
> 2. Overflow ($n \geq 8$) w/ CP330 and CP manager. Both software has limitation of calculation and if it is required over limitation (out of support) you may have error result.
> 3. In Matrix, If $n \geq 20$, get wrong result. We realize this problem w/ current software, it comes from "margin of error" when ref function make execution. This "margin of error" accumulate w/ multiple calculation, then give wrong result.

Task 5.1 *Check if bug has been actually fixed in new forthcoming releases of both the software emulator and the physical hand-held calculator.*

Acknowledgements

I want to thank, in alphabetic order, Rafael Moyano-Franco (Quantitative Methods for Economics and Business Administration, Univ. Pablo de Olavide, Sevilla, Spain, `rmoyfra@upo.es`) for our discussions about MAXIMA using `contrib_ode` (cf. §4), Ángel Santos-Palomo (Dept. Applied Mathematics, Univ. Málaga, Spain, `asantos@ctima.uma.es`) for showing me this differential problem ten years ago [7, 4], and Agustín Valverde-Ramos (Dept. Applied Mathematics, Univ. Málaga, Spain, `a_valverde@ctima.uma.es`) for our discussions about MAXIMA using `ode2` (cf. §4). Finally, I am also grateful to the calculator manufacturer for answering on my bug report.

References

[1] BURDEN, R., FAIRES, J.: *Numerical Analysis, 8th edition.* International Thomson Publishing, Belmont (CA, USA), 2005.

[2] GOOD, T., BROPHY, J.: *Contemporary Educational Psychology, 5th edition.* Longman Publishers, White Plains (NY, USA), 1995.

[3] GOLUB, G., VAN LOAN, C.: *Matrix Computations, 3rd edition.* The Johns Hopkins University Press, Baltimore (MD, USA), 1996.

[4] GUERRERO-GARCíA, P.: Slides for a Course on Numerical Methods (Spanish). Dept. Applied Mathematics, University of Málaga, December 2003.

[5] GUERRERO-GARCíA, P., TORIL, M.: Taking advantage of Sherman's march. *Proc. Technology and its Integration into Mathematics Education (TIME 2010) international symposium*, A052(1)–A052(10), July 2010. Available at http://www.matap.uma.es/investigacion/tr/ma10_01.pdf; full paper version submitted to journal *iJMEST*.

[6] HIGHAM, N.: *Accuracy and Stability of Numerical Algorithms, 2nd edition.* SIAM Publications, Philadelphia (PA, USA), 2002.

[7] SANTOS-PALOMO, Á.: Slides for a Course on Numerical Methods (Spanish). Dept. Applied Mathematics, University of Málaga, September 2001.

[8] SCHUNK, D.: *Learning Theories: An Educational Perspective, 5th edition.* Prentice-Hall, New York (NY, USA), 2007.

[9] SCHUNK, D., PINTRICH, P., MEECE, J.: *Motivation in Education: Theory, Research, and Applications, 3rd edition.* Prentice-Hall, New York (NY, USA), 2009.

Current addresses

Pablo Guerrero-García , Tchng. Fellow Numerical Methods
Department of Applied Mathematics, University of Málaga,
Complejo Tecnológico (ETSI Telecomunicación),
Campus de Teatinos s/n, 29010 Málaga (Spain),
Phone +34 95213 7168.
e-mail: pablito@ctima.uma.es

Journal of Applied Mathematics

volume IV (2011), number I

FINANCIAL EDUCATION WITH THE TECHNOLOGICAL SUPPORT

HAŠEK Roman, (CZ), PETRÁŠKOVÁ Vladimíra (CZ)

Abstract. An implementation of the financial literacy standards into the main curriculum documents of the primary and secondary education is finishing in the Czech Republic at present. It is clear that both in-service and pre-service teachers should be well prepared for such a situation. For this reason the standards of financial literacy are incorporated into curricula of fields of study related to the pre-service and in-service teacher training at universities throughout the country. The Faculty of Education of the University of South Bohemia in České Budějovice in reaction to this demand launched a subject 'An Introduction to Finance' in the fields of study concerning the teaching of mathematics. Due to the frequent changes in the offers of financial products and often unclear, insincere and confusing statements of commercial advertisements it is not easy to get a real grasp of the basic terms in the field of finance, their interrelations and functions. The authors of the article prepared a series of educational materials based on the Maple CAS software to help students to improve their grasp of the actual financial issues. The article introduces the reader into the design and use of these materials.

Key words. Financial literacy, financial education, computer algebra system, Maple, smart document, maplet, teacher training.

Mathematics Subject Classification: Primary 97U70, 97M30; Secondary 97D40.

1 Introduction

The number of indebted households continues to increase in the Czech Republic. With growing frequency we witness the inability of such households to clear their debts. They often tend to the handing over of the debtor to the executors or a personal bankruptcy. The main reason for their situation is frequently the low level of financial literacy of these debtors.

The phenomenon of such bad management of personal finance is not a problem only of the Czech Republic. The issue of financial literacy has also been discussed within the international Organization for Economic Co-operation and Development (OECD). In 2003, OECD initiated an intergovernmental project 'Financial Education Project' (see [8]) which focused on the introduction of an integrated system of financial education, the objective of which will be to increase the level of financial literacy. The results of the whole project were summarized in the publication *Improving Financial Literacy*, [9].

In the Czech Republic, attention to the activities in the field of financial education is paid by the Ministry of Finance of the Czech Republic concerned with consumer protection in the financial market. The Ministry of Finance followed the recommendations stated in the publication Improving Financial Literacy and issued the document *Strategie finančního vzdělávání* (*Strategy of financial education*), [4]. The objective of this strategy is to create an integrated system of financial education that will contribute to an increase in the level of financial literacy in the Czech Republic.

The document *Strategie finančního vzdělávání* (*Strategy of financial education*) is followed by the document *Systém budování finanční gramotnosti na základních a středních školách* (*The system of establishment of financial literacy at primary and secondary schools*), [5], that was approved by the government in December 2007. This is a common document of the Ministry of Finance, Ministry of Education, Youth and Sports and the Ministry of Industry and Commerce of the Czech Republic. This document includes particular standards determining the target status of financial education for primary and secondary education.

In compliance with the document [5], the faculties of education were invited to include the standards of financial literacy in the content of the relevant university programmes of education. The Faculty of Education of the University of South Bohemia in České Budějovice in reaction to such demand launched a subject 'An Introduction to Finance' in the fields of study concerning the teaching of mathematics. The content of this subject is in compliance with financial standards designed for primary and secondary education in the Czech Republic, [2].

2 Financial education at grammar and secondary schools

Three standards of financial literacy were defined in the above-cited document (MF ČR, 2007b), with respect to the target groups of affected pupils: *Financial literacy standard of a first-grade pupil* (ages 6 – 10 years), *Financial literacy standard of a second-grade pupil* (ages 11 – 15 years) and *Financial literacy standard of a secondary school pupil* (ages 16 – 19 years). The last one is the same as the financial literacy standard of an adult. All three standards include the following areas: money, household management and financial products. The secondary school pupil standard contains an extra topic of consumer rights. Let us look at the last two standards in detail. Detailed specification of these standards is given in [5].

In 2008, the specified financial literacy standards were fully implemented into *'General educational programs for gymnasiums and secondary schools'*, which plays the role of curriculum texts in the Czech Republic (see [6]). The standards are reflected especially in the next two educational areas: *'A man and the world of labour'* and *'Mathematics and its application'*. The incorporation of the financial literacy standards into the *'General educational programme for primary education'* has not been executed yet. Until this, primary schools and lower classes of long-term grammar schools can only implement the financial issue into their curriculum optionally.

3 Maple-based tools for financial education

For most people, including university students, it is not easy to have a good grasp of the basic terms in the field of finance, their interrelations and functions. The main reason lies in the fact that the offers of financial products keep changing and their statements are frequently unclear, insincere and confusing. The authors of the article prepared a series of educational materials based on the computer algebra system (CAS) Maple to help students to improve their grasp of the actual financial issues. Each of these Maple-based materials is related to the solution of an example – a real life problem that covers one of the up-to-date kinds of financial products.

These problems are thoroughly solved in a way that explains all aspects of a problem and uncovers possible risks which are mostly kept back from the eyes of an uninformed user. For these purposes the authors use the tools of the software Maple 13. They especially take advantage of the features of interactive so-called *smart documents* prepared in this software. The *smart documents* enable the author to combine text, symbolic and numerical computation and graphs in one worksheet all focused on solving, explaining and modelling some particular phenomenon from financial mathematics.

The materials are distributed to students through the web page, which besides these materials offers educational texts and links to other sources of information. The web page, which is available at http://www.pf.jcu.cz/stru/katedry/m/uf/ (see [10]), is continuously updated with new materials.

3.1 Smart document

The environment of the so-called *smart document* was selected as the main tool for the presentation of a pattern solution and simulation of the functioning of the relations hidden behind the computations of parameters of respective financial products. On the page, these interactive documents are titled *Tutorial* (see Fig. 1). They are available both in MW format that can be run in the mode of the so-called *Standard worksheet* of Maple version 11 and higher and in PDF format that can be used as an adequate teaching text detailing the respective topic. Particular tasks are also completed with the solution code in MWS format generated by the so-called *Classic worksheet* environment of Maple and, if it is beneficial, selected tasks are followed by specially programmed applications called *maplets*.

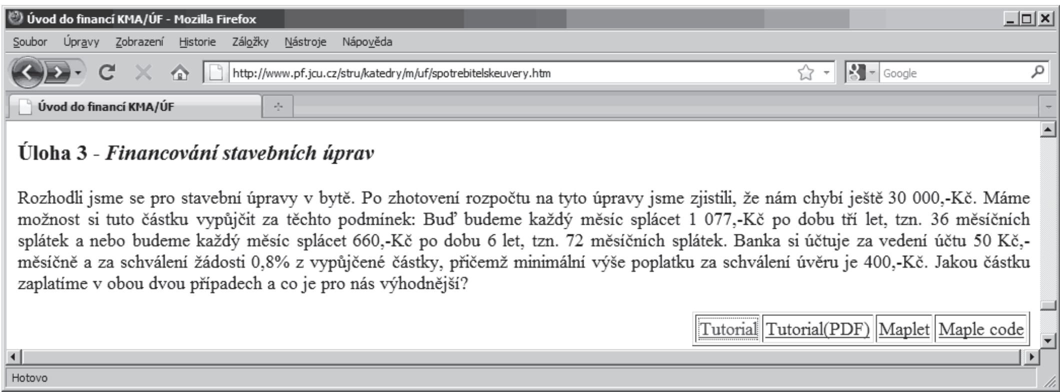


Figure 1. Consumer credit - assignment of Task 1

The *smart document* (see the screenshots on the next pages), also called an *interactive document*, which represents a unique environment enabling the user to combine the text, formulae for symbolic and numerical computations entered in the Maple code, graphs and tools for control of input parameters such as e.g. input and output fields, scroll bars or radio buttons. At first sight, a *smart document* seems to be a solution to a specific task with respective specification. Mathematical formulae used in the document are a part of the text and do not differ in color. However, its advantage lies in the possibility of a change in input values and subsequent conversion through the Maple function 'Execute the entire worksheet'. Thus the user gets a powerful tool for experimentation and examination of the dependence of a solution upon input values. *Smart documents*, also *called interactive* documents, are created in the environment *Document mode* of the program Maple. For this activity we can utilize both the detailed help of the program (after setting command *?Document mode*) and the materials available at the Internet, (Maplesoft), [7]. For illustration we state the solution of Task 2 from the chapter Mortgage credit on the next pages of the paper. Although the task is specified with particular values, respective *smart document* (*Tutorial*) enables the user to change these values arbitrarily and subsequently to recalculate the task in compliance with the actual specification. Thus by means of this single document the user can convert the parameters of mortgage credit for arbitrary values of input data and get a direct comparison to a result corresponding to different input data. Dealing with a mortgage credit a user can for example compare the relation between the amount of interest and the amount of credit for various parameters of the mortgage credit.

The Figure 2, which displays a detailed cut from the smart document devoted to solution of a task from the topic 'Mortgage credit' illustrates how the environment of a *smart document* (*Document mode*) can be combined with the classical working environment of Maple (*Worksheet mode*) in which we can execute supplementary symbolic or numerical computations. In this case, *Worksheet mode* was used for derivation of the relation between the amount of debt and amount of each installment.

Equation adjustment in Maple

```
> restart;
> Dl := a*v*(Sum(v^i, i = 0 .. n-1));
```

$$Dl := a\, v \left(\sum_{i=0}^{n-1} v^i \right) \qquad (1.1)$$

```
> a*v*(Sum(v^i, i = 0 .. n-1)) = a*v*(sum(v^i, i = 0 .. n-1));
```

$$a\, v \left(\sum_{i=0}^{n-1} v^i \right) = a\, v \left(\frac{v^n}{v-1} - \frac{1}{v-1} \right) \qquad (1.2)$$

```
> Dl := factor(a*v*(sum(v^i, i = 0 .. n-1)));
```

$$Dl := \frac{a\, v \left(v^n - 1\right)}{v-1} \qquad (1.3)$$

Figure 2. Mortgage credit – embedded classical code

Thus a student by means of Maple derives the formulae that are or should be used in computations of the parameters of respective financial products.

3.2 Maplet

Besides the environments of the *Standard worksheet* and *Classic worksheet* that are available in Maple we can take advantage of another Maple-based application, the so-called *maplet*. The *maplet* is an interactive application running in a separate window that utilises all computation means of the Maple core. The window of *maplet* can be equipped with a lot of various controls, input and output text and graphic fields, help, commentaries etc. Thus the user will get a tool by means of which he/she can utilise all possibilities of the program Maple without controlling this program. Program Maple need not be even started in the utilisation of maplet, however unfortunately it must be installed in the respective computer.

A user of Maple can make use of a lot of *maplets* that are components of the program installation (for more details see the help pages *?assistants* and *?tutors*) or he/she can build their own *maplets* (the reader acquires detailed information on the creation of maplets by setting the command *?roadmap*). The authors prepared several *maplets* devoted to selected financial issues, for example bond duration.

Among the topics mentioned in the standards of financial literacy we find the problem of the investment of surplus money. One of the possible instruments for the handling surplus money are stocks and bonds. Let's focus on bonds. They are the least risky and are generally the most popular.

One of the important parameters of bonds is the bond duration, which reflects the average time period of the investment to the bonds return. If the duration is no shorter than the term of expiration it is not worth investing in the bonds.
Despite of the duration's importance there are not online calculators of duration available on the Internet. Thus the authors decided to build a special *maplet* devoted to the duration (see Fig. 3).

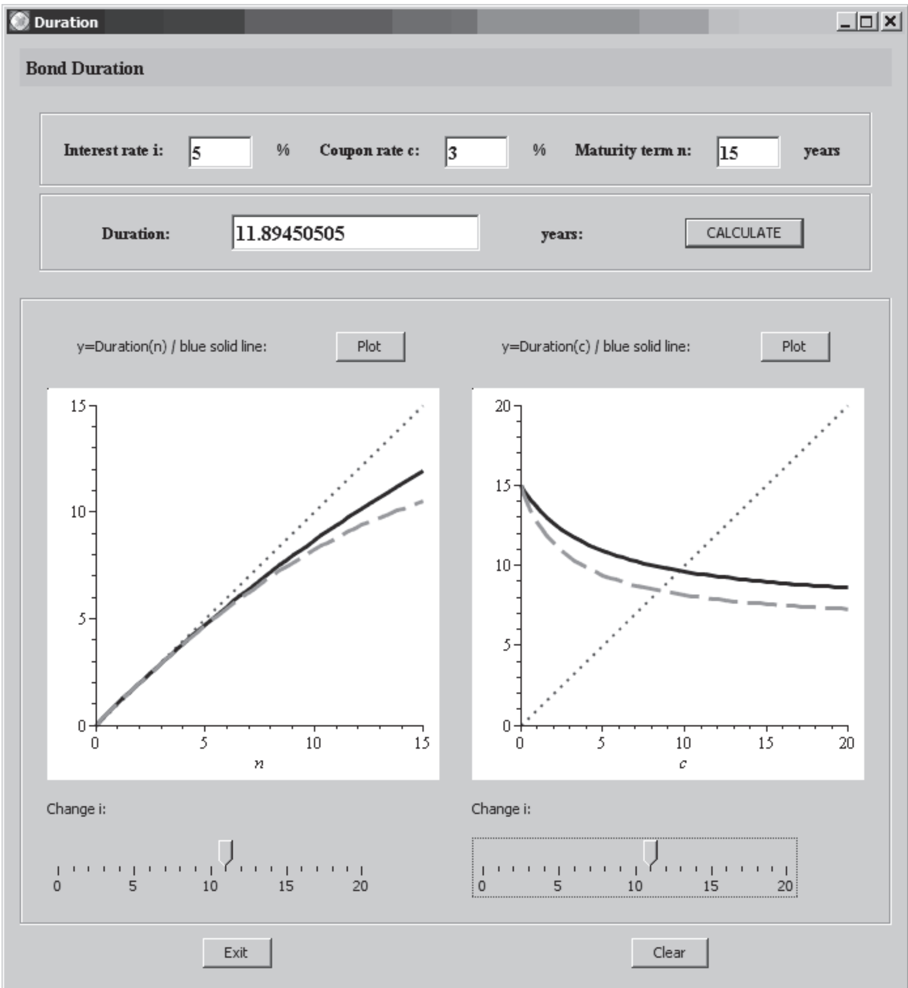


Fig. 3. Duration - maplet

Using this *maplet* it is possible to calculate bond duration in varying value of interest rate *i*, coupon rate *c* and maturity term n. The respective calculation is accompanied by graphic pictures from which the student can become aware of various regularities (e.g. an increase in coupon rate is accompanied by a decrease in duration).

3.3 Maple code

The selection *'Maple code'* (see Fig. 1) offers the task solution code entered in conventional line mode of program Maple. It allows the user familiarized with syntax of program Maple a prompt insight into the task solution and it enables him/her to detect hidden links between particular task parameters by means of experimentation with specification.

```
> restart:
> P:=900: c:=0.1: n:=10: F:=1000:
> i:=fsolve(P=sum(F*c/(1+i)^k,k=1..n-1)+(F*c+F)/(1+i)^n,i);
```

$$i := 0.1175190570$$

```
> Duration:=((sum(k*F*c/(1+i)^k,k=1..n-
1)+n*(F*c+F)/(1+i)^n)/(sum(F*c/(1+i)^k, k=1..n-1)+(F*c+F)/(1+i)^n));
```

$$Duration := 6.576325259$$

```
> deltai:=0.03:
> deltaP:=-(P*Duration*deltai)/(1+i);
```

$$deltaP := -158.8883705$$

```
> EstimP:=P+deltaP;
```

$$EstimP := 741.1116295$$

3.4 Application of the interactive tool

Next, we will illustrate the impact of the interactive tool on the deeper understanding of the functionality of the financial products via the specific example. Among the most frequently used financial products we can count the mortgage loan and consumer credit. Consider the next example from the interactive tool – Task 2 from the chapter Mortgage credit.

On the following screenshots we can see that the interactive aid enables us to change all given parameters of the mortgage loan (the initial credit amount, annual interest rate, number of instalments) and in accordance with these changes it computes the appropriate size of the annuity together with the corresponding curtail schedule (which depends on the size of the annuity).

A student can observe various relationships. For example the change of the annuity rate corresponding to the change of the payment period (number of instalments) or corresponding to the change of the interest rate.

Further, a student can easily learn about the functionality of the relationship between the total interest charges and the values of the task parameters (the amount of credit, the annual interest rate and the time of expiration) from the curtail schedule (the truncated one can be seen on the next screenshot of the smart document).

MORTGAGE CREDIT
Credit repayment with identical instalments – fixed annuity **TASK 2:** A young married couple from example 1 considers another alternative for paying of the debt in the amount of CZK 2,500.000.00. They consider mortgage credit with the same interest rate, i.e. 4.9% but the date of maturity is extended from 20 years to 30 years. The married couple takes out life insurance when in reaching a specified age (after 20 years from signature of the policy) they will receive CZK 980,600.00. They assume that after 20 years with this amount they will repay a part of the debt. Determine a) the amount of yearly overdue annuity instalment of mortgage credit, b) which of the above alternatives is a better one, whether the alternative with the date of maturity 20 years or the alternative with the date of maturity 30 years and repayment of a part of the credit by means of life insurance.
Solution: **ad a)** The amount of yearly overdue annuity instalment and curtail schedule Equation $\boldsymbol{Dl} = \boldsymbol{a}\cdot\boldsymbol{v} + \boldsymbol{a}\cdot\boldsymbol{v}^2 + \boldsymbol{a}\cdot\boldsymbol{v}^3 + \ldots + \boldsymbol{a}\cdot\boldsymbol{v}^n$ holds true, where $\boldsymbol{v} = \frac{1}{1+\boldsymbol{i}}$ is discount factor, $\boldsymbol{Dl}$ initial credit amount, $\boldsymbol{a}$ annuity, $\boldsymbol{i}$ annual interest rate in percentage expressed by decimal number. On the right side of equation there is finite geometric progression with ratio v and with extreme av. After the application of formula for its total we get equation in form $Dl = \frac{a\cdot(1-v^n)}{i}$, where $\frac{(1-v^n)}{i}$ is so-called overdue supplier. Then relation $a = \frac{Dl\cdot i}{(1-v^n)}$ holds true for annuity a. ▶ **Equation adjustment in Maple**

Solution step by step:		
1. We define the functions complying with equation parameters		$restart;$
	Discount factor:	$v := i \rightarrow \frac{1}{1+i}:$
	Overdue supplier:	$ani := (n, i) \rightarrow \frac{(1-v(i)^n)}{i}:$
	Annuity:	$an := (Dl, i, n) \rightarrow \frac{Dl\cdot i}{(1-v(i)^n)}:$

2. We set specific values of input variables of the task	**initial credit amount:**	$Dl := 2500000:$
	payment period (number of instalments):	$n := 20:$
	annual interest rate:	$i := 0.049:$
3. We calculate annuity	$Annuity := an(Dl, i, n) =$ 198909,04 CZK	

For example the fact that an extension of the term of expiration reduces the annuity rate is clear and understandable. But the interactivity of the aid clarifies another fact, namely that this reduction of an annuity is paid by a disproportionate increase of the total interest charge. Similarly we can discover another hidden fact; during the first five years the bigger part of the annuity is devoted to repaying the interest and only the lesser part of the annuity amortises the actual debt. Several years ago bank institutions took advantages of this fact and they permitted a premature redemption of a debt after at least five years of service. Now, in the time of the 'financial crisis' and with great competition within the financial market it has changed. Some mortgage loans can be redeemed as early as after one year.

4. We form curtail schedule

The table of curtail schedule is represented as a matrix

For simplification we lay the value of annuity into variable a: $a := an(Dl, i, n);$

$$198909{,}04 \tag{2}$$

$UP := matrix([['period', 'annuity', 'interest', 'amortization', 'balance'], [0, ``, ``, ``, Dl],$
$seq([j, a, a*(1 - v(i)^\wedge(n-j+1)), a*v(i)^\wedge(n-j+1), a*ani(n-j, i)], j = 1..n-1), [n, a, a*(1 - v(i)), a*v(i), ``], [``, ``, ``, ``, ``], [``, n*a, n*a - Dl, Dl, ``]]);$

period	*annuity*	*interest*	*amortization*	*balance*
,00				2500000,00
1,00	198909,04	122500,00	76409,04	2423590,96
2,00	198909,04	118755,96	80153,09	2343437,87
3,00	198909,04	114828,46	84080,59	2259357,28
4,00	198909,04	110708,51	88200,54	2171156,74
5,00	198909,04	106386,68	92522,36	2078634,38
6,00	198909,04	101853,08	97055,96	1981578,42
7,00	198909,04	97097,34	101811,70	1879766,72
19,00	198909,04	18148,54	180760,51	189617,77
20,00	198909,04	9291,27	189617,77	
	3978180,88	1478180,88	2500000,00	

Conclusion:

The interest amount is $u := UP[24, 3] = 1478180{,}88$ CZK. This means that for borrowing of CZK 2,500,000.00 we will pay almost 1, 500,000.00 CZK.

Computational and graphic capabilities of the Maple 11 software make the discovery of such relations more visual and clear. For example, thanks to the matching graph a student can easily see that it is possible to pay a total interest amount that is equal or even higher than the amount of the loan.

Notes

1. In the last line of the table of curtail schedule we can see how much we paid for respective credit on interest. It is interesting to watch how this amount depends on the term of credit reimbursement.

2. The payment period of mortgage credits is 5 – 30 years; we can pay back the credits in five-year cycles only.

Example: Dependence of amount of paid interest on term of reimbursement

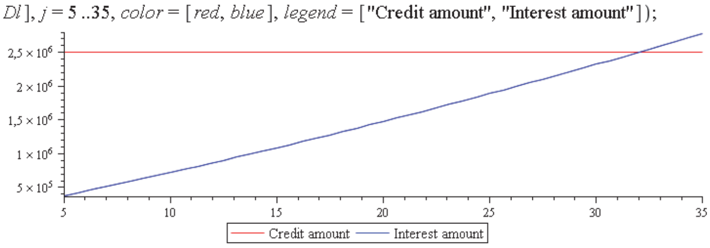
Table:

$matrix([['term\ of\ reimbursement', seq(j, j = 5..30, 5)], ['interest', seq(j*an(Dl, i, j) - Dl, j = 5..30, 5)]]);$

term of reimbursement	5,00	10,00	15,00	20,00	25,00	30,00
interest	379208,54	721910,37	1088456,55	1478180,88	1890187,49	2323384,20

(3)

$plot([Dl, j \cdot an(Dl, i, j) - Dl], j = 5..35, color = [red, blue], legend = ["Credit amount", "Interest amount"]);$



Solution:
ad b) Which of the above alternatives is a better one, whether the alternative with the 20 year maturity date or the alternative with the 30 year maturity date and repayment of a part of the credit by means of life insurance.

the 20 year maturity date		**the 30 year maturity date**	
Paid interest (see solution Task 1, curtail schedule)	1478180.88 $1.47818088\ 10^6$ (5) CZK.	**Paid interest** - the date of maturity 20 years	$\sum_{j=1}^{20} an(Dl, i, 30) * (1 - v(i)^\wedge(30 - j + 1)) = 1963137{,}16$ CZK
Balance	0 CZK	**Balance** after 20 years of repayment (see curtail schedule)	$an(Dl, i, 30) * ani(30 - 20, i) = 1.247547688\ 10^6$ CZK
Difference of interest paid after 20 years	$1963137.16 - 1478180.88 = 484956{,}28$ CZK		

Conclusion:
From the given table we can see that the mortgage credit with the 20 year maturity date is better than the one with the 30 year maturity date.

As we can see above (Solution ad b)) *smart documents* enable us to compare advantages and disadvantages of various solutions to particular situation of a money shortage.

Another important parameter of a loan is the charges. The charges change according to the actual situation within the financial market. The interactive aid enables a student to change this parameter and to explore the impact of such change on the price of the credit.

The smart document also illustrates the connection between financial mathematics and pure mathematics. A student must have knowledge of functions, sequences (particularly the geometric sequences) and series. From the above it is clear that the interactive tool enables a student, based on his/her own experiments, to explore various regularities and relationships in the area of financial products. In this way it helps to improve his/her financial capability.

4 Conclusion

The authors of the article presented a suit of original Maple-based tools that can be used for the teaching of the subjects related to the field of financial mathematics. These materials are components of the web page specially prepared to support such teaching. An advantage of this educational aid is the possibility of continuous updating according to the current condition of the financial products market. It reacts both to the creation of new products and to changes in the conditions (e.g. changes in interest rates, the creation of new products) in banking and financial spheres. Its efficiency was verified by means of the research executed with the students of the Faculty of Education of the University of South Bohemia in České Budějovice, whose results and conclusions are introduced in the article [2].

References

[1.] HAŠEK, R. - PETRÁŠKOVÁ, V. A Way to Improve Financial Literacy of Future Teachers. *Mathematics Education with Technology – Experiences in Europe*. Tamara Bianco, Volker Ulm (Ed.), University of Augsburg, Augsburg: 2010. pp. 199 – 218. ISBN 978-3-00-032628-8.

[2.] HAŠEK, R. - PETRÁŠKOVÁ, V. Issue of Financial Capability. *The International Journal for Technology in Mathematics Education.* Plymouth: 2010. Volum 17, Number 4 (to be published). ISSN 1744-2710.

[3.] Maplesoft. http://www.maplesoft.com/products/Maple/demo/aca.aspx.

[4.] MF ČR. *Strategie finančního vzdělávání* [online]. 2007. Praha: Ministerstvo financí ČR (Ministry of finance of the Czech Republic). [cit. 2010-03-12]. Available at: < http://www.mfcr.cz/cps/rde/xchg/mfcr/xsl/ft_strategie_financniho_vzdelavani.html?year=2009 >.

[5.] MF ČR. *Systém budování finanční gramotnosti na základních a středních školách* [online]. 2007. Praha: Ministry of education, youth and sports of the Czech Republic. [cit. 2010-06-28]. Available at http://www.msmt.cz/vzdelavani/system-budovani-financni-gramotnosti-na-zakladnich-a-strednich-skolach.

[6.] MF ČR. *Rámcové vzdělávací programy oborů středního vzdělání* [online]. 2007. Praha: Ministry of education, youth and sports of the Czech Republic. [cit. 2010-06-28]. Available at http://www.msmt.cz/vzdelavani/ramcove-vzdelavaci-programy-zaslani-do-vnejsiho-pripominkoveho-rizeni.

[7.] NOVÁK, M.. *Maple Basics* [online]. 2009. [cit. 2010-11-15]. Available at: <http://maple.michalnovak.eu/, 2009>.

[8.] OECD. OECD Project on Financial Education. I*nternational Gateway for Financial Education (IGFE)* [online]. 2009. [cit. 2010-10-30].
Available at: < http://www.financial-education.org >.

[9.] OECD. *Improving Financial Literacy: Analysis of Issues and Policies.* Paris: OECD Publishing. 2005. 177 p. ISBN 9264012567.

[10.] PETRÁŠKOVÁ, Vladimíra - HAŠEK, Roman. *Úvod do financí* [online]. Online elektronická učebnice. České Budějovice: Pedagogická fakulta JU. 2008. [cit. 2010-11-1]. Dostupné na WWW: < http://www.pf.jcu.cz/stru/katedry/m/uf/urokovani.htm >.

Current address

Vladimíra Petrášková, RNDr., Ph.D.
University of South Bohemia, Faculty of Education, Department of Mathematics, Jeronýmova 10, České Budějovice, 371 15, Czech Republic, +4207773093,
e-mail: petrasek@pf.jcu.cz

Roman Hašek, Ph.D.
University of South Bohemia, Faculty of Education, Department of Mathematics, Jeronýmova 10, České Budějovice, 371 15, Czech Republic, +4207773088,
e-mail: hasek@pf.jcu.cz

Journal of Applied Mathematics

volume IV (2011), number I

CHOSEN FEATURES OF MAPLE FOR MATHEMATICAL MODELING IN EDUCATION AND PRACTICE

HREBICEK Jiri, (CZ), CHVATALOVA Zuzana, (CZ)

Abstract. This paper deals with chosen features of the last versions of Maple system (Release 14 and 15) for teaching mathematical modeling online and its applications in scientific disciplines, especially in the area of economical/financial modeling; students learn mathematical modeling as an important interactive support for understanding and presenting solved real problems. In this paper is presented chosen basic ideas on process of mathematical modeling which is demonstrated in examples of several computational models in practice with the support of the system Maple.

Key words. System Maple, Mathematics, Education, Modeling, Economics.

Mathematics Subject Classification: Primary 97Q20, 97R20; Secondary 97M40.

1 Introduction

It would be possible to suppose, that one of the few solid pillars, on which in the period after the year 1989 at least in some views would link to and in new conditions also continuously and positively enlarge, was *Czechoslovak education,* because it had *traditionally high-quality base, international prestige and it was enough functional*. In addition to open connection with the world was the promise of rapid and wide penetration of the most modern technologies to user´s custom, also to schools of all degrees. It is a paradox that from the 20th and 21st century break, the education in the Czech Republic (CZ) became instable.

2 Mathematical Position in Present, Chosen Experience

1. From the sight of mathematical education, its position became more than weakened. The applications of its less and more advanced methods became uncomprehended, unjustly as far as disdainful. By the cancellation of leaving examination from mathematics, by releasing the curriculum the education of mathematics often resign to „undercutting“ by simplifying or

elimination of passages, where is needed the necessary degrees of active thinking. So eventually it is not possible to retrace neither philosophy nor logical line. Mathematics is than default the primary function not only in learned calculation methods, but also as support and training of abstract thinking. It smells of „cramming“, it is either changed for calculuses and it appears as the subject, which is difficult, whose „taught methods in real life will not be needed“. These kinds of opinions, unfortunately, transmit into more branches of university mathematics.

2. As well as all sorts of newly implementing, respectively again abandoned norms and systems, which rise so called from clerical table, frequent political changes, high diversification of school qualities, uncertainty, the sight of schools as on a company, which forms „products“, which should ensure above all profit, and not tangible, but rather existence. The friction of conservative and modern methods, instruments and opinions are entering into educational process lead rather to stagnation of present educational system. The estabilishing of the state leaving examinations, which will be with high probability newly realized in spring 2011, is mantled by many sorts of opinions. Especially their so-called pilot realizations, performed in autumn 2010 except some procedural absence, revealed pathetic fact.

3. Novinky.cz shows to this: „*At general leaving exam failed in the tests thirty three percent of students. At exam from mathematics*[1] *did not pass the easier version even forty eight percent of students*[2]*… The students could get into a trap, whether it is better for them to choose compulsory part of exams, either mathematics or a foreign language… According to expectations, gymnasium went much better than apprentice schools with leaving exams. Public schools went better than church and private schools.*“[3] At the same time on portal the Ministry of Education, Youth and Sports is information, that the plenum of the Czech Conference of rectors received a decree, on its conference, about the principle support of state leaving exams. The decree i.e. tells about: „*The Czech Conference of rectors supports the principle of state comparative leaving examination as a possible instrument of high-quality guarantee of general secondary education. After experiences with its establishing, leaving examination could become one of the bases for entrance exams to universities.*"[4]

4. Above mentioned finding, however, tells about further serious fact, which is necessary to speak about. To first years of universities, where mathematics is a frequented subject, get students with very different and as the school leaving exam showed, with a pretty low level of knowledge (especially concerning to apprentices with leaving exam, graduates of church schools or private schools). Regarding often to high capacity of accepted students, the teacher has a very difficult work. Shortly told by today language: At all sorts and relatively a great deal of poor quality “inputs” the pedagogue must in some weeks demonstrate “to produce” equivalent “outputs” for specified knowledge. Seminars and trainings of mathematical subjects are new and final officially proved fact interfere. If the trend of nowadays is the creativity support of students, it is necessary to notify, that the creativity without knowledge is not, surely, the sign of society. In mathematics is going mainly of many years standing successive knowledge developing.

---

[1] Mathematics was a compulsory optional subject.

[2] In the Czech language failed 22 percent of students, in English failed 34 percent of students, in Mathematics 48 percent and in German language 40 percent of students.

[3] Novinky.cz. *At General School Leaving Exams Failed one Third of Students.* [online]. 2010, [accessed 12 December, 2010]. Available from Internet: <http://www.novinky.cz/domaci/215467-u-generalky-maturit-neuspela-tretina-studentu.html>.

[4] MEYS. *Rectors Support the School Leaving Examination of Students.* [online]. 2010, [accessed 12 December, 2010]. Available from Internet: < http://www.msmt.cz/statni-maturita/rektori-podporuji-statni-maturitu>.

5. Significant, but nor overestimated instrument for finding of positive motivation of students again, but also the elimination of above mentioned problem in mathematics can evidently contribute to natural exploitation of suitable information and communication technology (ICT) tools. It supports the independency, but also cross-communication, team-work, the possibility of training and deeping of subject matter with the necessary measure by personal access. It is making possible the teacher at explaining of given problems for example "to show" the problem from practice, so that the students understand, the various mathematical passages are not only rapt isolated and unusable facts in general life (below in 3.1). The same way as sophistically quantifies and graphically visualizate the solution of problems from practice in final works or projects.
6.

3 Computer Support of Mathematical Education

7. Presently are developing many computer systems. Some of them offer a rear for complex solutions of problems in various computer literacy, [2], [3].
The Maple system also belongs to this kind of systems, whose actual version 14 was distributed in 2010. Presently is preparing the next version 15, which will be distributed in 2011. Maple is an efficient instrument for production and presentation of documents of various complexities. It uses the methods of mathematical disciplines, executes the numerical and also symbolic calculations. By high degree of utility user´s environment with the support of many interactive elements, always innovated clicking calculus and with inbuilt predefined components it becomes a strong instrument of mathematical education and its applications.

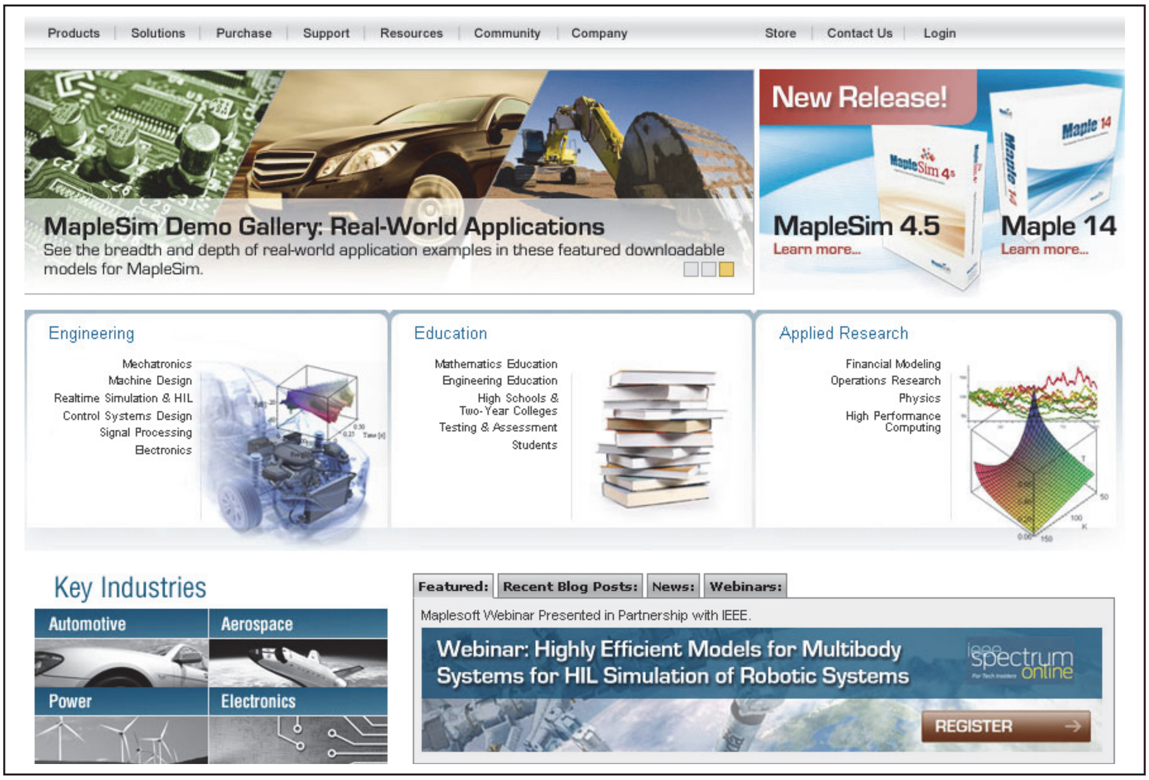


Fig. 1. Homepage of Maplesoft, Inc. *Maple 14, MapleSim 4.5* [online]. 2010, [accessed December, 2010].
(Source: Available from Internet: < http://www.maplesoft.com/>)

The Maple system had been developed more than thirty years by a Canadian company Maplesoft Inc. (http://www.maplesoft.com). Its websites give widely hypertext information about the system and about related products and activities. Home web page of Maplesoft, Inc. is shown on the Fig. 1. It looks forward to favour on many universities and scientific workplaces, as well as on many institutions in practice, public administrations and also commercial sphere, [2]. From the view of obtaining of literacy and skills for the solution of practical problems now this possibility of Maple enters into the process of education to its using in practice is its significant attribute for its setting-on.

From ninetieth in 20th century the Maple system gained place also at Czech and Slovak academic users. The reasons of its favour are discussed and presented on many national and international forums, [2].

The primary developmental line of the company Maplesoft Inc. from the beginning has been the support of mathematical education in different levels. Maplesoft produced further related components of Maple (as well as constantly developing of updated version of the Maple system) reacting to instigation from the sphere of education, science, research, practice, but it opened also activities and possibilities of communication among Maple users or its registered users, for example:

Technologies: *Maple T.A.*^{TM} is for training of subject matters at classical education, e-learning and delimitation of rules for tests making, *MapleSim* is determined mainly for engineering simulations, *MapleNet*™ enables to process Maplet Learning Objects by using Java™ browser on computer without installed Maple, however by access to server with Maple through on Internet or Intranet.

On the web Maplesoft is possible to find special Centers:

- the Student Help Center (http://www.maplesoft.com/studentcenter/) offering a Maple student forum, on-line mathematic Oracles, training videos and a mathematics homework resource guide;
- the Teacher Resource Center (http://www.maplesoft.com/teachercenter/) offering teachers to get the most out of their Maple teaching experience. It provides sample applications, course material, training videos, white papers, e-books, podcasts, and tips. [4];
- The Application Center (http://www.maplesoft.com/ applications/);
- the Documentation Center (http:// www.maplesoft.com/documentation_center/);
- the Media Center (http://www.maplesoft.com/ media/) etc.

The Maple setting-on in CZ added also supported activities by the club *Czech Maple User Group* (http://www.maplesoft.cz), which provides the support to Czech Maple users. Masaryk University in Brno (CZ), as well as Brno University of Technology (CZ) is using multilicence of the Maple system. Students can use the Maple system during their education, at processing of their master and bachelor final works, projects and also of researches evaluation.

3.1 Demonstration of Simple Motivational Examples by Maple Support

Integrals and matrix are redoubtable subject matter at universities. Simple „mathematization“ and comprehensible clarification of common problem from practice with computer support is possible for example to motivate at the beginning of subject matter the student, in order to understand, that the mentioned mathematical objects have a taste for studying just for their application in practice.

3.1.1 Definite Integral – Geometric Interpretation of Economical Exercise

Task: *Mr. Brown supposed, that in the first three years the oil consumption (y) in dependence on time (x) in his newly based trucking company will grow exponentially* $y = e^x$. *How was his surprise, when he found out, that after three years, the oil consumption in the first year really grew exponentially, however the second and the third year it grew quadratic i.e.* $y = x^2$. *Over managed Mr. Brown more or less oil in the second and in the third year against original supposition? How large amount of oil?*

Analysis and solution: The depiction of curve (Fig. 3) and the delimitation of area, whose content matches in demand different oil consumption (Fig. 4), in the Maple Document is possible easily perform the clicking calculus and the simple commands. By the first button on the mouse, using the interactive elements in upper bar or in open frame (Fig. 2) then it is possible operationally to sketch in the whole graphic output and callouts.

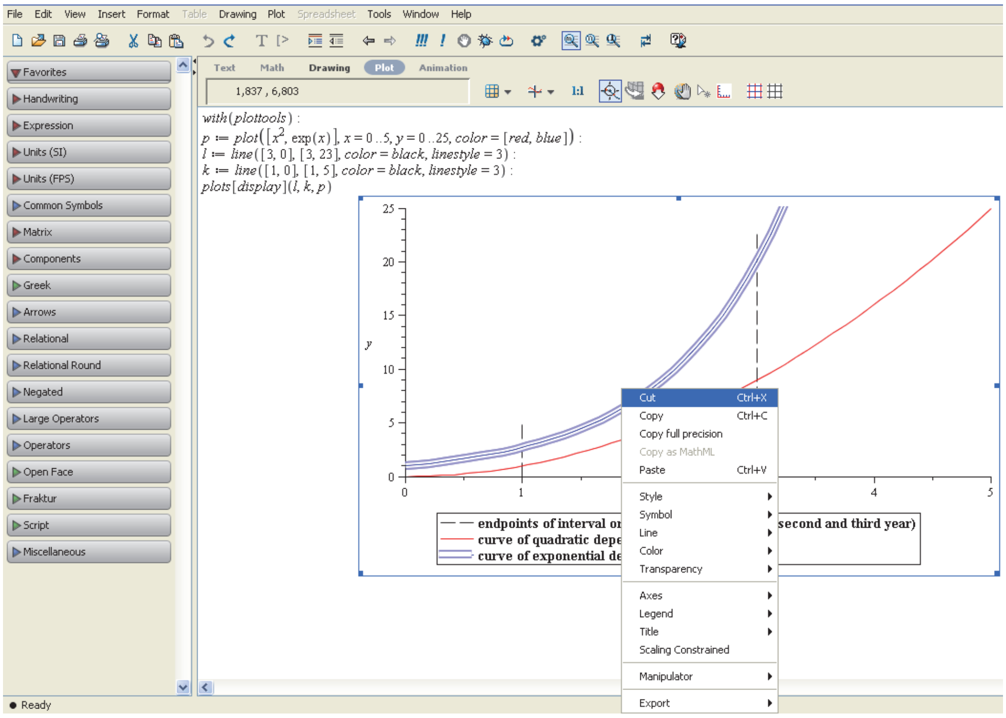


Fig. 2: Interactive tools for visualization in Maple Document (Source: Owen work in Maple)

Output – exported graphic plot from Maple:

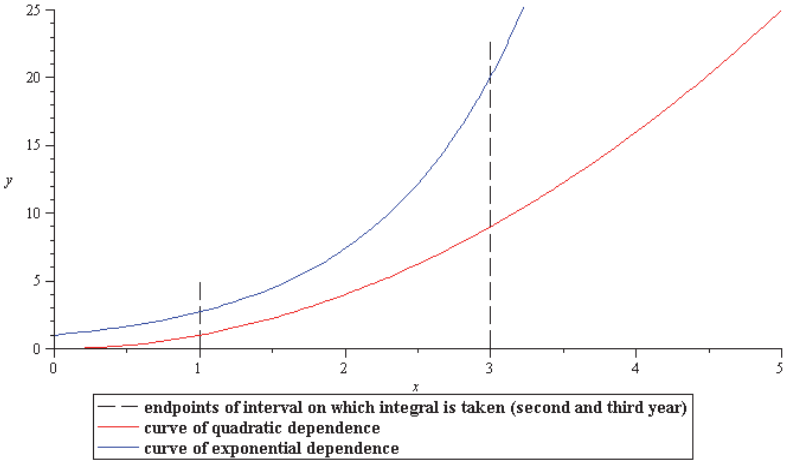


Fig. 3: Curve visualization (Source: Owen work in Maple)

$with(plottools):$
$p := plot\left([x^2, \exp(x)], x = 0..5, y = 0..25, thickness = 2\right):$
$l := line([3, 0], [3, 23], color = black, linestyle = 3):$
$k := line([1, 0], [1, 5], color = black, linestyle = 3):$
$f := plottools[transform]\left((x, y) \to [x, y + \exp(x)]\right):$
$p2 := plot\left(\exp(x) - x^2, x = 1..3, filled = true, color = cyan\right):$
$plots[display](l, k, f(p2))$

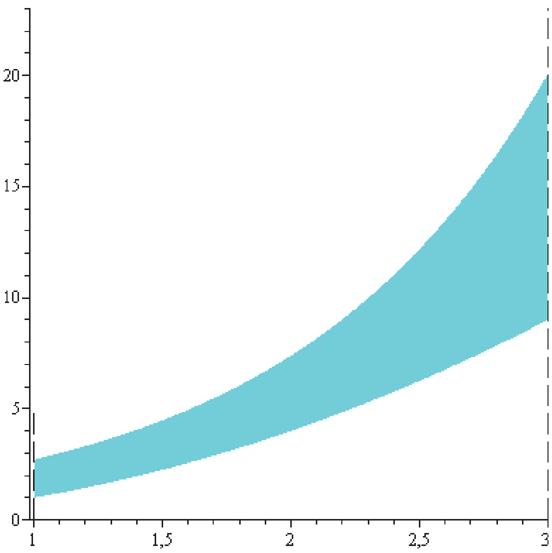


Fig. 4: Content visualization of area (Source: Owen work in Maple)

$\int_1^3 \left(\exp(x) - x^2\right) dx; evalf(\%)$ $\qquad -e - \frac{26}{3} + e^3 \qquad 8.70058842$

Conclusion: in the second and the third year Mr. Brown saved c. 8.7 units of oil capacity.

3.1.2 Matrix Product – Work with Tables from Financial Environment

Task: *The company of Mr. Novak has three workplaces A, B, C. In each workplace the staff in is divided into four categories with fixed salary allotment. The staff in K category get 10 000 Kc monthly, in category L 15 000 Kc monthly, in M category 20 000 Kc monthly and in the category N 22 000 Kc monthly. The number of workers in each category on particular workplace determines Table 1:*

Tab. 1: The numbers of workers of each category on particular workplace

	K	L	M	N
A	3	12	10	5
B	2	12	14	5
C	4	22	26	6

Mr. Novak promised that at fulfilment of a target in before the date, he will reward the worker in the next month. The height of the financial reward is different for each wage category and the same for all the 3 workplaces. (Tab. 2)

Tab. 2: The height of the financial reward of each category

	K	L	M	N
The height of the financial reward	20	15	16	22

The interest of Mr. Novak is, how much is the sum of rewards, which he must send to account of each workplace, where the sum will be separately divided and how much is the total sum.

Calculations with the tables we will demonstrate on the calculation of the matrix product. The matrix product in Maple we gain by the command *evalm* (A &*B), at the same time to primary aditing of the matrix, it is enough „to pull“ from the pallet of tools Matrix by a mouse intuitively the appropriate type of matrix and in the mode “as the natural hand-made” record to fill in its elements to corresponding position (Fig. 5):

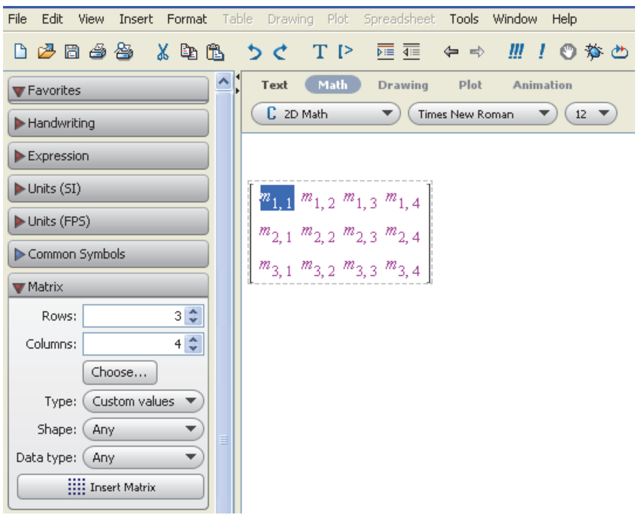


Fig. 5: Interactive tools for work with matrix in Maple Document (Source: Owen work in Maple)

The product of sum rewards, which is necessary to send on each workplace in Maple, we gain like this:

$$evalm\left(\begin{bmatrix} 3 & 12 & 10 & 5 \\ 2 & 12 & 14 & 5 \\ 4 & 22 & 26 & 6 \end{bmatrix} \&* \begin{bmatrix} 0.2\cdot 10000 \\ 0.15\cdot 15000 \\ 0.16\cdot 20000 \\ 0.22\cdot 22000 \end{bmatrix}\right) = \text{gained final amounts} \begin{pmatrix} 89200 \\ 100000 \\ 169740 \end{pmatrix} \begin{matrix} \text{workplace A} \\ \text{workplace B.} \\ \text{workplace C.} \end{matrix}$$

Total sum paid in rewards is obtained by the product of elements up on high of calculated matrix which is possible to ensure like this:

$$evalm\left(\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \&* \begin{bmatrix} 3 & 12 & 10 & 5 \\ 2 & 12 & 14 & 5 \\ 4 & 22 & 26 & 6 \end{bmatrix} \&* \begin{bmatrix} 0.2\cdot 10000 \\ 0.15\cdot 15000 \\ 0.16\cdot 20000 \\ 0.22\cdot 22000 \end{bmatrix}\right) = (358940) \text{ is the total sum of rewards.}$$

Mr. Novak will send 89 200 Kc as a reward to the workplace A, 100 000 Kc to the workplace B, 169 740 Kc to the workplace C. The total sum what he will pay in rewards is 358 940 Kc.
It is possible to solicit the students to sequential independent reasoning and the formulation of tasks, which would be possible to solve by using the matrix product.

3.2 The Demonstration of Chosen Outputs of the Solution of Economic Problems in Maple in Final Works on the Faculty of Business and Management Brno University of Technology

8. The Faculty of Business and Management Brno University of Technology (FBM BUT, CZ) is a school of economic character and in its entire syllabus, mathematical subject have their own fixed places. The faculty is interested in the employment of its students (graduates). That is why it continuously innovated the ICT tools. FBM supports the connection of education with practice of all sorts of forms – for example at processing of final works, researches etc. Using of mathematical moulding, optimization, regression analysis, econometrics applications and also other quantitative methods are more needed in economics, in the discipline before envisaged at the branch of social – science and above all using social – economic metric.
9. We can suppose, that the specialists working at the level of management in economics, financial and business sphere will edit more quality decision, by a firmer knowledge in the area of quantitative theory and the habits is necessary to process the instruments, ICT accepts already in a period of education and they will have a greater possibility to develop their knowledge, [5].

3.2.1 The Use of Neural Networks in Maple at Mathematical Models of Business Efficiency

Shortly: Input data are significant sums of particular company divided into three groups (teaching, testing, verifying). The output is the value gained by using of neural network, whose process copying the process of values of EVA[5] indicator. The basic architecture of the network is

[5] EVA was developed in 1993 by company Stern Stewart & Co. in the USA. EVA as a measure of surplus value created on an investment is based on the idea that a business has to cover both, the operating costs and the costs of capital. It stems from the estimate of economic, not the accounting profit. EVA can be defined in

a six layered perceptron working in the area of real numbers with linear rating of neurons. The network connection is absolute. The network is created in the programme Maple and it is composed of three coherent modules. The network uses for storage of necessary data and intermediate calculations vectors and matrix which are supported in Maple by many functions. The matrix and vector character of data enables a simple using of cycles. At learning of the network for EVA indicator was made as a whole 20 223 study epochs, [1].

Tab. 3: The final table of calculated values of EVA indicator classically and neural networks of system instruments Maple (Source: Adjusted according to [1])

Amount Names	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
EVA (classically)	2072	5749	14415	24179	32092	36658	22761	42265
EVA (neural network)	1927	7548	12672	22147	34445	43689	18961	36314
Relative Error	7.00 %	31.29 %	12.09 %	8.40 %	7.33 %	19.18 %	16.70 %	14.08 %

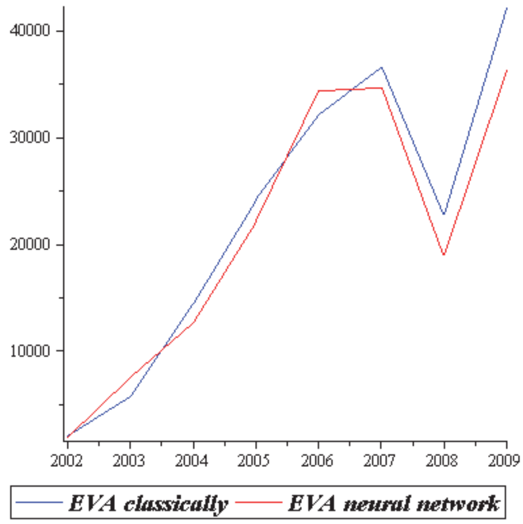


Fig. 6: EVA classically and neural network (Source: Adjusted according to **[1]**)

The juncture graph (Fig. 6) is constructed in Maple. It is seen, that the appropriate, and at the same time not complicated neural network is copying the trend of EVA indicator intended classically. This way processed network learnt how to transform “raw” input values from the balance on the concrete important EVA indicator [1].

3.2.2 Regression Model of Financial Data in Maple

It is very easy to import very intensive files by common office applications, and furthermore process them. In the Fig. 7 is caught up a constructed well-arranged Maplet, with some further before developed functions, for the determination of regression model of development dependencies of two significant titles of the Czech Stock Exchange including their time development, as it was processed, [9].

several ways, more in for example [7], [8]. By the reason of the article is not efficient to specialized the EVA indicator.

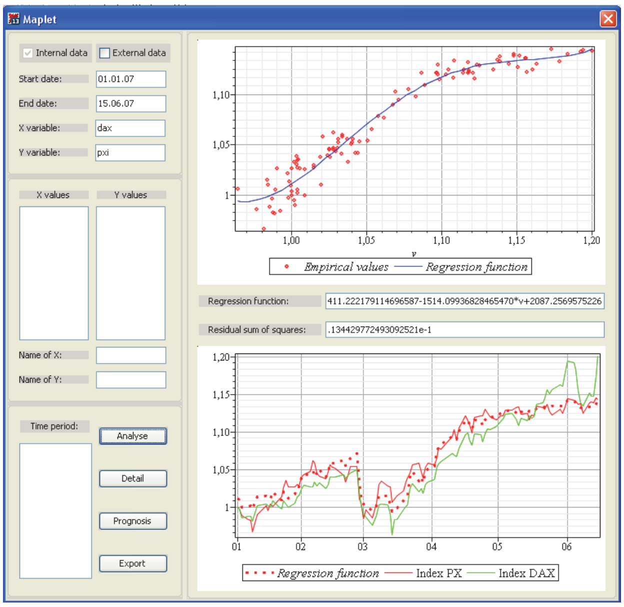


Fig. 7: Financial modeling (Source: Adjusted according to [9])

4 Chosen Features of Maple 14 and Maple 15

Maplesoft has introduced the concepts of *Clickable Mathematics with Clickable Calculus, Clickable Algebra, Clickable Engineering*[6] to make mathematics easy to do, easy to learn, and easy to teach through very visual, interactive point-and-click methods. This approach allows teachers and students to focus on concepts and eliminates the need to memorize command names and syntax.
Maple worksheet interfaces allow for an unprecedented level of point-and-click problem solving and visualization. Teachers and students can combine text and mathematics in the same line of Maple environment (worksheet), add tables to organize the content of his work, or insert images, sketch regions, and spreadsheets. They can visualize and animate problems in two and three dimensions, format text for academic papers or books, and insert hyperlinks to other Maple files, web sites, or email addresses. They can also embed and program graphical user interface components, as well as devise custom solutions using the Maple programming language [6]. The concept of Online Clickable Mathematics includes in Maple 14 and Maple 15:

- Smart, context-sensitive right-click menus for instant access to solvers and other command-free operations;
- Extensive range of palettes for visual editing of math expressions;
- Interactive plotting and animation controlled by the mouse and not by endless parameters and attributes in a command;
- "Drag and drop" operations on plots, expressions, text, and more;
- Interactive assistants that provide easy mechanisms to solve and explore advanced topics such as differential equation-solving, optimization, and advanced visualization;
- A Maple Portal for Students, which acts as a guide for hundreds of common tasks from mathematics courses;
- Built-in selection of interactive tutors that offer visual e-learning environments for many important mathematical topics in pre-calculus, calculus, linear algebra, and more;

[6] Available from Internet: < http://www.maplesoft.com/products/maple/academic/applications.aspx >

- Handwriting recognition of mathematical symbols and equations;
- WYSIWYG document processing features that let you create complex math documents more quickly and easily than in a word processor or LaTeX;
- An Exploration Assistant that allows you to instantly create interactive mini- applications to explore the parameters of expressions.

4.1 Maple Cloud Computing

The term Cloud represents the Internet in computer network diagrams as an abstraction of the underlying infrastructure. Cloud computing describes a new supplement, consumption, and delivery model for ICT services based on the Internet, and it typically involves over-the-Internet provision of dynamically scalable and often virtualized resources. Cloud computing providers deliver common software applications online which are accessed from a web browser, while the software and data are stored on servers [4].
The MapleCloud Document Exchange™ service being offered by Maplesoft is designed primarily as a means of sharing Maplesoft-related content. Teachers and student can use the MapleCloud system on Internet to share available worksheet content with other users (Fig. 8), view content shared by other users from given mathematical course, and store entire standard Maple worksheets or selected content from standard Maple worksheets[7].

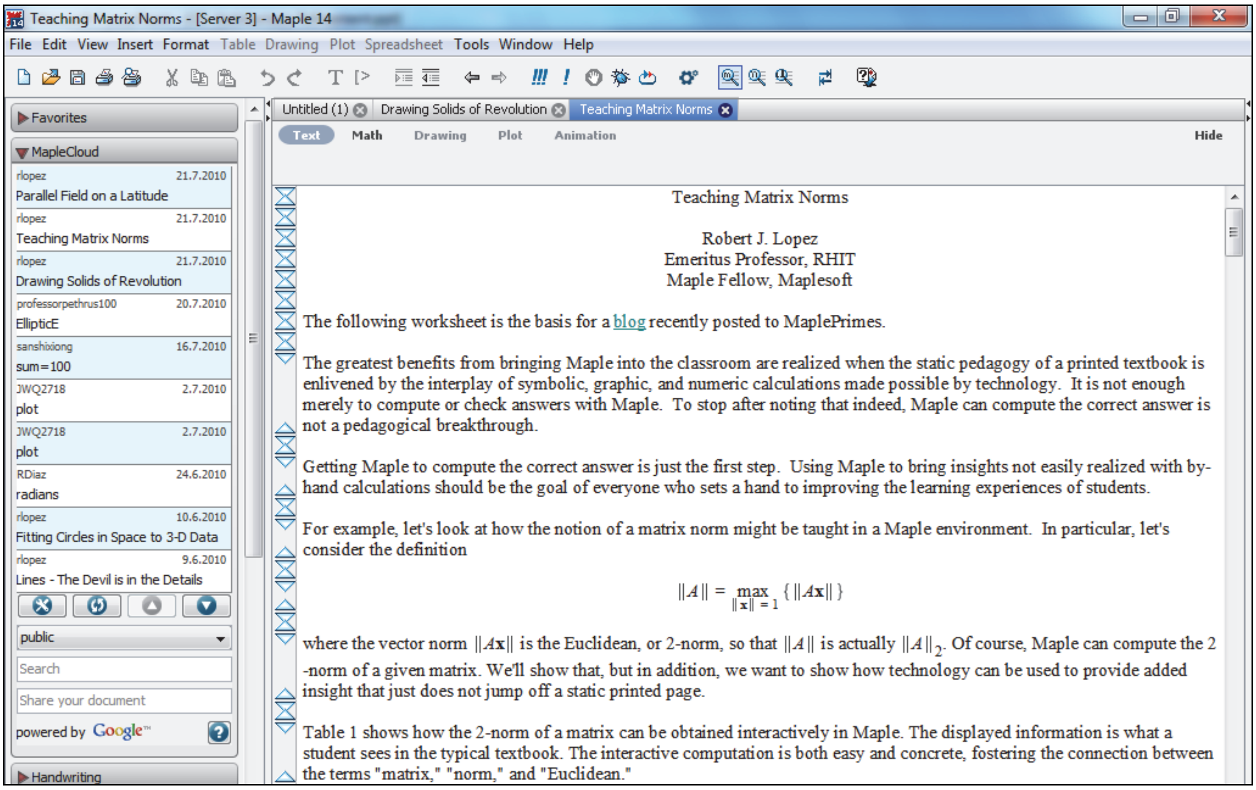


Fig. 8: MapleCloud Document Exchange service (Source: [4])

[7] Available from Internet: < http://www.maplesoft.com/support/help/Maple/view.aspx?path=worksheet%2fcloud%2foverview >.

The teacher can upload Standard Worksheet content and allow other users of he given group to download a copy of that content. Students can also upload and store content in a user-specific area that only they can access.
To share content with specific users (colleague teachers), the teacher can either create a user group or select an existing group and allow only those group members to access his content. Users need an internet connection to use the MapleCloud. To share worksheet content, create, manage and join user groups; and view group-specific content, users must log in to the MapleCloud using a Maplesoft.com, or Gmail™ account name and password.

4.1.1 MapleCloud Groups

To control who can view worksheet content that users upload to the MapleCloud Document Exchange, MapleCloud Groups user can share content in a group. Similarly, he can view content uploaded in a group if he is a member of that group and the group owner has assigned group members permissions to view content. In the *MapleCloud* palette, the drop-down list contains the groups that you can access. Maple Cloud user can view, share, and upload content in a MapleCloud group according to the permissions assigned to the group.
All Maple users can view and share content in the *public group*. Content uploaded in this group can be viewed by the entire Maple community, including users outside of your organization.
The *Maplesoft@admin group* is a read-only group in which Maplesoft shares example worksheets, applications, and other content that can help you with user's Maple projects. All Maple users are members of this group. User does not need to be logged in to view content shared in this group.

4.1.2 MapleCloud Terms of Service

The use of the *MapleCloud Document Exchange*™ is subject to the *MapleCloud Terms of Service* (http://www.maplesoft.com/cloud/terms.aspx). The MapleCloud Document Exchange™ service being offered by Maplesoft is designed primarily as a means of sharing Maplesoft-related content. There are certain conditions of the use of MapleCloud services, including but not limited to the following ones:

- By submitting, posting or displaying content through this service, MapleCloud service user gives Maplesoft a perpetual, irrevocable, worldwide, royalty-free, and non-exclusive license to use, reproduce, adapt, modify, translate, publish, publicly display and distribute any content which has been submitted, posted or displayed on or through this service.
- For user-created groups, the group moderator can remove any content at his/her discretion. For all content whether it was submitted in a public or private group, inappropriate content can be removed by Maplesoft. Content authors agree to not post malicious content to the MapleCloud that interferes or disrupts the proper functioning of the MapleCloud service or any other users' computers.

When user is logged in to the MapleCloud Document Exchange, only he can upload and view content in the *private group*. Content uploaded in this group cannot be viewed by other users. All Maple users have access to a user-specific *private group*. You can upload content in this group, for example, to store content that you plan to retrieve on another computer.
A Maplesoft.com membership account gives users access to thousands of free Maple resources and MaplePrimes, which is an active web community for sharing techniques and experiences with

Maple and related products. MapleCloud Best Practices web[8] describes some best practices for using the MapleCloud™.

4.2 Maple GRID Computing

GRIDs represent a form of distributed computing whereby a "super virtual computer" is composed of many networked loosely coupled computers acting in concert to perform very large tasks. Supercomputing Centre Brno (SCB) of MU is a workplace that deals with high-performance and high-throughput computing (HPC) and provides support and further development of HPC infrastructure of MU. SCB covers all MU activities concerning GRIDs and/or HPC in general. Regarding this fact SCB is a key member of the Czech local grid project METACentrum (http://meta.cesnet.cz/cms/opencms/en/index.html) whose aim is to enlarge the infrastructure of academic high speed network by support for applications requiring extensive computational resources. There is installed and used *Maple Grid Computing Toolbox* for teaching HPC in mathematical modeling courses at MU.

10. The *Maple Grid Computing Toolbox* provides tools for performing Maple computations in parallel, allowing its user to distribute computations across a network of workstations, a supercomputer, or the CPUs of a multiprocessor machine. It includes a *personal grid server*, allowing its user to simulate and test his parallel applications before running them on a real GRID network. The *Maple Grid Computing Toolbox* includes:

- A self-assembling grid in local networks, with an easy-to-use interactive interface for launching parallel jobs.
- Integration with job scheduling systems, such as Portable Batch System (PBS) of PBSworks (http://www.pbsworks.com/) and integration with Microsoft Windows HPC Server.
- High-level parallelization operations (e.g. *map* and *seq*), as well as a generic, parallel divide-and-conquer algorithm.
- Message Passing Interface (MPI) and MPI-like message passing Application Programming Interface (send, receive, etc.).
- Automatic deadlock detection and recovery.

4.2.1 Use of Maple Grid Computing Toolbox

Users of the Maple Grid Computing Toolbox can distribute computations across the nodes of a network of workstations, a supercomputer or, the CPUs of a multiprocessor computer, i.e. they have a support for heterogeneous networks. This allows them to handle problems that are not tractable on a single machine because of memory limitations or because it would simply take too long.

11. The Maple Grid Computing Toolbox is available in two different versions:

- The *Personal Edition* supports up to 8 CPUs in the cluster.
- The *Cluster Edition* supports an unlimited number of CPUs in the cluster.

12. The personal grid server allows its users to simulate a grid with any number of nodes on their desktop computers. They can develop and test their parallel applications before running them on the real grid.

To run Grid Computing applications on a network of machines, Maple Grid Computing Toolbox users must start Grid servers on each computer that they want to be part of the network. The toolbox will automatically detect new computers as they are added to the network, or they can turn off auto-

[8] Available from Internet: < http://www.maplesoft.com/cloud/bestpractices.aspx >.

discovery for use in a controlled environment, such as a PBS setup. They will need to re-configure the Grid so that they can create or be added to a Grid network [4].

Conclusion

Our experience and current research in the field of teaching mathematical modeling discovered a number of important aspects and trends. Primarily, the needs of contemporary mathematical modeling with the use of Maple reflect the growing complexity of problems in education of mathematical modelling in economic/finantial area as well as progress in ICT represented by grids, clouds, and virtualization in general. At the same time, the profile of a student, future researcher or engineer requires – more than ever before – the involvement of transversal competencies such as teamwork abilities, languages, communication and other soft-skills to be able to work in heterogeneous, often multinational teams. Teaching must be interactive, with intensive knowledge sharing and transfer supported by modern tools. The system Maple satisfies these needs. This paper presented chosen features of the system Maple for mathematical modeling in education and practice. Future research should be targeted on further, deeper incorporation of grid- and cloud-based approaches already in fundamental courses on mathematical modeling, at least in an introductory form.

Acknowledgement

This paper was also supported in part by the project No.3186/2010 "*Videomanual as a multimedia support innovation in the subjects Mathematics and Mathematical seminar*" of the Ministry of Education and Youth and it was also partly supported by Faculty of Business and Management Brno University of Technology. The supervisor of this project is Zuzana Chvatalova.

References

[1.] BARTULEC, T. *Mathematical Modeling of Company Efficiency Using Neural Networks in Maple*. *(master's thesis in Czech)*, supervisor RNDr. Zuzana Chvatalova, Ph.D.). Brno (CZ): FBM BUT (CZ), 20.

[2.] HREBICEK, J.; CHVATALOVA, Z. *Maple as an Instrument for Connection School and Practice*. In M. Kovacova (ed.), *Proceedings of 8th International Conference APLIMAT 2009.* Bratislava, Slovak Republic February 3-6, 2009. Bratislava: STU, Bratislava, 2009. pp. 667-682, 2009.

[3.] HREBICEK, J.; CHVATALOVA, Z. *Chosen Features of System Maple 13 in Education of Mathematics, Engineering and Applied Research*. In M. Kovacova (ed.), *Proceedings of 9th International Conference APLIMAT 2010.* Bratislava, Slovak Republic February 2-5, 2010. Bratislava: STU, Bratislava, 2010. pp. 749-764, 2010.

[4.] HREBICEK J., et al. *Teaching Mathematical Modelling in Biology Using Maple*. In. J. Hrebicek (ed.), *Proceedings of the 7th Summer School: Applied Informatics.* Bedrichov, Czech Republic September 3-5, 2010. Selected papers. Brno: Masaryk University, 2010, pp. 47 – 62, 2010.

[5.] CHVATALOVA, Z.; GAJDOKOVA, L. *Computer Support Applications of Mathematical Foundation in Logistics*. In. J. Hrebicek (ed.), *Proceedings of the 7th Summer School: Applied Informatics.* Bedrichov, Czech Republic September, 3-5 2010. Selected papers. Brno: Masaryk University, 2010, pp. 63 – 77, 2010.

[6.] MONAGAN, M. B.; GEDDES, K. O.; HEAL, K. M.; LABAHN, G.; VORKOETTER, S. M.; McCarron, J. & DeMarco, P. *Introductory Programming Guide*. Waterloo : Maplesoft, 2010.
[7.] MARIK, M. *Valuation of companies* (in Czech). 1st ed. Prague : Ekopress, 2005.
[8.] RUCKOVA, P. *Financial analysis: methods, indicators, used in practice* (in Czech). 3rd update ed. Prague : Grada Publishing, 2009.
[9.] ZIGARDY, M. *Program for Analyzing Economical Data via Mathematical Modeling in Maple (master's thesis in Czech)*, supervisor RNDr. Zuzana Chvatalova, Ph.D.). Brno (CZ): FBM BUT (CZ), 2010.

Current addresses

Jiri Hrebicek, Prof., RNDr., CSc.
Institute of Biostatistics and Analyses Faculty of Science, Masaryk University Brno,
Adress: Kotlarska 2, 611 37 Brno, Czech Republic
Phone number: +420 603217052, e-mail: hrebicek@iba.muni.cz
Zuzana Chvatalova, RNDr., Ph.D.
Institute of Informatics Faculty of Business and Management, Brno University of Technology
Adress: Kolejni 4, 612 00 Brno, Czech Republic
Phone number: +420 541142658, e-mail: chvatalova@fbm.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number I

POTENTIAL OF CAS FOR DEVELOPMENT OF MATHEMATICAL THINKING

JANČAŘÍK Antonín, (CZ), NOVOTNÁ Jarmila, (CZ)

Abstract. The paper presents the potential of the use of Computer Algebra Systems in mathematics education, especially in solving problems whose standard solution goes beyond the students' abilities, be it due to technical demands (making use of the computational power of computers) or due to high difficulty of the problem (using the possibility to model the solution numerically). The problems on which both these situations are illustrated are designed for higher secondary school students but can also be successfully used in pre-service teacher training.

Key words. CAS, Numerical Models, Computers in Education

Mathematics Subject Classification: Primary 97N80; Secondary 34K28.

1 Introduction

The use of computers in mathematics education on primary and secondary schools (e.g. [1], [2]) and in pre-service teacher training ([3], [4]) is a very up-to-date topic. Computers have become tool of motivation and foster comprehensible interdisciplinary links between mathematics and other subjects, including project teaching ([5]). However, the use of computers in teaching asks for new approaches to exposition and to the mathematical content ([6]). This might be one of the reasons why recent studies in Mathematics education show that, despite many national and international actions aiming at integration of ICT into mathematics classrooms, such integration in schools remains underdeveloped. The rate of this integration increases markedly slowly when compared to the speed of evolution of the technology.

One of the causes of this state is the huge diversity of ICT resources that often leaves teachers unsure of which to use, and when and how to use them. Another important retarder of successful use of ICT in teaching is a lack of information on the potential, advantages and dangers of inclusion of activities using ICT into teaching. All this despite the fact that the benefits of their use in mathematics education are well known (see e.g. [7]). The paper aspires to contribute to a decrease in the above described problems connected to implementation of ICT into mathematics education.

Integration of computers into teaching should always be governed by the principle of its efficiency. It is appropriate to use computers only in such situations in which it really brings benefit – opens new possibilities or significantly decreases the amount of time needed for technical calculations. The aim of this paper is to demonstrate the potential of both of these ways of use of ICT on two concrete problems. Both problems were designed for pre-service mathematics teacher training as well as for work with 12-19 year old students.

The first problem shows the demonstrational potential of ICT and intuitive approach to numerical solutions. It was developed within the frame of EdUmatic project. Despite the fact that the solution of the problem requires a thorough knowledge of integral calculus, it is within the frame of this project presented to 12-15 year old students. The problem also functions as a tool for introducing students and pre-service teachers to modeling (confer [1]) of real situations using numerical models.

The second problem demonstrates the use of the considerable computational power of CAS when looking for the deviation of an erroneous solution of a problem from the correct result.

2 Problem 1

The first problem is an example of a cascade of problems of increasing difficulty which can be solved as a project in several lessons.

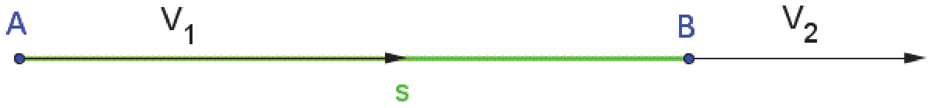
Motivation to the first task is the solution of the classical problem of the type bullet-target. The problem is assigned to secondary school students in different modifications. The students' task is to simulate an optimum path of the bullet under a set of given conditions – the initial point, speed of the bullet and the speed and trajectory of the target.

The difficulty of the tasks increases. To begin with, the students are assigned problems that can be solved using mathematics taught at secondary school. However, this is certainly no limit to the potential of this problem – activities carried out in this problem can be naturally stretched to university mathematics, which will be demonstrated in this paper.

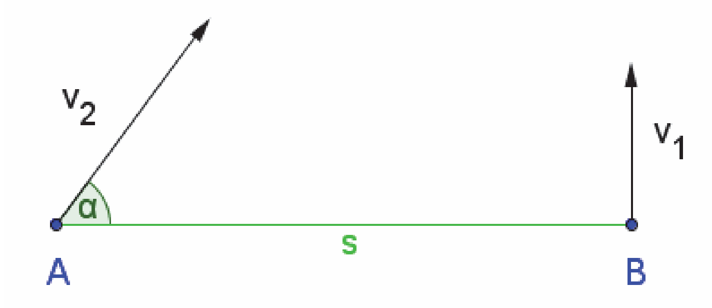
The main activity consists in the choice of the trajectory of the running and the following calculation of the bullet trace approximation. The main aim of the activity for the students is to learn about graphs of various (first of all trigonometric) functions in an indirect way (e.g. they want the target to zigzag or run around).

The activity in its basic design has a ***preparatory stage***. It consists of solving Problems 1 to 3, each solution followed by a discussion. Extensions 1 and 2 represent activities that give the activity a more general significance – they model situations from various domains of life and science.

Problem 1: There are two runners running one after the other at different speeds.

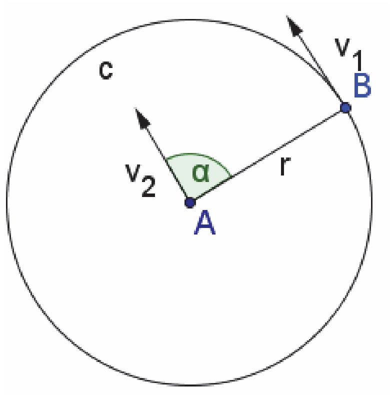


Problem 2: The first runner runs past the second runner in the distance s at the speed v_1; where should the second runner be heading if he is running at the speed v_2 and wants to catch the first runner? This may be interpreted as the chase of a flying target.



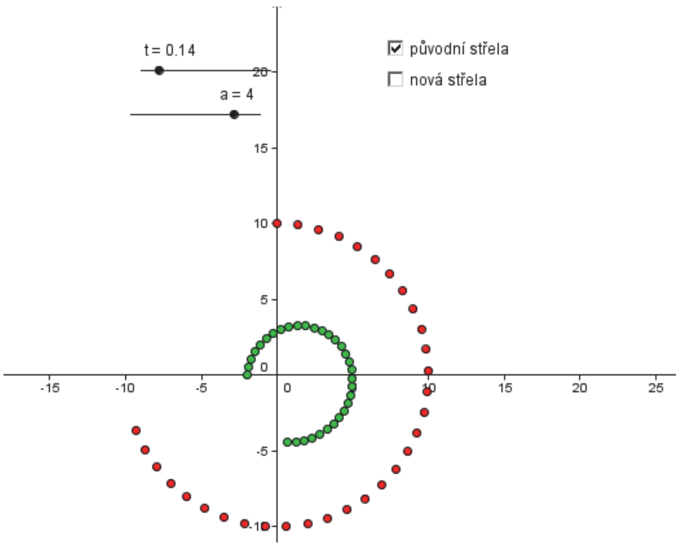
Desperado wants to raid a stagecoach, travelling speed 30 km per hour, which it passes at a distance of 40 km. When he assaults if he can ride at 50 km per hour?

Problem 3: The first runner overtakes the second one at the speed v on a circle of the radius r. Where should the first runner be hitting?

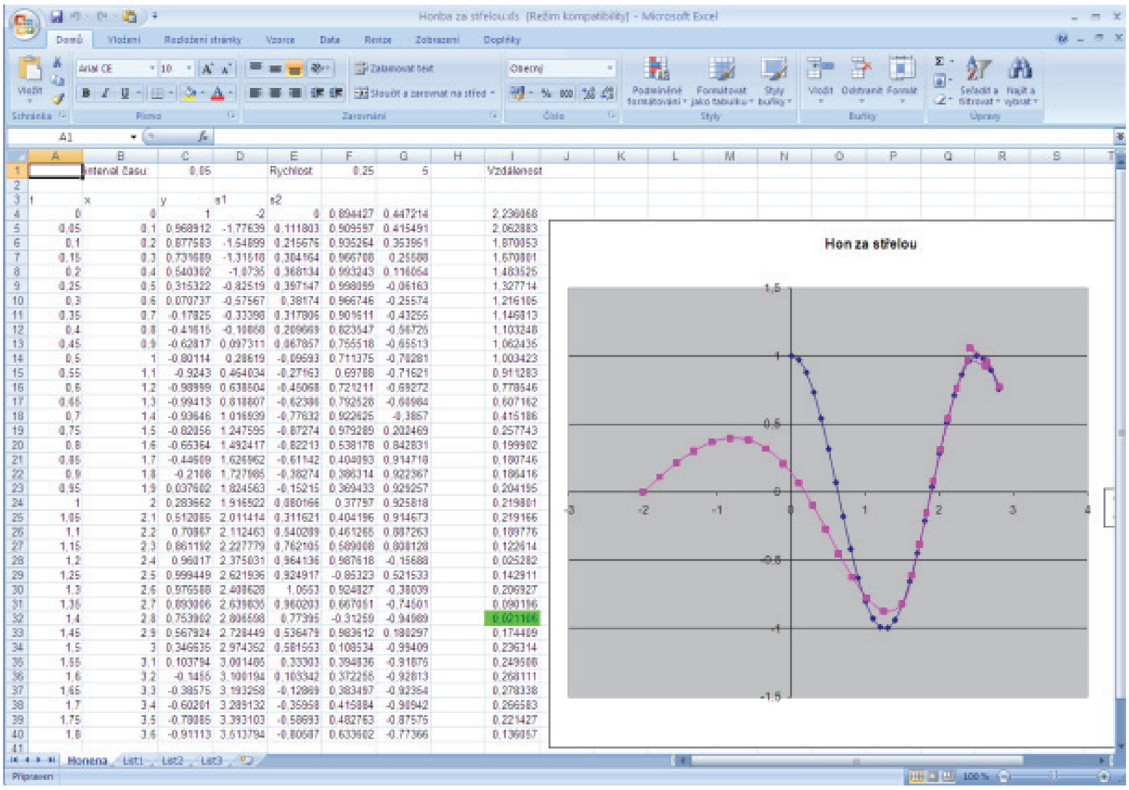


The activity requires the following mathematical knowledge: linear functions, Theorem of Pythagoras, circumference, length of a circular arch. The use of ICT is not required for this stage. In case of the necessity to solve a quadratic equation, it is possible to use a computer/calculator or spreadsheets. Students work with concrete numbers.

The activity itself can be introduced by the following questions posed by the teacher: What happens with the trajectories it the target changes the direction? How should the runner/bullet behave, under which conditions can it/he hit/catch the target? Where should the runner/bullet direct?



Subsequently, the students are assigned more complicated tasks that cannot be solved outright. These problems demand that the bullet continuously directs at the target and the solution is constructed numerically. In given intervals, the bullet redirects on the target and modifies its path. The solution is modeled with the help of various tools – MS Excel, Geogebra, TI Nspire. In our experiments, the students most often used MS EXCEL ([8]). The students' task is not only to model the bullet path, but also to model the requested trajectories of the target with the help of known mathematical formulas.



Extension 1: What about the situation when the target moves on a sine curve?

In more complicated cases, the paper-and-pencil calculations of the curve become more and more complex. The necessity to use ICT arises naturally from the development of the activity.

Let us illustrate the development of derivation of the formula for programming the bullet's path:

Step 1: From the origin ([0, 0]) to e.g. the point [3, 1] at the speed 1. Where will the runner be after the first step (the interval of the length 1)?

Step 2: Change the origin to another point (translation)

Step 3: From a general point to a general point.

Using the formula, the students themselves create a model of the situation in the selected program. They assign the target a position and the program calculates the positions of the bullet. The data are stored in a table with 4 columns (two for the position of the bullet, two for the target). At the beginning, both positions are entered manually; later students program the change of the position of the target and the program calculates the position of the bullet. Students propose various paths for the target and represent them by a function (e.g. a piecewise linear function, absolute value, trigonometric functions with various parameters); they observe the changes in the path of the bullet. They look for the changes of parameters for e.g. making the graph of the target "higher", "denser" etc.

There are various extensions of this activity. Let us mention some examples of suitable questions:

The bullet moves at a constant speed.

- What is the speed of the target?
- Is it also constant?
- Is it possible to calculate it?
- When is the speed lowest/highest?
- What are the changes in acceleration when the speed changes?

Students may observe/analyze the differences between the cases of constant speed and varying speed, the influence of the slope of the function, the relationship between the tangent and the instantaneous speed etc. The meaning of a point of inflexion as the transition between speeding-up and slowing-down becomes evident.

If we look at the three levels of "media competence" in solving problems with the use of ICT, i.e. knowing how to handle it, how to use it to solve a certain problem, and being able to choose and modify the tool according to the given problem, all three are present in the introduced activity. It does not mean that all students have to go through all its stages. It depends on the level of their competences both in mathematics and use of ICT that the teacher wants to develop in his/her students.

3 Problem 2

Motivation to the second problem is the following classical problem about a watering can: *We have a weekend house in point D close to a river bank represented by the line segment AB and a greenhouse in* point *C. ABCD is a square. The task is to take an empty watering can from the weekend house, fill it in the river and carry it to the greenhouse. The price for carrying an empty watering can from point D to point Q is 1 CZK/m, the price for carrying a full watering can from* *point Q to point C is c CZK/m. In which point Q on the river bank shall we fill the watering can if we want the price y for transport of water into the greenhouse to be minimum?*

3.1 Solution using the program Mathematica

Šišma in his paper [9] proposes the following solving procedure:
The total price equals $y = \sqrt{1+x^2} + c\sqrt{1+(1-x)^2}$. The sought minimum is the solution of the equation $y' = \frac{x}{\sqrt{1+x^2}} - \frac{c\cdot(1-x)}{\sqrt{1+(1-x)^2}} = 0$.

The last equality may be adjusted to an algebraic quartic equation. Here the author of the original article states that *the solution of the equation in dependence on the parameter is extremely difficult and therefore we must be satisfied with approximate calculations of roots for different values of the parameter c.*

Let us now study how difficult the solution of this problem is if we use also modern computer technologies in addition to classical methods. However, before we proceed to an exact solution, let us try to find an approximate solution. The main obstacle complicating the solution are the roots used for the calculation of the total price. When differentiated they get under another radical and make the problem extremely complicated. The solution that we get cannot be but erroneous. The question is how significant a mistake this erroneous adjustment causes. Instead of making calculation for the function *y*, we must look for minimum for $\overline{y} = 1 + x^2 + c(1+(1-x)^2)$. In that case we get a linear equation $\overline{y} = 2x - 2c(1-x) = 0$ .instead of a quartic equation to be differentiated. The solution to this linear equation is trivial. It is $x = \frac{c}{1+c}$.

Now it is important to find how large a mistake there is in the gained solution. To answer this question it is inevitable to use computer technologies. In this case the program Mathematica version 7.0.0. First we define the functions *reseni (solution)* and *odhad (estimation)* that in dependence on parameter c produce exact and approximate solutions of the problem:

```
In[1]:= koren = Solve[- c (1 - x)/Sqrt[1 + (1 - x)^2] + x/Sqrt[1 + x^2] == 0, x][[3]];
        reseni[c_] = x /. koren;
        odhad[c_] = c / (1 + c);
```

The program finds all complex roots from which we select the third root, which takes real values for the values $c \in (1,\infty)$. The evaluation of the root in dependence on the parameter *c* is quite difficult (see following figure). The complexness of the solution suggests that it is virtually impossible to carry the corresponding “paper-and-pencil” calculation and use of computer technology is inevitable. CAS systems have no difficulties working with these types of expressions. It is no problem to carry out the following calculations.

```
In[4]:= reseni[c]
```

Out[4]=

$$\frac{1}{2}+\frac{1}{2}\sqrt{-\frac{1}{3}+\frac{2\,2^{2/3}\left(-1+3c^2-3c^4+c^6\right)^{1/3}}{3\left(-4-15c^2+15c^4+4c^6+3\sqrt{3}\sqrt{8c^2-5c^4-6c^6-5c^8+8c^{10}}\right)^{1/3}}+\frac{2^{1/3}\left(-4-15c^2+15c^4+4c^6+3\sqrt{3}\sqrt{8c^2-5c^4-6c^6-5c^8+8c^{10}}\right)^{1/3}}{3\left(-1+3c^2-3c^4+c^6\right)^{1/3}}}-$$

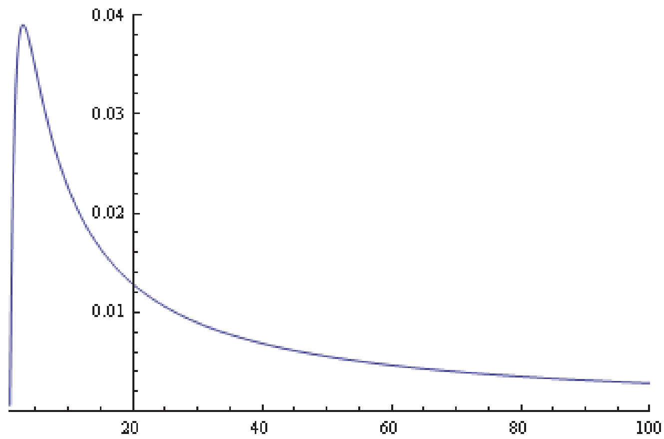
$$\frac{1}{2}\sqrt{\left(-\frac{2}{3}-\frac{2\,2^{2/3}\left(-1+3c^2-3c^4+c^6\right)^{1/3}}{3\left(-4-15c^2+15c^4+4c^6+3\sqrt{3}\sqrt{8c^2-5c^4-6c^6-5c^8+8c^{10}}\right)^{1/3}}-\frac{2^{1/3}\left(-4-15c^2+15c^4+4c^6+3\sqrt{3}\sqrt{8c^2-5c^4-6c^6-5c^8+8c^{10}}\right)^{1/3}}{3\left(-1+3c^2-3c^4+c^6\right)^{1/3}}+\right.}$$

$$\left.\frac{-8+\frac{16c^2}{-1+c^2}}{4\sqrt{-\frac{1}{3}+\frac{2\,2^{2/3}\left(-1+3c^2-3c^4+c^6\right)^{1/3}}{3\left(-4-15c^2+15c^4+4c^6+3\sqrt{3}\sqrt{8c^2-5c^4-6c^6-5c^8+8c^{10}}\right)^{1/3}}+\frac{2^{1/3}\left(-4-15c^2+15c^4+4c^6+3\sqrt{3}\sqrt{8c^2-5c^4-6c^6-5c^8+8c^{10}}\right)^{1/3}}{3\left(-1+3c^2-3c^4+c^6\right)^{1/3}}}}\right)$$

Let us now find out how significant the mistake stemming from the use of the solution $x=\frac{c}{1+c}$ is. We first construct a graph showing the difference between the approximate and exact solution in dependence on the parameter c:

```
In[5]:= Plot[reseni[c] - odhad[c], {c, 1.01, 100}, PlotRange → {{1, 100}, {0, 0.04}}]
```

Out[5]=



This graph clearly shows that the maximum difference is in the neighborhood of number 3. Using the command

```
In[6]:= maxvzd = FindRoot[reseni'[c] - odhad'[c] == 0, {c, 3}]

Out[6]= {c → 3.03277}
```

we get a more exact information that it is the case of the value of the parameter c approximately 3.03277. The difference for this parameter c is about 0.038987, i.e. less than four percent of the length of side AB.

```
In[7]:= N[reseni[c] - odhad[c]] /. maxvzd

Out[7]= 0.038987
```

However, this value is not an answer to the original question how different the price for transport of water is. If we want to state this difference in prices, both prices must be substituted into the original formula $y = \sqrt{1+x^2} + c\sqrt{1+(1-x)^2}$. For these ends we define two new functions that calculate (again in dependence on the parameter *c*) the minimum price and the price from approximate solution:

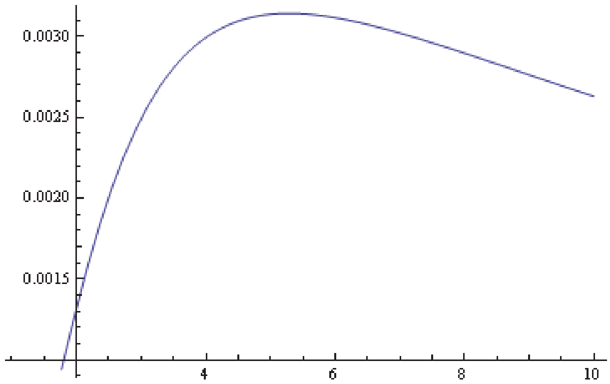
$$\text{In[8]:=}\ \mathtt{PCena[c\_]} := \sqrt{1 + \mathtt{reseni[c]}^2} + c\sqrt{1 + (1 - \mathtt{reseni[c]})^2}$$

$$\mathtt{OCena[c\_]} := \sqrt{1 + \mathtt{odhad[c]}^2} + c\sqrt{1 + (1 - \mathtt{odhad[c]})^2}$$

The relation of the difference of prices is again expressed as a graph in dependence on the parameter *c*.

```
In[10]:= Plot[{OCena[c] - PCena[c]}, {c, 1.1, 10}]
```

Out[10]=



With the help of the program Mathematica we find the values of parameter *c*, for which the difference of prices for a unit square is maximum and we evaluate the maximum value:

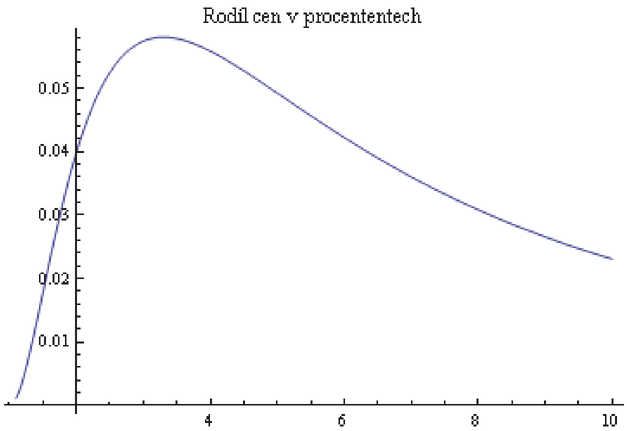
```
In[11]:= maxcn = FindRoot[OCena'[c] - PCena'[c] == 0, {c, 5}]
         N[OCena[c] - PCena[c]] /. maxcn

Out[11]= {c → 5.28419}

Out[12]= 0.00314158
```

The maximum difference of prices is 0.003 CZK and this is the case for the value of parameter `c = 5.28419`.

To finish this work we find out what percentage deviation in price we get if we leave out the roots in the original equation. The results can be again shown in a graph.

```
In[13]:= CenyProc[c_] := 100 (OCena[c] - PCena[c]) / PCena[c]
         Plot[CenyProc[c], {c, 1.1, 10}, PlotLabel → "Rodil cen v procententech"]
```

Out[14]=



This graph clearly shows that the difference in prices does not exceed six per mille of the total price. Therefore the approximation is very exact.

The result may be elaborated by a concluding calculation in which we use the function CenyProc defined in the previous step:

```
In[15]:= maxvpr = FindRoot[CenyProc'[c] == 0, {c, 3}]
         CenyProc[c] /. maxvpr

Out[15]= {c → 3.29978}

Out[16]= 0.0580615
```

4 Conclusion

The paper demonstrates the potential of the use of Computer Algebra Systems for solution of problems from calculus, which go beyond the abilities of most students. The use of programs such as Mathematica or Geogebra makes it possible to find a solution in the situation when pupils or students are not able to find an exact solution on their own, or when finding such a solution is technically extremely demanding.

The presented activities are aimed at improving the motivation, engagement and achievement of students of mathematics through rich mathematical tasks that use the best of the available technology in creative and innovative ways. It should result in extending students' depth and level of mathematics achievement which will have a positive effect on their personal development.

Acknowledgement
This research was supported by the project EdUmatics – European Development for the Use of Mathematics Technology in Classrooms, 503254-LLP-1-2009-1-UK-COMENIUS-CMP.

References

[1] GOLDENBERG, E.P.: *Thinking (and talking) about technology in math classrooms.* Issues in Mathematics Education. On-line : http://www2.edc.org/mcc/PDF/iss_tech.pdf, 2000.

[2] RUTHVEN, K.: In Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education (CERME), University of Cyprus, pp. 52-67, 2007.

[3] TURKER, B., SAGLAM, Y., UMAY, A.: *Preservice teachers' performances at mathematical modeling process and views on mathematical modelin*g. In Innovation and creativity in education, Vol. 2, Issue 2, pp. 4622-4628, 2010.

[4] GARCÍA-CAMPOS, M., ROJANO, T.: *Appropriation processes of CAS: A multidimensional study with secondaty school mathematics teachers.* In PME 32 and PME-NA XXX 2008, Mexico: Cinvestav-UMSNH, Vol. 1, p. 260, 2008.

[5] VANÍČEK, J.: *Mathematics and project-based learning employing technologies.* South Bohemia Mathematical Letters, Vol. 17, No. 1, pp. 77 – 92, 2010.

[6] ARTIGUE, M.: *Learning mathematics in a CAS environment: the genesis of a reflection about instrumentation and the dialectics between technical and conceptual work.* In International Journal of Computers for Mathematics Learning, n°7, pp. 245-274, 2002.

[7] AKTUMEN, M., KACAR, A. *Effects of computer algebra systems on attitutes towards mathematics.* In Hacettepe Universitesi Egitim Fakultesi Dergisi-Hacettepe University Journal Of Education, Issue: 35, pp. 13-26, 2008.

[8] JANČAŘÍK, A., HOŠPESOVÁ, A., DVOŘÁK, P.: *Využití programu MS Excel v práci učitele matematiky*. UK v Praze, PedF, Praha, 2007.

[9] [9] ŠIŠMA, P.: *Úloha o konvi*. In RMF 83/2, pp. 1-11, 2008.

Current address

RNDr. Antonín Jančařík, Ph.D.
Charles University in Prague
Faculty of Education, Dep. of Mathematics a Mathematics Education
M. D. Rettigové 4
116 39, Prague 1
Czech Republic
Tel: +420 221 900 252
e-mail: Antonin.Jancarik@pedf.cuni.cz

Doc. RNDr. Jarmila Novotná, CSc.
Charles University in Prague
Faculty of Education, Dep. of Mathematics a Mathematics Education
M. D. Rettigové 4
116 39, Prague 1
Czech Republic
Tel: +420 221 900 251
e-mail: Jarmila.Novotna@pedf.cuni.cz

Journal of Applied Mathematics

volume IV (2011), number I

MATRICES IN SIMPLE ECONOMIC TASKS

ULRYCHOVÁ Eva, (CZ)

Abstract. Several simple economic applications of matrices that are convenient as an illustration of use of linear algebra tools (eigenvector, Markov chains, systems of linear equations, least squares approximation) in math courses or textbooks are provided.

Key words and phrases. matrices, economic applications, least squares approximation.

Mathematics Subject Classification. Primary 97H60; Secondary 97D40.

1 Introduction

Math teachers often have to face questions asked by their students: "What do I need this for?" This is not the only reason why some simple applications of mathematical terms are convenient to show in mathematical courses. However, not only the lack of time in the courses is a problem. The mathematical courses usually anticipate other special classes, therefore the applications should not require any special knowledge. Solved problems and data given in examples can be simplified in comparison with real values – it is important to show possibilities of applications rather than to obtain an accurate result. On the other hand, it is interesting to include some real problems if possible.

This article provides several applications of the matrices used in economics.

2 Applications

2.1 Markov Chains, Eigenvector

Formulation of the problem. Suppose a market research monitoring a group of 300 people, 200 of them use a product A and 100 use a product B. In any month 80 % of product A users continue to use it and 20 % switch to the product B and 90 % of product B users continue to use it and 10 % switch to the product A. The percentages of users loyal to the original product are assumed to be constant in next months – it means the probability of changing from one product to the other is always the same. That is a simple example of so called Markov chains.

Following this research we determine, how many people will be using each product one resp. two resp. k month later, and estimate the state in the long run.

Solution. The number of product A users after one month is given by the following formula

$$0,8 \cdot 200 + 0,1 \cdot 100 = 170,$$

since 80 % of 200 A users (that is $0,8 \cdot 200$) stay with A and in addition 10 % of 100 B users (that is $0,1 \cdot 100$) convert to A.

Similarly, the number of product B users is given by the following formula

$$0,2 \cdot 200 + 0,9 \cdot 100 = 130.$$

We can rewrite these two formulas using the matrix (so called transition matrix)

$$\mathbf{T} = \begin{pmatrix} 0,8 & 0,1 \\ 0,2 & 0,9 \end{pmatrix}$$

(an entry T_{ij} represents the probability of moving from state corresponding to i to state corresponding to j):

$$\begin{pmatrix} 0,8 & 0,1 \\ 0,2 & 0,9 \end{pmatrix} \begin{pmatrix} 200 \\ 100 \end{pmatrix} = \begin{pmatrix} 170 \\ 130 \end{pmatrix}.$$

If we denote $\mathbf{x_0} = (200, 100)^T$ (so called initial vector) and $\mathbf{x_1} = (170, 130)^T$, we can write

$$\mathbf{x_1} = \mathbf{T}\mathbf{x_0}.$$

Numbers of each of A and B users after one month are given by the components of the vector $\mathbf{x_1}$ (let us note that these components are not necessarily integers – they are only approximations of numbers of people).

Similarly to computing numbers of users after one month we determine numbers of users after two months (represented by the vector $\mathbf{x_2}$):

$$\mathbf{x_2} = \mathbf{T}\mathbf{x_1} = \begin{pmatrix} 0,8 & 0,1 \\ 0,2 & 0,9 \end{pmatrix} \begin{pmatrix} 170 \\ 130 \end{pmatrix} = \begin{pmatrix} 149 \\ 151 \end{pmatrix}$$

(we can also write $\mathbf{x_2} = \mathbf{T}\mathbf{x_1} = \mathbf{T}(\mathbf{T}\mathbf{x_0}) = \mathbf{T}^2\mathbf{x_0}$).

It is obvious that numbers of A and B users after k months are determined by

$$\mathbf{x_k} = \mathbf{T}\mathbf{x_{k-1}} \qquad \text{or} \qquad \mathbf{x_k} = \mathbf{T}^k\mathbf{x_0}$$

(an entry $(T^k)_{ij}$ of this matrix $\mathbf{T}^k$ represents the probability of moving from state corresponding to i to state corresponding to j in k transitions).

It is possible to show (see [2]) that if the transition matrix is a matrix which has some power positive then vectors $\mathbf{x_k}$ (so called state vectors) converge for large k to a unique vector $\mathbf{x}$ (so called steady state vector); when this vector is reached, it will not change by multiplying by $\mathbf{T}$:

$$\mathbf{x} = \mathbf{Tx}$$

(it means that $\mathbf{T}$ has 1 as an eigenvalue and the steady state vector is one of eigenvectors corresponding to this eigenvalue). Moreover, steady state vector does not depend on the choice of the initial vector $\mathbf{x_0}$.

To determine numbers of A and B users after a long time we compute the steady state vector $\mathbf{x}$:

$$\mathbf{x} = \mathbf{Tx} \iff \mathbf{Tx} - \mathbf{x} = \mathbf{o} \iff (\mathbf{T} - \mathbf{I})\mathbf{x} = \mathbf{o}$$

($\mathbf{I}$ is a unit matrix). Since

$$\mathbf{T} - \mathbf{I} = \begin{pmatrix} -0,2 & 0,1 \\ 0,2 & -0,1 \end{pmatrix},$$

we obtain the following homogeneous linear system (represented by the augmented matrix):

$$\left(\begin{array}{rr|r} -0,2 & 0,1 & 0 \\ 0,2 & -0,1 & 0 \end{array}\right) \sim \left(\begin{array}{rr|r} -0,2 & 0,1 & 0 \end{array}\right) \sim \left(\begin{array}{rr|r} -2 & 1 & 0 \end{array}\right)$$

the general solution of which is $\mathbf{x} = (t, 2t)^T$, $t \in R$. Since components of each of state vectors represent numbers of A resp. B users, the sum of these components must be equal to the global number of users; in our case components of the steady stage vector must satisfy

$$t + 2t = 300,$$

from which it follows that $t = 100$ and $\mathbf{x} = (100, 200)^T$.

After a long time, 100 people will be using the product A and 200 people will be using the product B (and – as was said – this result does not depend on the initial distribution of A and B users).

2.2 Eigenvector

Formulation of the problem. Suppose that three producers (denoted by P_1, P_2, P_3) organized in a simple closed society produce three commodities c_1, c_2, c_3. Each of these producers sells and buys from each other, all their products are consumed by them and no other commodities enter the system (the closed model – see [1]). The proportions of the products consumed by each of P_1, P_2, P_3 are given in the Table 1:

	c_1	c_2	c_3
P_1	0,7	0,4	0,5
P_2	0,1	0,5	0,2
P_3	0,2	0,1	0,3

Table 1

For example, the first column lists that 70 % of the produced commodity c_1 are consumed by P_1, 10 % by P_2 and 20 % by P_3. It is obvious, that the sum of elements in each column is equal to 1.

We compute what proportions of producers' incomes have to be to ensure that this society survives.

Solution. If we denote x_1, x_2, x_3 the incomes of the producers P_1, P_2, P_3, then the amount spent by P_1 on c_1, c_2, c_3 is $0,7x_1 + 0,4x_2 + 0,5x_3$. The assumption that the consumption of each producer equals his income leads to the equation $0,7x_1 + 0,4x_2 + 0,5x_3 = x_1$, similarly for producers P_2, P_3. We obtain the system of linear equations

$$\begin{array}{ccccccc} 0,7x_1 & + & 0,4x_2 & + & 0,5x_3 & = & x_1 \\ 0,1x_1 & + & 0,5x_2 & + & 0,2x_3 & = & x_2 \\ 0,2x_1 & + & 0,1x_2 & + & 0,3x_3 & = & x_3. \end{array}$$

This system can be rewritten as the equation $\mathbf{Ax} = \mathbf{x}$, where

$$\mathbf{A} = \begin{pmatrix} 0,7 & 0,4 & 0,5 \\ 0,1 & 0,5 & 0,2 \\ 0,2 & 0,1 & 0,3 \end{pmatrix} \qquad \text{and} \qquad \mathbf{x} = (x_1, x_2, x_3)^T,$$

(it means that $\mathbf{A}$ has 1 as an eigenvalue and $\mathbf{x}$ is an eigenvector corresponding to this eigenvalue). Moreover – since we assume the income to be nonnegative, the components x_i must be nonnegative, i.e. $x_i \geq 0$ for $i = 1, 2, 3$. We can rewrite the equation $\mathbf{Ax} = \mathbf{x}$ into the equivalent form:

$$\mathbf{Ax} = \mathbf{x} \iff \mathbf{Ax} - \mathbf{x} = \mathbf{o} \iff (\mathbf{A} - \mathbf{I})\mathbf{x} = \mathbf{o}.$$

That is a homogeneous linear system (for more comfortable computing we multiply each row of the matrix $\mathbf{A} - \mathbf{I}$ by 10) :

$$\left(\begin{array}{rrr|r} -0,3 & 0,4 & 0,5 & 0 \\ 0,1 & -0,5 & 0,2 & 0 \\ 0,2 & 0,1 & -0,7 & 0 \end{array}\right) \sim \left(\begin{array}{rrr|r} -3 & 4 & 5 & 0 \\ 1 & -5 & 2 & 0 \\ 2 & 1 & -7 & 0 \end{array}\right) \sim \cdots \sim \left(\begin{array}{rrr|r} 1 & 0 & -3 & 0 \\ 0 & 1 & -1 & 0 \end{array}\right).$$

The general solution of the system is $\mathbf{x} = t(3, 1, 1)^T$, $t \in R$; according to the condition $x_i \geq 0$ we choose $t \geq 0$. This result means, that to ensure that the society survives, the incomes of the producers P_1, P_2, P_3 have to be in proportions 3:1:1.

2.3 System of Linear Equations, Least Squares Approximation

Formulation of the problem. Average salaries in Czech Republic for the years 1990, 1995, 2000 and 2005 (in thousands of crowns) are given in the Table 2 (source: [3]). The values are rounded in order to compute easily without calculator.

	thousands of crowns
1990	3
1995	8
2000	13
2005	18

Table 2

Based on these data we estimate the average salary in 2010.

Solution. We consider the time to be an independent variable and the salary a dependent variable – we obtain a function of one variable. If we denote the year 1990 as 0, 1995 as 1, 2000 as 2 and 2005 as 3, then we have four points $[0, 3]$, $[1, 8]$, $[2, 13]$, $[3, 18]$ in plain. We are looking for a function whose graph approximately goes through these points. If we draw the picture, we can see that the points seem to lie on a straight line (Figure 1).

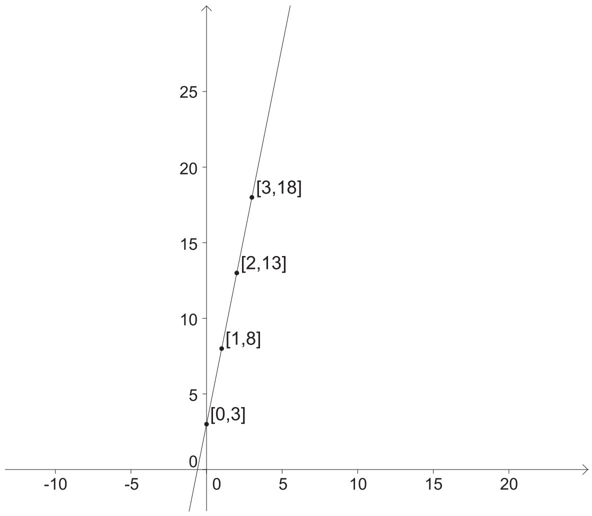


Figure 1

We will try to find a linear function so that coordinates of the given points satisfy the equation of this function $y = ax + b$:

$$\begin{array}{rcrcr} a \cdot 0 & + & b & = & 3 \\ a \cdot 1 & + & b & = & 8 \\ a \cdot 2 & + & b & = & 13 \\ a \cdot 3 & + & b & = & 18. \end{array}$$

We obtained the system of linear equations (represented by the augmented matrix)

$$\left(\begin{array}{cc|c} 0 & 1 & 3 \\ 1 & 1 & 8 \\ 2 & 1 & 13 \\ 3 & 1 & 18 \end{array}\right) \sim \cdots \sim \left(\begin{array}{cc|c} 1 & 1 & 8 \\ 0 & 1 & 3 \end{array}\right).$$

the solution of which is $a = 5$, $b = 3$. It is realy surprising that the linear system has a solution – the given points lie exactly on a stright line[1]! The equation of this line is

$$y = 5x + 3.$$

To estimate the average salary in 2010 we substitute $x = 4$ (the value corresponding to the year 2010 – the next to the value corresponding to the year 2005) into the equation $y = 5x + 3$ and obtain $y = 23$ (let us note that this value corresponds to the real value in this year).

Formulation of the problem. The ratios of households having computers (in percentage) are given in the Table 3 (source: [4]). The values are rounded in order to compute easily without calculator.

	%
1990	3
1995	7
2000	21
2005	42

Table 3

We find the least squares approximating quadratic for these data, compute the norm of the least squares error and estimate the ratio of households having computers in the year 2010.

Solution. As in the previous problem we denote the year 1990 as 0, 1995 as 1, 2000 as 2 and 2005 as 3. In this case it is obvious that the points $[0, 3]$, $[1, 7]$, $[2, 21]$, $[3, 42]$ do not lie on a straight line (Figure 2).

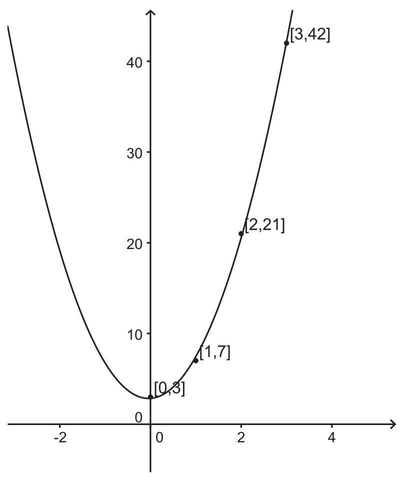


Figure 2

We will try to find a parabola that gives the least squares approximation to the given points. Substituting these points into the equation of a parabola $y = ax^2 + bx + c$, we obtain the system

[1] In the next problem we will show how to proceed if given points are not laying on a stright line.

of linear equations

$$\begin{array}{ccccccc} a\cdot 0 & + & b\cdot 0 & + & c & = & 3 \\ a\cdot 1 & + & b\cdot 1 & + & c & = & 7 \\ a\cdot 4 & + & b\cdot 2 & + & c & = & 21 \\ a\cdot 9 & + & b\cdot 3 & + & c & = & 42 \end{array}$$

represented by the augmented matrix

$$\left(\begin{array}{ccc|c} 0 & 0 & 1 & 3 \\ 1 & 1 & 1 & 7 \\ 4 & 2 & 1 & 21 \\ 9 & 3 & 1 & 42 \end{array}\right) \sim \cdots \sim \left(\begin{array}{ccc|c} 1 & 1 & 1 & 7 \\ 0 & 2 & 3 & 7 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 3 \end{array}\right).$$

This system is inconsistent – the four given points do not lie on a parabola. To find a parabola that fits the points as possible, we use the method called least squares approximation (see [2]). The last squares solution of the system $\mathbf{Ax} = \mathbf{b}$ ($\mathbf{A}_{m\times n}$, $\mathbf{x} \in R^n$, $\mathbf{b} \in R^m$) is the vector $\tilde{\mathbf{x}}$ such that

$$||\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}|| \le ||\mathbf{b} - \mathbf{A}\mathbf{x}|| \quad \text{for all} \quad \mathbf{x} \in R^n$$

and can be obtained as a solution of the equation (so called normal equation)

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}.$$

In our case we have

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 3 \\ 7 \\ 21 \\ 42 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

We compute

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix} 0 & 1 & 4 & 9 \\ 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 98 & 36 & 14 \\ 36 & 14 & 6 \\ 14 & 6 & 4 \end{pmatrix}$$

$$\mathbf{A}^T\mathbf{b} = \begin{pmatrix} 0 & 1 & 4 & 9 \\ 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 7 \\ 21 \\ 42 \end{pmatrix} = \begin{pmatrix} 469 \\ 175 \\ 73 \end{pmatrix}$$

and the normal equation takes the form

$$\begin{pmatrix} 98 & 36 & 14 \\ 36 & 14 & 6 \\ 14 & 6 & 4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 469 \\ 175 \\ 73 \end{pmatrix}$$

that is the system of linear equations

$$\left(\begin{array}{ccc|c} 98 & 36 & 14 & 469 \\ 36 & 14 & 6 & 175 \\ 14 & 6 & 4 & 73 \end{array}\right).$$

The solution of this system is

$$a = 4,25, \quad b = 0,35, \quad c = 2,85 \qquad \text{otherwise} \qquad \tilde{\mathbf{x}} = (4,25,\ 0,35,\ 2,85)^T$$

and the required equation of parabola is

$$y = 4,25x^2 + 0,35x + 2,85.$$

To obtain the least squares error

$$||\mathbf{e}|| = ||\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}||$$

we compute the product $\mathbf{A}\tilde{\mathbf{x}}$ at first:

$$\mathbf{A}\tilde{\mathbf{x}} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 2,85 \\ 7,45 \\ 20,55 \\ 42,15 \end{pmatrix}.$$

Then

$$||\mathbf{e}|| = ||\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}|| = ||(3,7,21,42)^T - (2,85,\ 7,45,\ 20,55,\ 42,15)^T|| =$$

$$= ||(0,15,\ -0,45,\ 0,45,\ -0,15)^T|| = \sqrt{0,0225 + 0,2025 + 0,2025 + 0,0225} = \sqrt{0,45}.$$

We estimate the ratio of households having computers in the year 2010 substituting $x = 4$ into the equation $y = 4,25x^2 + 0,35x + 2,85$: $\quad y = 72,25 \cong 72$. Thus, using the data in the table 3 the ratio of households having computers is supposed to be 72%.

3 Conclusions

It is not quite easy to find sufficiently simple applications, because the requirement of their simplicity is rather limiting. To use the tools of linear algebra for solving some real problems it is usually necessary to have deeper knowledge of linear algebra or other branches of science.

We provided a few examples motivated by tasks in economy. Their solutions illustrate the power of linear algebra tools such as eigenvectors, matrix formalism, Markov chains, system of linear equations, least squares approximation. These examples can be used in mathematical courses or textbooks for high school or bachelor students at non-technical universities.

References

[1] FRIEDBERG, S. H., Insel, A. J., Spence, L. E.: *Linear Algebra.* Prentice Hall, 2003
[2] POOLE, D.: *Linear Algebra: A Modern Introduction.* Brooks/Cole, 2003.
[3] http://www.kurzy.cz/makroekonomika/mzdy/
[4] http://czso.cz/csu/redakce.nsf/i/cr_od_roku_1989#05

Current address

Eva Ulrychová, RNDr.
University of Economics
Department of Mathematics
Ekonomická 957, 148 00 Prague 4
tel. number: 224 094 234
e-mail: uleva@vse.cz