

ISSN: 1337-6365
© Slovak University of Technology in Bratislava
All rights reserved

APLIMAT – JOURNAL OF APPLIED MATHEMATICS

VOLUME 3 (2010), NUMBER 3

APLIMAT – JOURNAL OF APPLIED MATHEMATICS

VOLUME 3 (2010), NUMBER 3

Edited by: Slovak University of Technology in Bratislava

Editor - in - Chief: KOVÁČOVÁ Monika (Slovak Republic)

Editorial Board: CARKOVŠ Jevgenijs (Latvia)
CZANNER Gabriela (Great Britain)
CZANNER Silvester (Great Britain)
DE LA VILLA Augustin (Spain)
DOLEŽALOVÁ Jarmila (Czech Republic)
FEČKAN Michal (Slovak Republic)
FERREIRA M. A. Martins (Portugal)
FRANCAVIGLIA Mauro (Italy)
KARPÍŠEK Zdeněk (Czech Republic)
KOROTOV Sergey (Finland)
LORENZI Marcella Giulia (Italy)
MESIAR Radko (Slovak Republic)
TALAŠOVÁ Jana (Czech Republic)
VELICHOVÁ Daniela (Slovak Republic)

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

Correspondence concerning subscriptions, claims and distribution:

F.X. spol s.r.o
Azalková 21
821 00 Bratislava
journal@aplimat.com

Frequency: One volume per year consisting of three 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

Copyright © STU 2007-2010, Bratislava

All rights reserved. No part may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior written permission from the Editorial Board. All contributions published in the Journal were reviewed with open and blind review forms with respect to their scientific contents.

APLIMAT – JOURNAL OF APPLIED MATHEMATICS

VOLUME 3 (2010), NUMBER 3

ENGINEERING APPLICATIONS AND SCIENTIFIC COMPUTATIONS

BEŇOVÁ Mariana, DOBRUCKÝ Branislav, MARČOKOVÁ Mariana: ON SPECIFIC UTILIZATION OF INFINITE SERIES FOR PERIODICAL NON-HARMONIC FUNCTIONS IN ELECTRICAL ENGINEERING	15
BÖHM Patrik: MODIFICATIONS TO IV SETUP OF STREAM CIPHER SNOW 2.0	25
ČERNÁ Dana, FIŇEK Václav: CONSTRUCTION OF INTERVAL CUBIC SPLINE-WAVELET BASES ON COARSE SCALES	35
JANČO Roland, KOVÁČOVÁ Monika: USING PROGRAM MATHEMATICA FOR SOLUTION OF BEAM ON ELASTIC FOUNDATION	47
KOVÁČOVÁ Monika, JANČO Roland: LARGE SERIES-PARALLEL STRUCTURES AND METHODS USABLE FOR APPROXIMATE THEIR MTTF IN REPAIRABLE SYSTEMS	55
KRIŽAN Peter, ŠOOŠ Ľubomír, MATÚŠ Miloš, SVÁTEK Michal, VUKELIČ Djordje: EVALUATION OF MEASURED DATA FROM RESEARCH OF PARAMETERS IMPACT ON FINAL BRIQUETTES DENSITY	69
MAJSTOROVIČ Snježana: K-DOMINATION SETS ON DOUBLE LINEAR HEXAGONAL CHAINS	77
MATÚŠ Miloš, KRIŽAN Peter: INFLUENCE OF STRUCTURAL PARAMETERS IN COMPACTING PROCESS ON QUALITY OF BIOMASS PRESSINGS	87
NAVRÁTIL Vladislav, NOVOTNÁ Jiřina: SOLID SOLUTION HARDENING IN CADMIUM SINGLE CRYSTALS	97
POTŮČEK Radovan: COMPUTATION OF BOUNDARIES AND SUMS OF REDUCED HARMONIC SERIES FORMED BY ZERO AND ANOTHER DIGIT	103
SKORKOVSKY Petr: MINIMALIST IMPLEMENTATION OF A GENETICALLY PROGRAMMING PROCESS WRITTEN IN ASSEMBLY LANGUAGE	113

APLIMAT – JOURNAL OF APPLIED MATHEMATICS

VOLUME 3 (2010), NUMBER 3

NEW TRENDS IN MATHEMATICS EDUCATION

BOHÁČ Zdeněk, DOLEŽALOVÁ Jarmila, KREML Pavel: STATISTICAL ANALYSIS II OF RESULTS OF PROJECT ESF STUDY SUPPORTS WITH PREVAILING DISTANCE FACTORS FOR SUBJECTS OF THE THEORETICAL BASE FOR STUDY	131
JUKIC Ljerka: DIFFERENCES IN REMEMBERING CALCULUS CONCEPTS IN UNIVERSITY SCIENCE STUDY PROGRAMMES	137
ORSZÁGHOVÁ Dana: THE USE OF INFORMATION TECHNOLOGY IN MATHEMATICS EDUCATION	147

APLIMAT – JOURNAL OF APPLIED MATHEMATICS

VOLUME 3 (2010), NUMBER 3

STATISTICAL METHODS IN TECHNICAL AND ECONOMIC SCIENCES AND PRACTICE

ANDRADE Marina, FERREIRA Manuel Alberto M.: CIVIL IDENTIFICATION PROBLEMS WITH BAYESIAN NETWORKS USING OFFICIAL DNA DATABASES	155
BARTOŠOVÁ Jitka, FORBELSKÁ Marie: MIXTURE MODEL CLUSTERING FOR HOUSEHOLD INCOMES	163
BARTOŠOVÁ Jitka, BÍNA Vladislav: SOME FACTORS AFFECTING EXPENDITURE ON HOUSING IN THE CZECH REPUBLIC	173
BARTOŠOVÁ Jitka, NOVÁK Michal: INCOME AND EXPENDITURE OF CZECH HOUSEHOLDS	171
BUSS Ginters: ECONOMIC FORECASTS WITH BAYESIAN AUTOREGRESSIVE DISTRIBUTED LAG MODEL: CHOOSING OPTIMAL PRIOR IN ECONOMIC DOWNTURN	191
DOLEŽALOVÁ Jarmila, VALOVÁ Marie: CHEMICAL ANALYSIS OF TREE BARK COMPOSITION IN NON-BALLAST REGIONS	201
FERREIRA Manuel Alberto M. , ANDRADE Marina: $M G _{\infty}$ SYSTEM TRANSIENT BEHAVIOR WITH TIME ORIGIN AT THE BEGINNING OF A BUSY PERIOD MEAN AND VARIANCE	207
FERREIRA Manuel Alberto M., FILIPE José António: SOLVING LOGISTICS PROBLEMS QUEUE SYSTEMS BUSY PERIOD $\infty G $ USING M	213
CHVOSTEKOVÁ Martina: SIMULTANEOUS TOLERANCE INTERVALS IN A LINEAR REGRESSION	223
JANKOVÁ Mária: CONFIDENCE INTERVAL FOR COMMON MEAN - A COMPARISON OF TWO METHODS	231
JAROŠOVÁ Eva: ESTIMATION WITH THE LINEAR MIXED EFFECTS MODEL	239
KLŮFA Jindřich: NEW LTPD PLANS FOR INSPECTION BY VARIABLES - CALCULATION AND ECONOMICAL ASPECTS	247

POBOČÍKOVÁ Ivana: THE SIMPLE ALTERNATIVES OF THE CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO BINOMIAL PROPORTIONS	257
TREŠL Jiří: APPLICATION OF NEURAL NETWORKS IN FINANCE	269

APLIMAT – JOURNAL OF APPLIED MATHEMATICS

LIST OF REVIEWERS

Andrade Marina , Professor Auxiliar	ISCTE - Lisbon University Institute, Lisboa, Portugal
Ayud Sani Muhammed	NGO, University of Cape Coast, Accra
Barbagallo Annamaria , Dr.	University of Catania, Italy
Bartošová Jitka , RNDr., PhD.	University of Economics Prague, Jindřichův Hradec, Czech Republic
Bayaraa Nyamish ,	Ulaanbaatar, Mongolia
Beránek Jaroslav , Doc. RNDr. CSc.	Masaryk University, Brno, Czech Republic
Brueckler Franka Miriam , DrSc.	University of Zagreb, Zagreb, Poland
Budíková Marie , RNDr., Dr.	Masarykova univerzita, Brno, Czech Republic
Budkina Natalja , Dr. math	Riga, Latvia Technical university, Riga, Latvia
Buy Dmitriy B. , PhD.	Kyiv Taras Shevchenko National University, Kyiv
Conversano Elisa	Architettura, Roma, Italy
Csatáryová Mária , RNDr., PhD.	University of Prešov, Prešov, Slovak Republic
Daniele Patrizia , Associate Professor	University of Catania, Italy,
Dobrákovová Jana , Doc., RNDr., CSc.	Slovak University of Technology in Bratislava, Slovak Republic
Doležalová Jarmila , Doc., RNDr., CSc.	VŠB - TU, Ostrava, Czech Republic
Dorociaková Božena , RNDr., PhD.	University of Žilina, Žilina, Slovak Republic
Doupovec Miroslav , doc. RNDr.. CSc.	University of Technology, Brno, Czech Republic
Ferreira Manuel Alberto M. , Professor Catedrático	ISCTE- Lisbon University Institute, Lisboa, Portugal
Filipe José António , Prof. Auxiliar	ISCTE - Lisbon University Institute, Lisboa, Portugal

Francaviglia Mauro , Full Prof.	University of Calabria, Calabria, Italy
Franců Jan , prof. RNDr., CSc.	University of Technology, Brno, Czech Republic
Fulier Jozef , prof. RNDr. CSc.	UKF Nitra, Slovak Republic
Habiballa Hashim , RNDr., PaedDr., PhD.	University of Ostrava, Ostrava, Czech Republic
Hennyeyová Klára , doc. Ing. CSc.	Slovak University of Agriculture, Nitra, Slovak Republic
Hinterleitner Irena , PhD.	University of Technology, Brno, Czech Republic
Hošková Šárka , Doc. RNDr., PhD.	University of Defence, Brno, Czech Republic
Hunka Frantisek , Doc. (Assoc. Prof)	University of Ostrava, Ostrava, Czech Republic
Chvosteková Martina , Mgr.	Slovak Academy of Science, Bratislava, Slovak Republic
Jan Chvalina , Prof. DrSc.	University of Technology, Brno, Czech Republic
Janková Mária , Mgr.	Slovak Academy of science, Bratislava, Slovak Republic
Jukic Ljerka	University of Osijek, Osijek, Croatia
Kalina Martin , Doc. RNDr., CSc.	Slovak University of Technology in Bratislava, Slovak Republic
Kapička Vratislav , Prof. RNDr., DrSc.	Masaryk university, Brno, Czech Republic
Klobucar Antoaneta ,	University in Osijek, Osijek, Croatia
Klůfa Jindřich , prof. RNDr., CSc.	University of Economics, Prague, Czech Republic
Kováčová Monika , Mgr., PhD.	Slovak University of Technology in Bratislava, Slovak Republic
Kulčár Ladislav , Doc. RNDr. CSc.	M. Bel University Banská Bystrica, Poprad, Slovak Republic
Kureš Miroslav , Assoc. Prof.	Brno, Czech Republic University of Technology, Brno, Czech Republic
Kvasz Ladislav , Doc. PhD. (Assoc. Prof)	University Komenskeho, Bratislava, Slovak Republic
Lopes Ana Paula , PhD.	Polytechnic Institute of Oporto – IPP, Oporto, Portugal
Marcella Giulia Lorenzi , prof.	Università della Calabria, Calabria, Italy

Lungu Nicolae , Prof.	Technical University of Cluj- Napoca, Cluj-Napoca, Romania
Majstorović Snježana	J.J.Strossmayer University, Osijek, Croatia
Mařík Robert , doc. PhD.	Mendel University in Brno, Czech Republic, Brno, Czech Republic
Matvejevs Andrejs , Ing., DrSc.	Riga, Latvia Technical university, Riga, Latvia
Michálek Jiří , RNDr, CSc.	Czech Academy of Sciences, Prague, Czech Republic
Mikeš Josef , Prof. RNDr., DrSc.	Palacky University, Olomouc Czech Republic
Myšková Helena , RNDr., PhD.	Technical University of Košice, Košice
Nemoga Karol , doc. RNDr., PhD.	Slovak Academy of Sciences, Bratislava, Slovak Republic
Neuman František , Prof., RNDr., DrSc.	Mathematical Institute of the Academy, Brno, Czech Republic
Omachelová Milada , Mgr., PhD	Slovak University of Technology in Bratislava, Slovak Republic
Paun Marius , Prof. Dr.	Transilvania University of Brasov, Brasov, Romania
Pavlenko Oksana , Dr. math	Riga, Latvia Technical university, Riga, Latvia
Plavka Ján	Technical University Košice, Košice, Slovak Republic
Plch Roman , RNDr., PhD.	Masaryk University, Brno, Czech Republic
Pokorný Michal , Prof.	University of Zilina, Zilina, Slovak Republic
Potůček Radovan , RNDr., PhD.	University of Defense, Brno, Czech Republic
Půlpán Zdeněk , Prof., RNDr., PhD.r., CSc.	University of Hradec Králové, Hradec Králové, Czech Republic
Ráčková Pavlína , PhD.r., PhD.	University of Defense, Brno, Czech Republic
Rus Ioan A. , Professor	Babes-Bolyai University, Cluj-Napoca, Romania
Růžicková Miroslava , RNDr. Doc. CSc.	University of Žilina, Slovak Republic, Žilina, Slovak Republic
Segeth Karel , Prof., RNDr., CSc.	Academy of Sciences, Praha, Czech Republic
Stankovičová Iveta , Ing., PhD.	Univerzita Komenského, Bratislava, Slovak Republic

Sterba Jan , Ing.	University of Economics in Bratislava, Slovak Republic
Svatošová Libuše , Prof. Ing., CSc.	Czech University of Life Sciences, Praha, Czech Republic
Šooš Lubomír , Prof. Ing., PhD.	Slovak University of Technology in Bratislava, Slovak Republic
Such Ondrej , PhD.	Inštitút Matematiky a Informatiky, Banská Bystrica, Slovak Republic
Tarabella Leonello , Researcher	Research Area pf CNR, Pisa, Italy
Terek Milan , prof., Ing., PhD.	University of Economics, Bratislava, Slovak Republic
Trokanova Katarina , Doc.	Slovak University of Technology in Bratislava, Slovak Republic
Tseytlin Georgiy E. , PhD.	Institute of Program System , Kyiv, Ukraine
Urban František , Doc. Ing., CSc.	Slovak University of Technology, Bratislava, Slovak Republic
Vanžurová Alena , Doc. RNDr., CSc.	Palacký University, Olomouc, Czech Republic
Velichová Daniela , Doc. RNDr., CSc.	Slovak University of Technology, Bratislava, Slovak Republic
Wikovsky Viktor , Doc. RNDr., CSc.	Slovak Academy of Sciences, Bratislava, Slovak Republic
Wimmer Gejza , Mgr.	Slovak Academy of Sciences, Bratislava, Slovak Republic

ON SPECIFIC UTILIZATION OF INFINITE SERIES FOR PERIODICAL NON-HARMONIC FUNCTIONS IN ELECTRICAL ENGINEERING

BEŇOVÁ Mariana, (SK), DOBRUCKÝ Branislav, (SK), MARČOKOVÁ Mariana, (SK)

Abstract. The paper deals with using of infinite series for analysis and determination of root-mean-square values and harmonic distortion factor of the periodical non-harmonic function. The investigated quantities are output voltages and currents of power electronic converter supplying resistive-inductive or capacitive load. Those voltages are strongly non-harmonic ones, so they must be pulse-modulated due to requested nearly sinusoidal currents with low requested total harmonic distortion. New method of calculation based on the theory of numerical series verified by actual examples is presented in the paper.

Key words: infinite series, Fourier series, periodical non-harmonic function, total harmonic distortion, root-mean-square value.

Mathematics Subject Classification: Primary: 42A16, 42A20; Secondary: 42A24.

1 Introduction

It is well known that for a 2π -periodic function $f(x)$ that is integrable on $[-\pi, \pi]$, the series

$$\frac{a_0(f)}{2} + \sum_{n=1}^{\infty} [a_n(f) \cos(nx) + b_n(f) \sin(nx)]$$

is called the Fourier series of $f(x)$ on $[-\pi, \pi]$ and the numbers $a_n(f), b_n(f)$ are called the Fourier coefficients of $f(x)$ on $[-\pi, \pi]$ (cf. [2] and [7]). One introduces the partial sums of the Fourier series for $f(x)$, often denoted by

$$(S_N f)(x) = \frac{a_0(f)}{2} + \sum_{n=1}^N [a_n(f) \cos(nx) + b_n(f) \sin(nx)] ,$$

where $N \geq 0$ is integer. These partial sums are trigonometric polynomials. One expects that the functions $S_N f$ approximate the function $f(x)$ and that the approximation is improved when N tends to infinity.

Knowing properties of non-harmonic function it is possible to determine harmonic distortion of the investigated function. It will be done using Fourier and other infinite series.

2 Harmonic analysis of output voltage of the voltage sourced inverter

Let us assume square-wave non-harmonic periodical functions of the output voltage with simple width control (Fig. 1).

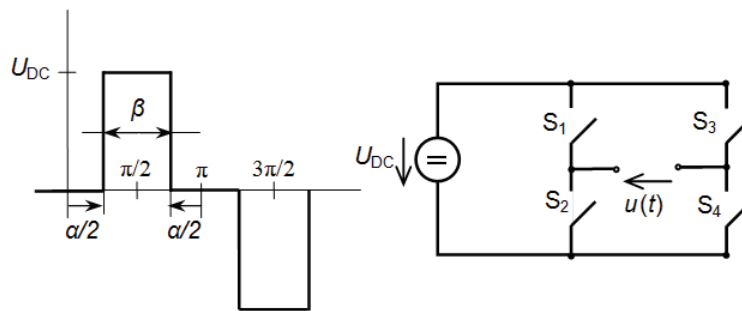


Fig. 1. Circuit diagram of inverter and its output voltage waveform, where α is control angle, β is pulse-width, U_{DC} is supply voltage

Using Fourier analysis (cf. [5], [7]) we obtain relation for amplitudes of harmonic components

$$A_v = \frac{4}{\pi} \frac{1}{v} U_{DC} \cos\left(v \frac{\alpha}{2}\right),$$

where v is odd number equal to $2n+1$, $n=0,1,2,3,\dots$, α is control angle and U_{DC} is supply voltage. So, Fourier series of the voltage $u(t)$ is following (cf. [3], [5]):

$$u(t) = \frac{4}{\pi} U_{DC} \sum_{v=1}^{\infty} \left[\frac{\cos(v \alpha/2)}{v} \sin(v \omega t) \right],$$

where ω is the angular frequency and t is the time.

From the mathematical point of view the Fourier series does not always converge, and even when it does converge for a specific value t_0 of t , the sum of the series at t_0 may differ from the value $f(t_0)$ of the function. It is one of the main questions in harmonic analysis to decide when Fourier series converge, and when the sum is equal to the original function. If a function is square-integrable on the interval $[-\pi, \pi]$ then the Fourier series converges to the function at almost every point (cf. [2], [7]).

Checking convergence for e.g. $t = \frac{\pi}{2\omega}$, $\alpha = 0$ (thus $\beta = \pi$) we have

$$u\left(\frac{\pi}{2\omega}\right) = \frac{4}{\pi} U_{DC} \sum_{v=1}^{\infty} \left[\frac{\cos 0}{v} \sin\left(v \frac{\pi}{2}\right) \right] = U_{DC},$$

because (cf. [7])

$$\sum_{\nu=1}^{\infty} \left[\frac{1}{\nu} \sin\left(\nu \frac{\pi}{2}\right) \right] = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots = \frac{\pi}{4}.$$

When $t = \frac{\pi}{\omega}$, the Fourier series converges to 0, which is the half-sum of the left- and right-limit of $u(t)$ at $t = \frac{\pi}{\omega}$ (according to the Dirichlet theorem for Fourier series).

Root-mean-square value (RMS) of each harmonic component will be

$$U_{\nu} = \frac{A_{\nu}}{\sqrt{2}} = \frac{2\sqrt{2}}{\pi} U_{\text{DC}} \frac{1}{\nu} \cos\left(\nu \frac{\alpha}{2}\right)$$

and RMS value of total voltage waveform

$$U = \sqrt{\sum_{\nu=1}^{\infty} U_{\nu}^2} = \sqrt{\sum_{\nu=1}^{\infty} \left[\frac{2\sqrt{2}}{\pi} U_{\text{DC}} \frac{1}{\nu} \cos\left(\nu \frac{\alpha}{2}\right) \right]^2} = \frac{2\sqrt{2}}{\pi} U_{\text{DC}} \sqrt{\sum_{\nu=1}^{\infty} \frac{1}{\nu^2} \cos^2\left(\nu \frac{\alpha}{2}\right)}.$$

For $\alpha = 0$ ($\beta = \pi$) we have

$$U = \frac{2\sqrt{2}}{\pi} U_{\text{DC}} \sqrt{\sum_{\nu=1}^{\infty} \frac{1}{\nu^2}} = \frac{2\sqrt{2}}{\pi} U_{\text{DC}} \sqrt{\frac{\pi^2}{8}} = U_{\text{DC}},$$

where (cf. [7])

$$\sum_{\nu=1}^{\infty} \frac{1}{\nu^2} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \frac{1}{9^2} + \dots = \frac{\pi^2}{8}.$$

For $\alpha = \frac{\pi}{6}$ ($\beta = \frac{2\pi}{3}$) we have

$$U = \frac{2\sqrt{2}}{\pi} U_{\text{DC}} \sqrt{\sum_{\nu=1}^{\infty} \frac{1}{\nu^2} \cos^2\left(\nu \frac{\pi}{6}\right)} = \frac{2\sqrt{2}}{\pi} U_{\text{DC}} \sqrt{\frac{\pi^2}{12}} = \frac{2\sqrt{2}}{\pi} U_{\text{DC}} \frac{\pi}{\sqrt{12}} = U_{\text{DC}} \sqrt{\frac{2}{3}},$$

where

$$\sum_{\nu=1}^{\infty} \frac{1}{\nu^2} \cos^2\left(\nu \frac{\pi}{6}\right) = \frac{3/4}{1} + 0 + \frac{3/4}{5^2} + \frac{3/4}{7^2} + 0 + \dots = \frac{3}{4} \left(\sum_{\nu=1}^{\infty} \frac{1}{\nu^2} - \frac{1}{3^2} \sum_{\nu=1}^{\infty} \frac{1}{\nu^2} \right) = \frac{3}{4} \frac{\pi^2}{8} \left(1 - \frac{1}{3^2} \right) = \frac{\pi^2}{12}.$$

3 Determination of the voltage total harmonic distortion factor

Total harmonic distortion of the voltage THD_U (Fig. 1) will be

$$THD_U = \sqrt{\left(\frac{U}{U_1}\right)^2 - 1} = \sqrt{\frac{\sum_{v=1}^{\infty} \frac{1}{v^2} \cos^2\left(v \frac{\alpha}{2}\right)}{\cos^2\left(\frac{\alpha}{2}\right)} - 1}.$$

For $\alpha = 0$ ($\beta = \pi$) we have

$$THD_U / \pi = \sqrt{\sum_{v=1}^{\infty} \frac{1}{v^2} - 1} = \sqrt{\frac{\pi^2}{8} - 1} = 0,48, \text{ i.e. } 48\%.$$

For $\alpha = \frac{\pi}{6}$ ($\beta = \frac{2\pi}{3}$) we have

$$THD_U / \frac{2\pi}{3} = \sqrt{\frac{\sum_{v=1}^{\infty} \frac{1}{v^2} \cos^2(v \cdot \pi/6)}{\cos^2(\pi/6)} - 1} = \sqrt{\frac{\pi^2/12}{(\sqrt{3}/2)^2} - 1} = \sqrt{\frac{4}{3} \cdot \frac{\pi^2}{12} - 1} = 0,31, \text{ i.e. } 31\%.$$

4 Determination of the current total harmonic distortion factor – the case of knowing steady-state current time waveforms

Time waveforms for supposed loads are shown in Fig. 2.

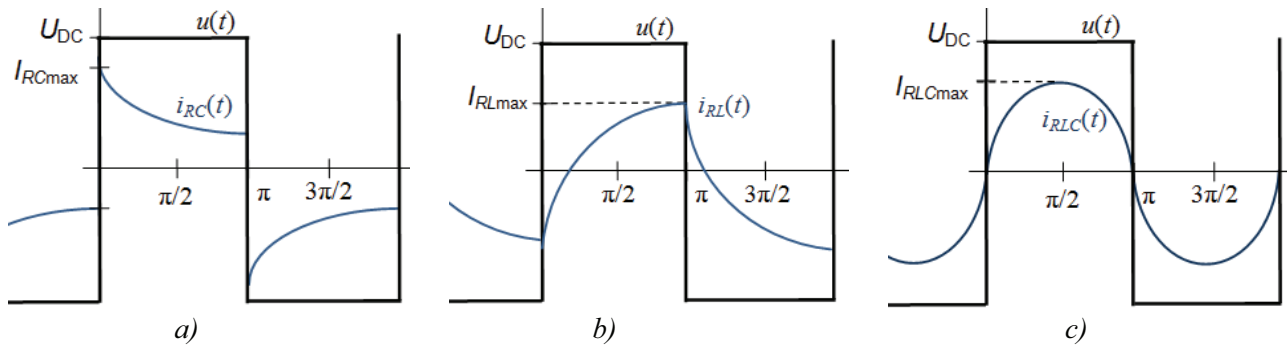


Fig. 2. Steady-state current time waveforms for different supposed loads: a) for resistive-capacitance load, b) for resistive-inductive load and c) for resistive-inductive-capacitance load under resonance condition.

Supposing R-C load and $\alpha = 0$ ($\beta = \pi$) the current i_{RC} during positive half period will be

$$i_{RC}(t) = I_{RCmax} e^{-t/\tau} = \frac{U_{Rmax}}{R} e^{-t/\tau} = \frac{U_{DC} + U_{C(0)}}{R} e^{-t/\tau},$$

where $t \in [0, \pi]$, $\tau = RC$ and

$$U_{C(0)} = U_{DC} \frac{1 - e^{-T/2\tau}}{1 + e^{-T/2\tau}} = U_{DC} \tanh \frac{T}{4\tau}.$$

Using time discretization $t \rightarrow n\Delta T$

$$i(n\Delta T) = \frac{U_{Rmax}}{R} e^{-(\Delta T/\tau)n},$$

where $\Delta T = \frac{T}{360}$ and $n = 0, 1, 2, 3, \dots, 180$. Then RMS value of the current i_{RC} can be numerically expressed by the [6] using geometric series as

$$I = \frac{U_{Rmax}}{R} \sqrt{\frac{2}{N} \sum_{n=1}^{180} e^{-(2\Delta T/\tau)n}} = \frac{U_{DC}}{R} \sqrt{\frac{2}{N} \sum_{n=1}^{180} r^{-n}},$$

where $r = e^{(2\Delta T/\tau)}$ and quotient $q = \frac{1}{r}$. Partial sum of this series is

$$\sum_{n=0}^N \frac{1}{r^n} = 1 + \frac{1}{r^1} + \frac{1}{r^2} + \frac{1}{r^3} + \dots + \frac{1}{r^N} = \frac{r^N - 1}{r^{N-1}(r - 1)}.$$

Based on above equations, the total harmonic distortion of current I is

$$THD_1 = \sqrt{\left(\frac{I}{I_1}\right)^2 - 1} = \sqrt{\frac{(1 + 0,462)^2 \frac{2}{N} \sum_{n=1}^{180} r^{-n}}{\left(\frac{2\sqrt{2}}{\pi} \frac{1}{\sqrt{1 + 1/\pi^2}}\right)^2} - 1}, \quad (1)$$

where $I_1 = \frac{2\sqrt{2}}{\pi|Z_{RC}|} U_{DC}$ and $|Z_{RC}| = \sqrt{R^2 + (1/\omega C)^2}$,

and for $\tau = \frac{T}{2}$ and $\omega = \frac{2\pi}{T}$ will be $\frac{|Z_{RC}|}{R} = \sqrt{1 + (1/\pi)^2}$, $r = e^{1/90}$

and

$$U_{C(0)} = -U_{DC} \frac{1 - e^{-1}}{1 + e^{-1}} = -U_{DC} \tanh \frac{1}{2} \doteq -0,462 U_{DC}.$$

So, we can finally write for current THD_1

$$THD_1 = \sqrt{\frac{(1 + 1/\pi^2) \cdot 1,462^2 \cdot \frac{1}{180} \frac{(1 - e^{-2})}{(1 - e^{-1/90})}}{8/\pi^2} - 1} \doteq 0,374, \text{ i. e. } 37,4\%.$$

Supposing R - L load and $\alpha = 0$ ($\beta = \pi$) the current i_{RL} during positive half - period will be

$$i_{RL}(t) = I_{RLmax} (1 - e^{-t/\tau}) + i_{RL}(0) e^{-t/\tau} = \frac{U_{DC}}{R} (1 - e^{-t/\tau}) + i_{RL}(0) e^{-t/\tau} = \frac{U_{DC}}{R} \left[1 - \left(1 + \tanh \frac{T}{4\tau} \right) \right] e^{-T/\tau}.$$

According waveforms in Fig. 2b

$$i_{RL}(0) = -i_{RL}(T/2) = \frac{U_{R\max}}{R} \frac{(1 - e^{-T/2\tau})}{(1 + e^{-T/2\tau})} = -\frac{U_{DC}}{R} \tanh \frac{T}{4\tau},$$

where $\tau = \frac{L}{R}$ and $\tanh x$ can also be developed into infinite series

$$\tanh x = x - \frac{1}{3}x^3 + \frac{2}{12}x^5 - \frac{17}{315}x^7 + \dots \quad (|x| < \pi/2).$$

RMS value of total current waveform

$$\begin{aligned} I &= \sqrt{\frac{2}{T} \int_0^{T/2} i^2(t) dt} = \frac{U_{DC}}{R} \sqrt{\frac{2}{T} \int_0^{T/2} \left(1 - \left[1 + \tanh \frac{T}{4\tau} \right] \cdot e^{-t/\tau} \right)^2 dt} = \\ &= \frac{U_{DC}}{R} \sqrt{\frac{2}{T} \int_0^{T/2} \left(1 - 2 \left[1 + \tanh \frac{T}{4\tau} \right] \cdot e^{-t/\tau} + \left[1 + \tanh \frac{T}{4\tau} \right]^2 \cdot e^{-2t/\tau} \right) dt}. \end{aligned}$$

For simplification let us take $\tau = \frac{T}{2}$. Thus $\frac{T}{4\tau} = \frac{1}{2}$ and $\tanh \frac{1}{2} \doteq 0,462$

$$I \doteq \frac{U_{DC}}{R} \sqrt{\frac{2}{T} \int_0^{T/2} (1 - 2,924 \cdot e^{-2/T} + 1,462^2 \cdot e^{-4/T}) dt}$$

and after discretization

$$I_{RMS} / \frac{\beta=\pi}{\tau=T/2} \doteq \frac{U_{DC}}{R} \sqrt{\frac{2}{N} \sum_{n=1}^{N/2} 1 - 2,924 \frac{2}{N} \sum_{n=1}^{N/2} r_1^{-n} + 1,462^2 \frac{2}{N} \sum_{n=1}^{N/2} r_2^{-n}}$$

where $N = 360$ and

$$r_1^{-n} = e^{-(2/T)\Delta T n} = e^{-(2/T)(T/360)n} = e^{-(1/180)n}, \quad r_2^{-n} = e^{-(4/T)\Delta T n} = e^{-(4/T)(T/360)n} = e^{-(1/90)n}.$$

Then

$$I_{RMS} / \frac{\beta=\pi}{\tau=T/2} \doteq \frac{U_{DC}}{R} \sqrt{\frac{2}{N} \frac{N}{2} - 2,924 \frac{2}{N} \frac{1 - e^{-1}}{1 - e^{-1/180}} + 1,452^2 \frac{2}{N} \frac{1 - e^{-2}}{1 - e^{-1/90}}}.$$

Based on above equations the total harmonic distortion of current is

$$THD_1 \doteq \frac{U_{DC}}{R} \sqrt{\frac{\frac{2}{N} \left(\frac{N}{2} - 2,924 \frac{1 - e^{-1}}{1 - e^{-1/180}} + 1,452^2 \frac{1 - e^{-2}}{1 - e^{-1/90}} \right)}{\frac{8}{\pi^2} \frac{1}{1 + \pi^2}}} < 0,22, \quad \text{i.e. } 22\%, \quad (2)$$

$$\text{where } I_{1RMS} = \frac{2\sqrt{2}}{\pi} \frac{U_{DC}}{R} \frac{1}{\sqrt{1 + \pi^2}}.$$

Supposing R - L - C load and $\alpha = 0$ ($\beta = \pi$), the current i_{RLC} during positive half period will be (under condition of resonance and high quality of the load circuit). The current waveform i_{RLC} is nearly harmonic one (i.e. sinusoidal, depending on the quality of the circuit), i.e.

$$i_{RLC}(t) = I_{RLC\max} \sin(\omega_0 t) \cdot e^{-t/\tau} = \frac{U_{DC}}{\omega_0 L} \sin(\omega_0 t) \cdot e^{-t/\tau}.$$

Let $\omega_0 = \frac{2\pi}{T}$ and $\omega_0 = \sqrt{\frac{1}{LC} - \frac{1}{\tau^2}}$, then RMS value of total current waveform

$$I = \sqrt{\frac{2}{T} \int_0^{T/2} i^2(t) dt} = \sqrt{\frac{2}{T} \int_0^{T/2} I_{RLC\max}^2 \sin^2(\omega t) \cdot e^{-2t/\tau} dt} = I_{RLC\max} \sqrt{\frac{2}{N} \int_{n=1}^{N/2} \sin^2\left(\frac{\pi}{180} n\right) \cdot e^{-(1/90)n} dn}$$

and total harmonic distortion factor *THD* is very small, less than 2%.

5 Determination of the current total harmonic distortion factor - case of without knowing steady-state current time waveforms

Supposing *R-L*, load the current i_{RL} can be expressed using Fourier series as

$$i(t) = \frac{4}{\pi} U_{DC} \sum_{v=1}^{\infty} \left[\frac{\cos(v\alpha/2)}{\sqrt{R^2 + (v\omega L)^2}} \frac{\sin(v\omega t - \varphi_v)}{v} \right]$$

with RMS value of each current harmonic component

$$I_v = \frac{I_{v\max}}{\sqrt{2}} = \frac{2\sqrt{2}}{\pi} U_{DC} \frac{\cos(v\alpha/2)}{v\sqrt{R^2 + (v\omega L)^2}}$$

and RMS value of total current waveform

$$I = \sqrt{\sum_{v=1}^{\infty} I_v^2} = \sqrt{\sum_{v=1}^{\infty} \left[\frac{2\sqrt{2}}{\pi} U_{DC} \frac{\cos(v\alpha/2)}{v\sqrt{R^2 + (v\omega L)^2}} \right]^2} = \frac{2\sqrt{2}}{\pi} \frac{U_{DC}}{R} \sqrt{\sum_{v=1}^{\infty} \left[\frac{\cos(v\alpha/2)}{v\sqrt{1^2 + (v\omega\tau)^2}} \right]^2}.$$

The total harmonic distortion of current will be similar to (1).

For $\alpha = 0$, $\beta = \pi$, $\tau = \frac{T}{2}$ root-mean-square value of total current waveform

$$I = \frac{2\sqrt{2}}{\pi} \frac{U_{DC}}{R} \sqrt{\sum_{v=1}^{\infty} \left[\frac{1}{v\sqrt{1^2 + (v\pi)^2}} \right]^2}.$$

The total harmonic distortion of current will be the same as it is in (2)

$$THD_I / \pi = \sqrt{1^2 + \pi^2} \sqrt{\sum_{v=1}^{\infty} \frac{1}{v^2(1 + v\pi)^2}} - 1 < 0,22, \text{ i. e. } 22\%.$$

THD for other types of the voltages with pulse-width modulation (PWM) (e.g. Fig. 3) can be solved by similar way, based on above mentioned method (cf.[1], [4]).

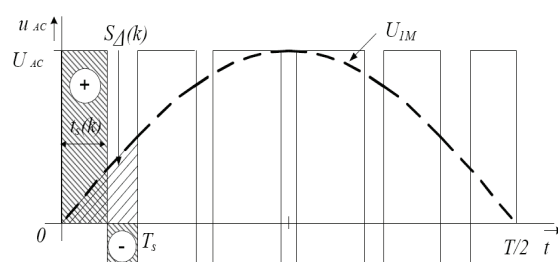


Fig. 3. Time waveforms of PWM voltage

The amplitudes of the voltage time waveform will be (cf. [8]):

$$A_v = \frac{4U_{DC}}{\nu \pi} \sum_{k=0}^{(m_f/4)-1} \left\{ \left[\cos\left(\nu k \frac{2\pi}{m_f}\right) - \cos\left(\nu k \frac{2\pi}{m_f} + \nu \omega t_s(k)\right) \right] - \right. \\ \left. - \left[\cos\left(\nu(k+1) \frac{2\pi}{m_f} - \nu \omega t_s(k)\right) - \cos\left(\nu k \frac{2\pi}{m_f}\right) \right] \right\}$$

where m_f is frequency modulation ratio of PWM (as a ratio of switching and fundamental frequency) and switching instant time t_s is calculated in [1] and [4]. Then, implicating approach introduced above, the total harmonic distortion factor will be similar to R - L - C resonant circuit with THD about 2 %.

6 Conclusions

New method of calculation based on the theory of series verified by actual examples has been presented. The solution given in the paper makes it possible to calculate the root-mean-square values and harmonic distortion factors more effectively and analyse more exactly effect of each harmonic component comprised in total waveform on resistive-inductive or capacitive load quantities more precisely.

Acknowledgement

The authors wish to thank for the financial support to Slovak grant agency VEGA through the projects No. 1/0470/09 and No.1/0867/08 and R&D operational program Centre of excellence of power electronics systems and materials for their components No. OPVaV-2008/2.1/01-SORO, ITMS 26220120003 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*. Logistika, 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) (in print).

Current address

Mariana Beňová, Ing., PhD.;

University of Zilina, Faculty of Electrical Engineering,
Dept. of Theoretical Electrotechnics,
Univerzitná 1, 010 26 Žilina, Slovak Republic
e-mail: benova@fel.uniza.sk

Branislav Dobrucký, prof., Ing., PhD.

University of Zilina, Faculty of Electrical Engineering,
Dept. of Mechatronics and Electronics,
Univerzitná 1, 010 26 Žilina, Slovak Republic
e-mail: dobrucky@fel.uniza.sk

Mariana Marčoková, doc., RNDr., CSc.

University of Zilina, Faculty of Science,
Dept. of Mathematics,
Univerzitná 1, 010 26 Žilina, Slovakia
e-mail: mariana.marcokova@fpv.uniza.sk

MODIFICATIONS TO IV SETUP OF STREAM CIPHER SNOW 2.0

BÖHM Patrik, (SK)

Abstract. A cipher is usually defined as an algorithm that transforms a plaintext into ciphertext under the control of a key. Due to synchronization purposes, another input value called initialization vector (IV) has gained importance in the design of stream ciphers in recent years. In our paper we list the security criteria for IV setup. In addition, we examine the IV setup of stream cipher SNOW 2.0, one of the two stream ciphers included in the ISO/IEC 18033-4 standard. We propose four modifications of the initialization algorithm that increase the speed of the cipher.

Key words. Stream cipher, SNOW 2.0, initialization value

Mathematics Subject Classification: Primary 94A60

1 SNOW 2.0

Cipher SNOW 2.0 [4] is synchronous additive stream cipher (see Figure 1). It allows two key sizes, 128 and 256 bits. Initialization vector has 128 bits. SNOW 2.0 consists of two components: linear feedback shift register (LFSR) of length 16 (denoted $S^0 \dots S^{15}$) over the field $F_{2^{32}}$ and finite state machine (FSM), that consists of two 32-bit registers R1 and R2.

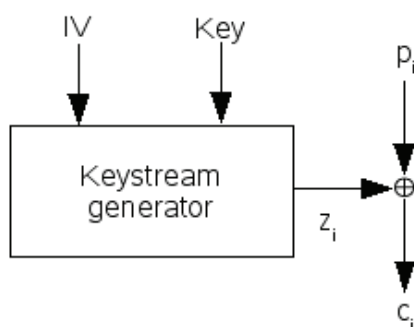


Figure 1. Synchronous additive stream cipher

1.1 Cipher initialization

At the start of the encryption algorithm, key and IV are loaded into the inner state. Initialization vector is represented by four 32-bit words $IV = (IV_3, IV_2, IV_1, IV_0)$. Similarly, key is represented by 4 32-bit words $K = (k_3, k_2, k_1, k_0)$ for 128-bit key or 8 32-bit words $K = (k_7, k_6, k_5, k_4, k_3, k_2, k_1, k_0)$ for 256-bit key.

Initialization of inner state for 128-bit key:

$$\begin{array}{llll} S^{15} = k_3 \oplus IV_0 & S^{14} = k_2 & S^{13} = k_1 & S^{12} = k_0 \oplus IV_1 \\ S^{11} = k_3 \oplus \bar{1} & S^{10} = k_2 \oplus \bar{1} \oplus IV_2 & S^9 = k_1 \oplus \bar{1} \oplus IV_3 & S^8 = k_0 \oplus \bar{1} \\ S^7 = k_3 & S^6 = k_2 & S^5 = k_1 & S^4 = k_0 \\ S^3 = k_3 \oplus \bar{1} & S^2 = k_2 \oplus \bar{1} & S^1 = k_1 \oplus \bar{1} & S^0 = k_0 \oplus \bar{1} \end{array}$$

Initialization of inner state for 256-bit key:

$$\begin{array}{llll} S^{15} = k_7 \oplus IV_0 & S^{14} = k_6 & S^{13} = k_5 & S^{12} = k_4 \oplus IV_1 \\ S^{11} = k_3 & S^{10} = k_2 \oplus IV_2 & S^9 = k_1 \oplus IV_3 & S^8 = k_0 \\ S^7 = k_7 \oplus \bar{1} & S^6 = k_6 \oplus \bar{1} & \dots & S^0 = k_0 \oplus \bar{1} \end{array}$$

In both cases registers $R1$ and $R2$ are assigned zero values. After values of registers $S^0 - S^{15}$, $R1$ and $R2$ are set, cipher is clocked 32 times with output of FSM taken as an input to LFSR, as shown on Figure 2. Detailed description of cipher SNOW 2.0 can be found in [4]. In this paper we consider 128-bit keys only.

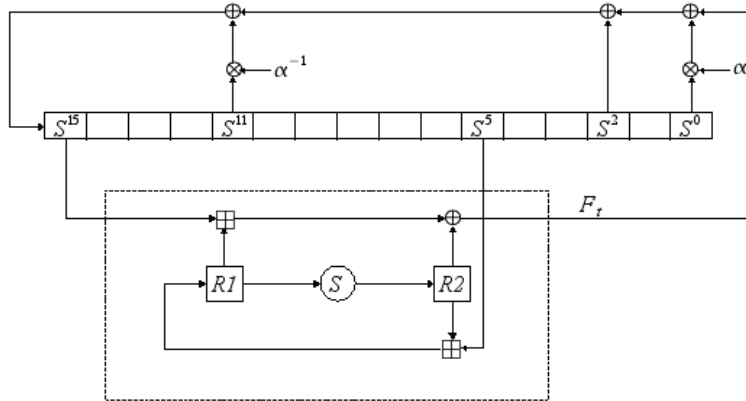


Figure 2. SNOW 2.0 initialization

2 Key/IV setup

Most newly proposed stream ciphers use initialization value for synchronization purposes. Initialization vector is publicly known and new value is used for each new message.

Designing the stream cipher with IV proved to be difficult task. Initialization vector is publicly known and is under the control of the attacker (with the restriction that each IV value can be used

only once). This increases the range of attacks available. For example, 10 stream ciphers out of 34 submitted to ECRYPT eStream project [5] were broken due to problems with the initialization of the cipher.

2.1 Key/IV setup security requirements

In order to consider the security of key/IV setup of stream cipher, we have to take the properties of the keystream generator (KG) into account. The stronger the KG is, the weaker the key/IV setup of cipher can be. For example, the key/IV setup of block cipher in CTR mode consists of just writing the values of key and IV into appropriate parts of cipher [3], but this mode of operation is considered secure. This relationship between the key/IV setup and the strength of KG makes the design and cryptanalysis of key/IV setup of stream cipher complicated.

If we assume that the KG part of the stream cipher is secure, a sufficient condition for stream cipher security is that the output of the cipher setup cannot be distinguished from random [8].

2.2 Key/IV setup criteria

There are three main criteria for evaluation of statistical properties of key/IV setup. We will state the definitions as given in [1] and [7].

For a vector x , the vector $x^{(i)}$ denotes the vector obtained by complementing the i -th bit of x . The hamming weight $w(x)$ of vector x is the number of nonzero components of x .

- The *dependence matrix* of a function $f : F_{2^n} \rightarrow F_{2^m}$ is an $n \times m$ matrix A whose (i, j) -th element denotes the number of inputs for which complementing the i -th input results in a change of j -th output bit:

$$a_{ij} = \#\{x \in X \mid (f(x^{(i)}))_j \neq (f(x))_j\} \quad (1)$$

- The *distance matrix* of a function $f : F_{2^n} \rightarrow F_{2^m}$ is an $n \times (m+1)$ matrix B whose (i, j) -th element denotes the number of inputs for which complementing the i -th input results in a change of j output bits:

$$b_{ij} = \#\{x \in X \mid w(f(x^{(i)}) - f(x)) = j\} \quad (2)$$

If the number of input bits is big, it is infeasible to compute these two matrices. Therefore we usually choose only subset of all input values and compute reduced matrices only. After we calculate two (reduced) matrices, we define three security criteria. Good statistical properties require all three degrees close to 1.

- The *degree of completeness* is defined as:

$$d_c = 1 - \frac{\#\{(i, j) \mid a_{ij} = 0\}}{nm} \quad (3)$$

- The *degree of avalanche effect* is defined as:

$$d_a = 1 - \frac{\sum_{i=1}^n \left| \frac{1}{\#X} \sum_{j=1}^m 2^j b_{ij} - m \right|}{nm} \quad (4)$$

- The *degree of strict avalanche criterion* is defined as:

$$d_{sa} = 1 - \frac{\sum_{i=1}^n \sum_{j=1}^m \left| \frac{2a_{ij}}{\#X} - 1 \right|}{nm} \quad (5)$$

3 Key/IV setup for SNOW 2.0

We tested the properties of stream initialization in [2]. The results for LFSR are in Table 1.

Table 1: Results for SNOW 2.0

Number of rounds	d_a	d_c	d_{sa}
10	0,237147	0,305283	0,993846
11	0,292854	0,364822	0,994333
12	0,351410	0,425247	0,994819
13	0,412912	0,487198	0,995306
14	0,475416	0,549698	0,995792
15	0,537910	0,612198	0,996279
16	0,599431	0,674210	0,996765
17	0,660683	0,736221	0,997252
18	0,716564	0,790649	0,997739
19	0,771741	0,836853	0,998225
20	0,825703	0,881805	0,998712
21	0,875597	0,925430	0,999199
22	0,923679	0,959335	0,999686
23	0,950645	0,977203	0,999712
24	0,969907	0,984344	0,999912
25	0,984049	0,990875	0,999912
26	0,989989	0,995392	0,999975
27	0,995795	0,997864	0,999974
28	0,999648	0,999939	0,999975
29	0,999858	1,000000	0,999975
30	0,999867	1,000000	0,999975
31	0,999867	1,000000	0,999975
32	0,999862	1,000000	0,999976

We see that the degree of avalanche effect is $d_{sa} = .999858$ after 29 rounds. If we increased the number of rounds, we didn't received significantly better results (we tested 48, 64, 128 and 192 rounds too). Also, after 29 rounds, degree of completeness is 1. Similarly, the degree of strict avalanche criterion reached the value $d_{sa} = .999975$ after 26 rounds. This value was not significantly improved for larger number of rounds.

So we conclude that the IV setup of cipher LFSR has good statistical properties if the following conditions are satisfied:

- i.) $d_a > .9998$ (6)
- ii.) $d_c = 1$ (7)
- iii.) $d_{sa} > .99997$ (8)

It is necessary to consider the statistical properties of registers $R1$ and $R2$ too. However, our calculations show [2] that if conditions (6) – (8) are satisfied for LFSR, they are satisfied for both registers $R1$ and $R2$ too. So in the next sections of this paper we will evaluate the statistical properties of LFSR only.

3.1 IV setup modifications

As a part of ECRYPT eSTREAM project, speed of stream ciphers was tested in [6], for example. AES in CTR mode and SNOW 2.0 were used as benchmark ciphers. The testing is done for three different stream lengths – 40 bytes, 576 bytes and 1500 bytes. In general, AES-CTR is faster then stream ciphers for short stream lengths (especially for 40 bytes) , while many stream ciphers are faster than AES-CTR for longer streams. It is caused by fast IV setup in AES-CTR cipher, but slower generation of keystream. Our speed comparison of ciphers AES-CTR and SNOW 2.0 is shown in Table 2.

Table 2: Performance testing (cycles per 32-bit word)

Cipher	40 bytes	576 bytes	1500 bytes
AES in CTR mode	100	73	72
SNOW 2.0	128	25	22

We tried to alter the IV setup of SNOW 2.0 so that it is faster than AES in CTR mode even for short 40 byte streams. It requires to decrease the number of rounds in IV setup. We tried several changes to the SNOW 2.0 cipher initialization that reduced the number of rounds while keeping the statistical properties (6) – (8) satisfied.

3.1.1 Modification 1

In the SNOW 2.0 initialization vector is XOR-ed to registers S^{15} , S^{12} , S^{10} and S^9 . We XOR-ed the IV to other registers and calculated the values of the three degrees d_a , d_c and d_{sa} . We tested all 1820 possibilities and in many cases we get better results then original version of SNOW 2.0. We received the best values for registers S^5 , S^6 , S^7 and S^8 , as shown in Table 3.

Table 3: Modification 1

Number of rounds	d_a	d_c	d_{sa}
10	0,388071	0,427765	0,994332
11	0,450573	0,490265	0,994817
12	0,513081	0,552765	0,995304
13	0,575575	0,615265	0,995790
14	0,638079	0,677765	0,996277
15	0,700580	0,740265	0,996764
16	0,763079	0,802765	0,997250
17	0,825581	0,865265	0,997737
18	0,886066	0,921646	0,998224
19	0,930878	0,962006	0,998711
20	0,964500	0,987595	0,999198
21	0,982568	0,997375	0,999684
22	0,989701	0,998276	0,999710
23	0,996971	0,999176	0,999910
24	0,998400	0,999588	0,999910
25	0,999851	1,000000	0,999973
26	0,999858	1,000000	0,999973
27	0,999857	1,000000	0,999974
28	0,999858	1,000000	0,999975
29	0,999864	1,000000	0,999975
30	0,999866	1,000000	0,999975
31	0,999861	1,000000	0,999975
32	0,999857	1,000000	0,999975

We see that all three conditions (6) – (8) are satisfied after 25 rounds, which is 4 rounds less then original version of SNOW 2.0. The initialization of the inner state will look like this:

$$\begin{array}{llll}
S^{15} = k_3 & S^{14} = k_2 & S^{13} = k_1 & S^{12} = k_0 \\
S^{11} = k_3 \oplus \bar{1} & S^{10} = k_2 \oplus \bar{1}_2 & S^9 = k_1 \oplus \bar{1} & S^8 = k_0 \oplus \bar{1} \oplus IV_3 \\
S^7 = k_3 \oplus IV_2 & S^6 = k_2 \oplus IV_1 & S^5 = k_1 \oplus IV_0 & S^4 = k_0 \\
S^3 = k_3 \oplus \bar{1} & S^2 = k_2 \oplus \bar{1} & S^1 = k_1 \oplus \bar{1} & S^0 = k_0 \oplus \bar{1}
\end{array}$$

3.1.2 Modification 2

In this modification we XOR-ed the parts of IV to each register $S^0 - S^{15}$. The initialization of the inner state will look like this:

$$\begin{array}{llll}
S^{15} = k_3 \oplus IV_3 & S^{14} = k_2 \oplus IV_2 & S^{13} = k_1 \oplus IV_1 & S^{12} = k_0 \oplus IV_0 \\
S^{11} = k_3 \oplus \bar{1} \oplus IV_3 & S^{10} = k_2 \oplus \bar{1}_2 \oplus IV_2 & S^9 = k_1 \oplus \bar{1} \oplus IV_1 & S^8 = k_0 \oplus \bar{1} \oplus IV_0 \\
S^7 = k_3 \oplus IV_3 & S^6 = k_2 \oplus IV_2 & S^5 = k_1 \oplus IV_1 & S^4 = k_0 \oplus IV_0 \\
S^3 = k_3 \oplus \bar{1} \oplus IV_3 & S^2 = k_2 \oplus \bar{1} \oplus IV_2 & S^1 = k_1 \oplus \bar{1} \oplus IV_1 & S^0 = k_0 \oplus \bar{1} \oplus IV_0
\end{array}$$

The results of statistical properties are in Table 4:

Table 4: Modification 2

Number of rounds	d_a	d_c	d_{sa}
10	0,491564	0,518234	0,995164
11	0,553085	0,580246	0,995650
12	0,614607	0,642258	0,996137
13	0,676136	0,704269	0,996624
14	0,737650	0,766281	0,997111
15	0,799171	0,828293	0,997597
16	0,860354	0,890305	0,998084
17	0,912471	0,947708	0,998570
18	0,952453	0,976135	0,999027
19	0,977315	0,990189	0,999468
20	0,988467	0,995819	0,999897
21	0,995617	0,998306	0,999903
22	0,997461	0,999039	0,999954
23	0,999248	0,999664	0,999954
24	0,999633	0,999893	0,999975
25	0,999864	1,000000	0,999975
26	0,999862	1,000000	0,999975
27	0,999865	1,000000	0,999976
28	0,999863	1,000000	0,999976
29	0,999872	1,000000	0,999976
30	0,999865	1,000000	0,999976
31	0,999855	1,000000	0,999976
32	0,999844	1,000000	0,999977

We see that there is no improvement with respect to *Modification 1*. All three conditions (6) – (8) are satisfied after 25 rounds too.

3.1.3 Modification 3

In addition to results received in *Modification 1* (values of IV are XOR-ed to registers S^5 , S^6 , S^7 and S^8), we decided to XOR the output of FSM to one of the registers S^0 - S^{15} . We tested all 16 possibilities and it showed up that the best results were reached when output of FSM was XOR-ed to the register S^2 . So the cipher initialization is now defined as in *Modification 1*, with a change in update function of LFSR:

$$S^2(t+1) = S^3(t) \oplus FSM(t) \quad (9)$$

Table 5: Modification 3

Number of rounds	d_a	d_c	d_{sa}
10	0,602852	0,619003	0,996040
11	0,665326	0,681503	0,996527
12	0,727797	0,744003	0,997017

13	0,790290	0,806503	0,997504
14	0,790286	0,806503	0,997504
15	0,790287	0,806503	0,997504
16	0,835096	0,847122	0,997991
17	0,926536	0,934250	0,998964
18	0,999485	1,000000	0,999937
19	0,999763	1,000000	0,999963
20	0,999873	1,000000	0,999977
21	0,999873	1,000000	0,999977
22	0,999873	1,000000	0,999976
23	0,999876	1,000000	0,999976
24	0,999877	1,000000	0,999976
25	0,999868	1,000000	0,999976
26	0,999865	1,000000	0,999977
27	0,999865	1,000000	0,999976
28	0,999856	1,000000	0,999976
29	0,999867	1,000000	0,999976
30	0,999874	1,000000	0,999977
31	0,999868	1,000000	0,999977
32	0,999869	1,000000	0,999977

All conditions (6) – (8) are satisfied after 20 rounds.

3.1.4 Modification 4

This modification is similar to *Modification 3*, but values of IV are XOR-ed to all registers $S^0 - S^{15}$, not just registers S^5, S^6, S^7 and S^8 . Results are in Table 6.

Table 6: Modification 4

Number of rounds	d_a	d_c	d_{sa}
10	0,679549	0,706985	0,996687
11	0,741057	0,768997	0,997173
12	0,802578	0,831009	0,997661
13	0,864105	0,893021	0,998148
14	0,864110	0,893021	0,998148
15	0,864117	0,893021	0,998147
16	0,891187	0,906998	0,998588
17	0,958113	0,971130	0,999507
18	0,999839	1,000000	0,999973
19	0,999855	1,000000	0,999975
20	0,999860	1,000000	0,999975
21	0,999862	1,000000	0,999976
22	0,999870	1,000000	0,999976
23	0,999871	1,000000	0,999976
24	0,999862	1,000000	0,999976
25	0,999855	1,000000	0,999976
26	0,999862	1,000000	0,999976
27	0,999873	1,000000	0,999976
28	0,999869	1,000000	0,999977

29	0,999871	1,000000	0,999977
30	0,999871	1,000000	0,999977
31	0,999875	1,000000	0,999976
32	0,999876	1,000000	0,999976

All statistical criteria (6) – (8) are satisfied after 18 rounds, which is the best result of all four modifications.

3.2 Performance testing

We measured the performance of modified versions of the cipher SNOW 2.0 for 40 byte streams. The performance results are shown in Table 7 (in cycles per 32-bit word).

Table 7: Performance in cycles/word

AES-CTR	SNOW 2.0	Modification 1	Modification 2	Modification 3	Modification 4
100	122	102	104	107	101

All four modifications give very similar results. There is a significant speed increase in respect to SNOW 2.0 cipher. Moreover, the speed of modified versions of SNOW 2.0 compares well to AES in CTR mode, even though we didn't get better speeds than AES-CTR.

4 Conclusions

In our paper we proposed four modifications of the key/IV setup of stream cipher SNOW 2.0. All modified versions of the cipher reach high encryption speeds 101 – 107 cycles per word even for small 40-byte keystream lengths. Since all speeds are almost the same, we suggest using *Modification 1* with encryption speed 102 cycles per 32-bit word, since this modification requires only small changes to the cipher SNOW 2.0:

- XOR initialization vector to registers S^5 , S^6 , S^7 and S^8 instead of registers S^{15} , S^{12} , S^{10} and S^9 .
- Clock the cipher 25 times, not 32 times.

With this modification we reach the same encryption speed as AES-CTR cipher even for small 40-byte keystream lengths.

Acknowledgement

The paper was supported by VEGA grant no. 2/7138/27.

References

- [1.] *Analysis of the Key Setup Function in Rabbit*. http://www.cryptico.com/Files/filer/wp_key_setup_analysis.pdf

- [2.] BÖHM, P.: *Analysis of IV setup of stream cipher SNOW 2.0*. In Annals of the Constantin Brancusi University of Targu - Jiu, Engineering Series, No. 3. Academica Brancusi Publisher, 2009.
- [3.] DWORKIN, M.: *Recommendation for Block Cipher Modes of Operation*. NIST special Publication 800-38A 2001 Edition. <http://csrc.nist.gov/publications/nistpubs/800-38a/sp800-38a.pdf>
- [4.] EKDAHL, P. – JOHANNSON, T.: *A new version of the stream cipher SNOW*. In K. Nyberg and H. M. Heys editors, *Selected Areas in Cryptography – SAC 2002*, Lecture Notes in computer Science. Springer-Verlag, 2002.
- [5.] eStream – the ECRYPT stream cipher project. <http://www.ecrypt.eu.org/stream>
- [6.] eStream testing framework. <http://www.ecrypt.eu.org/stream/perf/#results>
- [7.] SERF, P.: *The degrees of completeness, of avalanche effect, and of strict avalanche criterion for Mars, RC5, Rijndael, Serpent, and Twofish with reduced number of rounds*. Nessie Technical report, 2000.
- [8.] ZENNER, E.: *Why IV Setup for Stream Ciphers is Difficult*. In Proc. Dagstuhl Seminar “Symmetric Cryptography”, 2007.

Current address

Patrik Böhm, Mgr.

FPEDAS, KKMAHI, ŽU v Žiline,
Univerzitná 1, 010 26 Žilina, 041/5133268,
e-mail: patrik.bohm@fpedas.uniza.sk

CONSTRUCTION OF INTERVAL CUBIC SPLINE-WAVELET BASES ON COARSE SCALES

ČERNÁ Dana, (CZ), FINĚK Václav, (CZ)

Abstract. Constructions of wavelet bases on the unit interval assume that the supports of the left and right boundary functions do not overlap. It enables to construct wavelets only on scales larger than j_0 , where the limit scale j_0 depends on the type of a wavelet. In our paper, we propose a construction of the cubic spline wavelet bases on the interval on coarser scales. We show that the decomposition on coarser scales improves the condition number of the wavelet bases. Finally, we compare the efficiency of an adaptive wavelet scheme for several spline-wavelet bases and we show the superiority of our construction.

Key words. Spline wavelets, cubic, interval, coarse scale, the condition number, boundary conditions.

Mathematics Subject Classification. Primary 65T60, 65D07; Secondary 65N12.

1 Introduction

Wavelet bases on a bounded domain are usually constructed in the following way: Wavelets on the real line are adapted to the interval and then extended by tensor product technique to the n - dimensional cube. Finally by splitting the domain into overlapping or non-overlapping subdomains which are images of a unit cube under appropriate parametric mappings one can obtain a wavelet basis or a wavelet frame on a fairly general domain. Thus, the properties of the employed wavelet basis on the interval are crucial for the properties of the resulting bases or frames on a general domain.

Constructions of wavelet bases on the unit interval assume that the supports of the left and right boundary functions do not overlap. It enables to construct wavelets only on scales larger than j_0 , where the limit scale j_0 depends on the type of a wavelet. It means that the minimum number of basis functions in a subdomain is $2^{j_0 d}$, where d is the spatial dimension. It might be too restrictive, especially in higher dimensions. In [1] the construction of orthonormal wavelet

bases on the interval at large scales has been proposed. However, orthonormal wavelet bases are usually avoided in numerical treatment of partial differential equations, because they are not accessible analytically, the complementary boundary conditions can not be satisfied and it is not possible to increase the number of vanishing wavelet moments independent from the order of accuracy. For these reasons spline basis functions would be convenient.

In our paper, we focus on the cubic spline wavelets and we propose a construction of the cubic spline wavelet bases on the interval on coarser scales. We show that the decomposition on coarser scales improves the condition number of the wavelet bases. Finally, we compare the efficiency of an adaptive wavelet scheme for several spline-wavelet bases and we show the superiority of our construction.

2 Wavelet bases

This section provides a short introduction to the concept of wavelet bases in Sobolev spaces. We consider the domain $\Omega \subset \mathbb{R}^d$ and the Sobolev space or its subspace $H \subset H^s(\Omega)$ for nonnegative integer s with an inner product $\langle \cdot, \cdot \rangle_H$, a norm $\|\cdot\|_H$ and a seminorm $|\cdot|_H$. In case $s = 0$, we consider the space $L^2(\Omega)$ and we denote the L^2 -inner product by $\langle \cdot, \cdot \rangle$ and the L^2 -norm by $\|\cdot\|$, respectively. Let \mathcal{J} be some index set and let each index $\lambda \in \mathcal{J}$ take the form $\lambda = (j, k)$, where $|\lambda| := j \in \mathbb{Z}$ is a *scale* or a *level*. Let

$$l^2(\mathcal{J}) := \left\{ \mathbf{v} : \mathcal{J} \rightarrow \mathbb{R}, \sum_{\lambda \in \mathcal{J}} |\mathbf{v}_\lambda|^2 < \infty \right\}. \quad (1)$$

A family $\Psi := \{\psi_\lambda \in \mathcal{J}\} \subset H$ is called a *wavelet basis* of H , if

- i) Ψ is a *Riesz basis* for H , i.e. Ψ is complete in H and there exist constants $c, C \in (0, \infty)$ such that

$$c \|\mathbf{b}\|_{l^2(\mathcal{J})} \leq \left\| \sum_{\lambda \in \mathcal{J}} b_\lambda \psi_\lambda \right\|_H \leq C \|\mathbf{b}\|_{l^2(\mathcal{J})}, \quad \mathbf{b} := \{b_\lambda\}_{\lambda \in \mathcal{J}} \in l^2(\mathcal{J}). \quad (2)$$

Constants $c_\psi := \sup \{c : c \text{ satisfies (2)}\}$, $C_\psi := \inf \{C : C \text{ satisfies (2)}\}$ are called *Riesz bounds* and $\text{cond } \Psi := C_\psi / c_\psi$ is called the *condition* of Ψ .

- ii) The functions are *local* in the sense that $\text{diam}(\Omega_\lambda) \leq C 2^{-|\lambda|}$ for all $\lambda \in \mathcal{J}$, where Ω_λ is the support of ψ_λ , and at a given level j the supports of only finitely many wavelets overlap in any point $x \in \Omega$.

By the Riesz representation theorem, there exists a unique family

$$\tilde{\Psi} = \{\tilde{\psi}_\lambda, \lambda \in \mathcal{J}\} \subset H \quad (3)$$

biorthogonal to Ψ , i.e.

$$\langle \psi_{i,k}, \tilde{\psi}_{j,l} \rangle_H = \delta_{i,j} \delta_{k,l}, \quad \text{for all } (i,k), (j,l) \in \mathcal{J}. \quad (4)$$

This family is also a Riesz basis for H . The basis Ψ is called a *primal* wavelet basis, $\tilde{\Psi}$ is called a *dual* wavelet basis.

In many cases, the wavelet system Ψ is constructed with the aid of a multiresolution analysis. A sequence $S = \{S_j\}_{j \geq j_0}$, of closed linear subspaces $S_j \subset H$ is called a *multiresolution* or *multiscale analysis*, if

$$S_{j_0} \subset S_{j_0+1} \subset \dots \subset S_j \subset S_{j+1} \subset \dots \subset H \quad (5)$$

and $\cup_{j \geq j_0} S_j$ is complete in H . The dual wavelet system $\tilde{\Psi}$ generates a *dual* multiresolution analysis $\tilde{S} = \{\tilde{S}_j\}_{j \geq j_0}$.

The nestedness and the closedness of the multiresolution analysis implies the existence of the *complement spaces* W_j such that

$$S_{j+1} = S_j \oplus W_j, \quad W_j \perp \tilde{S}_j. \quad (6)$$

We now assume that S_j and W_j are spanned by sets of basis functions

$$\Phi_j := \{\phi_{j,k}, k \in \mathcal{I}_j\}, \quad \Psi_j := \{\psi_{j,k}, k \in \mathcal{J}_j\}, \quad (7)$$

where $\mathcal{I}_j, \mathcal{J}_j$ are finite or at most countable index sets. We refer to $\phi_{j,k}$ as *scaling functions* and $\psi_{j,k}$ as *wavelets*. The multiscale basis is given by

$$\Psi_{j_0,s} = \Phi_{j_0} \cup \bigcup_{j=j_0}^{j_0+s-1} \Psi_j \quad (8)$$

and the wavelet basis of H is obtained by

$$\Psi = \Phi_{j_0} \cup \bigcup_{j \geq j_0} \Psi_j \quad (9)$$

Dual scaling basis $\tilde{\Phi}_{j_0}$ and dual wavelet basis $\tilde{\Psi}_j$ are defined in a similar way.

Polynomial exactness of order $N \in \mathbb{N}$ for the primal scaling basis and of order $\tilde{N} \in \mathbb{N}$ for the dual scaling basis is another desired property of wavelet bases. It means that

$$\mathbb{P}_{N-1}(\Omega) \subset S_j, \quad \mathbb{P}_{\tilde{N}-1}(\Omega) \subset \tilde{S}_j, \quad j \geq j_0, \quad (10)$$

where $\mathbb{P}_m(\Omega)$ is the space of all algebraic polynomials on Ω of degree less or equal to m .

By Taylor theorem, the polynomial exactness of order \tilde{N} on the dual side is equivalent to \tilde{N} vanishing wavelet moments on the primal side, i.e.

$$\int_{\Omega} P(x) \psi_{\lambda}(x) dx = 0, \quad P \in \mathbb{P}_{\tilde{N}-1}, \quad \psi_{\lambda} \in \bigcup_{j \geq j_0} \Psi_j. \quad (11)$$

It is known [9] that if $\Psi \subset H^s(\Omega)$, $s > 0$, is a wavelet basis for $L^2(\Omega)$ with a polynomial exactness at least s , then

$$\mathbf{D}^{-s} \Psi := \{d_{\lambda}^{-s} \psi_{\lambda}, \lambda \in \mathcal{J}\} \quad (12)$$

is a wavelet basis for $H^s(\Omega)$, where $\mathbf{D} := \text{diag} \{d_{\lambda} = 2^{|\lambda|}, \lambda \in \mathcal{J}\}$.

3 Optimized construction of cubic spline-wavelet bases on the interval

In this section, we briefly review the construction of stable cubic spline-wavelet basis on the interval from [6]. The primal scaling bases will be the same as bases designed in [2], because they are known to be well-conditioned. Let $\mathbf{t}^j = (t_k^j)_{k=-3}^{2^j+3}$ be a sequence of knots defined by

$$\begin{aligned} t_k^j &= 0 \quad \text{for } k = -3, \dots, 0, \\ t_k^j &= \frac{k}{2^j} \quad \text{for } k = 1, \dots, 2^j - 1, \\ t_k^j &= 1 \quad \text{for } k = 2^j, \dots, 2^j + 3. \end{aligned}$$

The corresponding cubic B-splines are defined by

$$B_k^j(x) := (t_{k+4}^j - t_k^j) [t_k^j, \dots, t_{k+4}^j]_t (t - x)_+^3, \quad x \in [0, 1], \quad (13)$$

where $(x)_+ := \max\{0, x\}$ and $[t_1, \dots, t_4]_t f$ is the fourth divided difference of f . The set Φ_j of primal scaling functions is then simply defined as

$$\phi_{j,k} = 2^{j/2} B_k^j, \quad \text{for } k = -3, \dots, 2^j - 1, \quad j \geq 0. \quad (14)$$

Thus there are $2^j - 3$ inner scaling functions and 3 functions on each boundary. The inner functions are translations and dilations of a function ϕ which corresponds to the primal scaling function constructed by Cohen, Daubechies, and Feauveau in [7]. In the following, we consider ϕ from [7] which is shifted so that its support is $[0, 4]$.

The desired property of the dual scaling basis $\tilde{\Phi}$ is biorthogonality to Φ and polynomial exactness of order $\tilde{N} \geq 4$, \tilde{N} even. Let $\tilde{\phi}$ be dual scaling function which was designed in [7] and which is shifted so that its support is $[-\tilde{N} + 1, \tilde{N} + 3]$. We assume that $j \geq 4$ so that the supports of the left and right boundary functions do not overlap. We define inner scaling functions as translations and dilations of $\tilde{\phi}$:

$$\theta_{j,k} = 2^{j/2} \tilde{\phi}(2^j \cdot -k), \quad k = \tilde{N} - 1, \dots, 2^j - \tilde{N} - 3. \quad (15)$$

There will be two types of basis functions at each boundary. Basis functions of the first type are defined to preserve polynomial exactness in the same way as in [10]:

$$\theta_{j,k} = 2^{j/2} \sum_{l=-2-\tilde{N}}^{\tilde{N}-2} \left\langle p_{k+3}^{\tilde{N}-1}, \phi(\cdot - l) \right\rangle \tilde{\phi}(2^j \cdot -l) |_{[0,1]}, \quad k = -3, \dots, \tilde{N} - 4. \quad (16)$$

where $\{p_0, \dots, p_{\tilde{N}-1}\}$ is a monomial basis of $\mathbb{P}_{\tilde{N}-1}([0, 1])$, i.e. $p_i(x) = x^i$, $x \in [0, 1]$, $i = 0, \dots, \tilde{N} - 1$. The basis functions of the second type are defined as

$$\theta_{j,k} = 2^{\frac{j+1}{2}} \sum_{l=\tilde{N}-1-2k}^{\tilde{N}+3} \tilde{h}_l \tilde{\phi}(2^{j+1} \cdot -2k - l) |_{[0,1]}, \quad k = \tilde{N} - 3, \dots, \tilde{N} - 2, \quad (17)$$

where \tilde{h}_l are scaling coefficients corresponding to $\tilde{\phi}$.

The boundary functions at the right boundary are defined to be symmetrical with the left boundary functions:

$$\theta_{j,k} = \theta_{j,2^j-3-k} (1 - \cdot), \quad k = 2^j - 2 - \tilde{N}, \dots, 2^j - 1. \quad (18)$$

Since the set $\Theta_j := \{\theta_{j,k} : k = -3, \dots, 2^j - 1\}$ is not biorthogonal to Φ_j , we derive a new set $\tilde{\Phi}_j$ from Θ_j by biorthogonalization. Let $\mathbf{A}_j = (\langle \phi_{j,k}, \theta_{j,l} \rangle)_{j,l=-N+1}^{2^j-1}$, then viewing $\tilde{\Phi}_j$ and Θ_j as column vectors we define

$$\tilde{\Phi}_j := \mathbf{A}_j^{-T} \Theta_j. \quad (19)$$

Our next goal is to determine the corresponding wavelets. We follow a general principle called stable completion which was proposed in [3]. We found the initial stable completion by the method from [10] with some modifications. Since the construction of the initial stable completion is long and technical we can not provide it here. A detailed description can be found in [4].

3.1 Adaptation to complementary boundary conditions

We review the adaptation of the constructed bases to complementary boundary conditions of the first order from [5]. This means that the primal wavelet basis is adapted to homogeneous Dirichlet boundary conditions of the first order, whereas the dual wavelet basis preserves the full degree of polynomial exactness.

Let $\Phi_j = \{\phi_{j,k}, k = -3, \dots, 2^j - 1\}$ be defined as above. Note that the functions $\phi_{j,-3}$ and $\phi_{j,2^j-1}$ are the only two functions which do not vanish at zero. Therefore, defining

$$\Phi_j^{comp} = \{\phi_{j,k}, k = -2, \dots, 2^j - 2\} \quad (20)$$

we obtain primal scaling bases satisfying complementary boundary conditions of the first order.

On the dual side, we also need to omit one scaling function at each boundary, because the number of primal scaling functions must be the same as the number of dual scaling functions. Since we want to preserve the full degree of polynomial exactness, we omit one function of the second type at each boundary. Thus, we define

$$\begin{aligned} \theta_{j,k}^{comp} &= \theta_{j,k-1}, & k &= -2, \dots, \tilde{N} - 3, \\ \theta_{j,k}^{comp} &= \theta_{j,k}, & k &= \tilde{N} - 2, \dots, 2^j - \tilde{N} - 2, \\ \theta_{j,k}^{comp} &= \theta_{j,k+1}, & k &= 2^j - \tilde{N} - 1, \dots, 2^j - 2. \end{aligned}$$

Since the set $\Theta_j^{comp} := \{\theta_{j,k}^{comp} : k = -2, \dots, 2^j - 2\}$ is not biorthogonal to Φ_j , we derive a new set $\tilde{\Phi}_j^{comp}$ from Θ_j^{comp} by biorthogonalization. Finally, we again determine the corresponding wavelets $\Psi_j^{comp} := \{\psi_{j,k}^{comp}, k = 1, \dots, 2^j\}$ by the method of stable completion.

4 Decomposition on coarser scales

In this section, we decompose the scaling basis Φ_4 into two parts Φ_3 and Ψ_3 . We consider cubic spline wavelet basis with six vanishing wavelet moments, i.e. $\tilde{N} = 6$. Six vanishing wavelet moments on the primal side is equivalent to the polynomial exactness of order six on the dual side. We choose polynomial exactness of this order, because the dual scaling function of order four does not belong to $L^2(\mathbb{R})$ and the polynomial exactness of order greater than six leads to a larger support of primal wavelets which makes the computation more expensive.

Scaling functions in Φ_3 are defined by (13) for $j = 3$. Functions in Ψ_3 are defined by

$$\psi_{3,k}(x) := \frac{(B_{t_k}^{10})^{(6)}(x)}{\|(B_{t_k}^{10})^{(6)}\|}, \quad k = 1, \dots, 8, \quad x \in [0, 1], \quad (21)$$

where $B_{t_k}^{10}$ is a B-spline of order ten on the sequence of knots t_k and $^{(6)}$ denotes the sixth derivative. The sequences of knots are given by:

$$\begin{aligned} t_1 &= [0, 0, 0, 0, 1/16, 1/8, 2/8, 3/8, 4/8, 5/8, 6/8], \\ t_2 &= [0, 0, 0, 1/8, 3/16, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8], \\ t_3 &= [0, 0, 0, 1/8, 2/8, 5/16, 3/8, 4/8, 5/8, 6/8, 7/8], \\ t_4 &= [0, 0, 1/8, 2/8, 3/8, 7/16, 4/8, 5/8, 6/8, 7/8, 1], \\ t_5 &= [0, 1/8, 2/8, 3/8, 4/8, 9/16, 5/8, 6/8, 7/8, 1, 1], \\ t_6 &= [1/8, 2/8, 3/8, 4/8, 5/8, 11/16, 6/8, 7/8, 1, 1, 1], \\ t_7 &= [1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 13/16, 7/8, 1, 1, 1], \\ t_8 &= [2/8, 3/8, 4/8, 5/8, 6/8, 7/8, 15/16, 1, 1, 1, 1]. \end{aligned} \quad (22)$$

Functions in Ψ_3 have the same number of vanishing wavelet moments as wavelets in Ψ .

Theorem 4.1 *Under the above assumptions, $\text{span } \Phi_4 = \text{span } \Phi_3 \cup \Psi_3$. Furthermore, functions $\psi_{3,k}$, $k = 1, \dots, 8$, have six vanishing wavelet moments.*

Proof. It is known that Φ_4 is a basis of the space of all cubic splines on the knots $\mathbf{t}^4 = [0, 0, 0, 0, 1/16, 2/16, \dots, 15/16, 1, 1, 1, 1]$. Functions in Φ_3 are cubic splines on the subsets of these knots. Functions in Ψ_3 are also cubic splines, because they are sixth derivative of the spline of order ten, and they are defined on the subsets of knots \mathbf{t}^4 . Therefore $\Phi_3 \cup \Psi_3 \subset \text{span } \Phi_4$.

Functions in Φ_3 are linearly independent. Functions $\psi_{3,i}$ cannot be written as linear combinations of functions from $\Phi_3 \cup \Psi_3 \setminus \{\psi_{3,i}\}$, because it is a cubic spline on sequence of the knots t_i containing an additional knot $i/16$. Hence, $\Psi_3 \cup \Phi_3$ is a linearly independent subset of $\text{span } \Phi_4$, which proves the first assertion.

To prove that the functions $\psi_{3,k}$, $k = 1, \dots, 8$, have six vanishing moments, we use the integration by parts:

$$\int_0^1 x^n (B_{t_k}^{10})^{(6)}(x) dx = \left[x^n (B_{t_k}^{10})^{(5)}(x) \right]_0^1 - \int_0^1 n x^{n-1} (B_{t_k}^{10})^{(5)}(x) dx, \quad n = 0, \dots, 5. \quad (23)$$

Since $(B_{t_k}^{10})^{(n)}$ is the spline of order $10 - n$ on the knots of multiplicity at most four in points 0 and 1, we have

$$(B_{t_k}^{10})^{(n)}(0) = (B_{t_k}^{10})^{(n)}(1) = 0, \quad n = 0, \dots, 6, \quad (24)$$

and thus

$$\int_0^1 (B_{t_k}^{10})^{(6)}(x) dx = 0 \quad (25)$$

and

$$\int_0^1 x^n (B_{t_k}^{10})^{(6)}(x) dx = - \int_0^1 n x^{n-1} (B_{t_k}^{10})^{(5)}(x) dx, \quad n = 1, \dots, 5. \quad (26)$$

Using (24) and the integration by parts five times, we obtain for $n = 1, \dots, 5$:

$$\int_0^1 x^n (B_{t_k}^{10})^{(6)}(x) dx = (-1)^n n! \left[(B_{t_k}^{10})^{(6-n)}(1) - (B_{t_k}^{10})^{(6-n)}(0) \right] = 0, \quad (27)$$

which proves the assertion.

4.1 Decomposition of bases with complementary boundary conditions

Scaling functions in Φ_3^{comp} are defined by (20) for $j = 3$. Functions in Ψ_3^{comp} are defined by

$$\psi_{3,k}^{comp}(x) := \frac{(B_{t_k}^{10})^{(6)}(x)}{\|(B_{t_k}^{10})^{(6)}\|}, \quad k = 1, \dots, 8, \quad x \in [0, 1], \quad (28)$$

where the sequences of knots are given by:

$$\begin{aligned} t_1 &= [0, 0, 0, 1/16, 1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8], \\ t_2 &= [0, 0, 0, 1/8, 3/16, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8], \\ t_3 &= [0, 0, 0, 1/8, 2/8, 5/16, 3/8, 4/8, 5/8, 6/8, 7/8], \\ t_4 &= [0, 0, 1/8, 2/8, 3/8, 7/16, 4/8, 5/8, 6/8, 7/8, 1], \\ t_5 &= [0, 1/8, 2/8, 3/8, 4/8, 9/16, 5/8, 6/8, 7/8, 1, 1], \\ t_6 &= [1/8, 2/8, 3/8, 4/8, 5/8, 11/16, 6/8, 7/8, 1, 1, 1], \\ t_7 &= [1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 13/16, 7/8, 1, 1, 1], \\ t_8 &= [1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8, 15/16, 1, 1, 1]. \end{aligned} \quad (29)$$

Theorem 4.2 *Under the above assumptions, $\text{span } \Phi_4^{comp} = \text{span } \Phi_3^{comp} \cup \Psi_3^{comp}$. Furthermore, functions $\psi_{3,k}^{comp}$, $k = 1, \dots, 8$, have six vanishing wavelet moments.*

Proof. The proof is similar to the proof of Theorem 4.1.

Scaling functions on the coarse scale are shown in Figure 1.

Left boundary wavelets are shown in Figure 2. The right boundary wavelets are symmetrical to the left boundary wavelets.

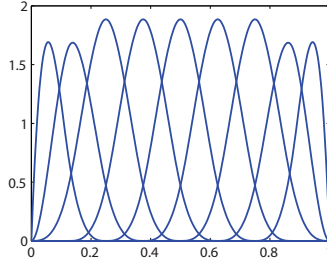


Figure 1: Scaling functions $\phi_{3,k}^{comp}$, $k = -2, \dots, 6$.

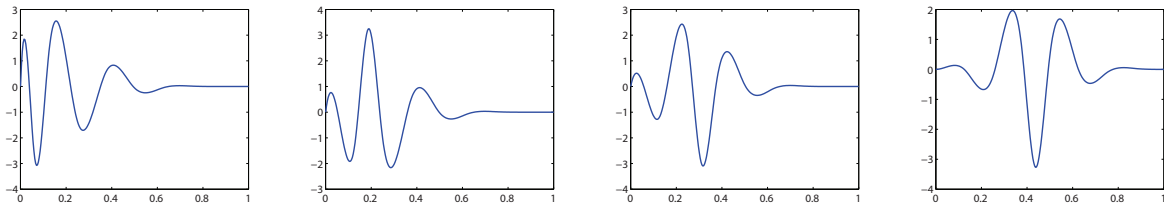


Figure 2: Left boundary wavelet functions $\psi_{3,k}^{comp}$, $k = 1, \dots, 4$.

5 Quantitative properties of constructed bases

As we already mentioned, the decomposition of wavelet bases on coarse scales is important especially in higher dimensions, because it enables to reduce the number of basis functions on the minimum level. In this section, we show that the additional advantage of wavelet bases on coarse scales adapted to complementary boundary conditions consists in their better condition with respect to H^1 -seminorm in comparison with bases from [5, 12]. In [5], we orthogonalize the scaling functions on the coarsest scale in H^1 -seminorm. This approach leads to improved condition number of the resulting bases and stiffness matrices, but the computation with these bases is more expensive, because the multiplication by the orthogonalization matrix must be accomplished in each iteration. The condition of multi-scale wavelet basis $\Psi_{j_0,6}^{comp}$ is shown in Table 1. Wavelet basis from [5] is denoted by CF, the basis constructed in this paper is denoted by CFcoarse and a basis with orthogonalization of scaling functions is denoted by CFort.

Other criteria for the effectiveness of wavelet bases is the condition number of a corresponding stiffness matrix. Here, let us consider the stiffness matrix for the Poisson equation:

$$\mathbf{A}_{j_0,s} = (\langle \psi'_{j,k}, \psi'_{l,m} \rangle)_{\psi_{j,k}, \psi_{l,m} \in \Psi_{j_0,s}^{comp}}. \quad (30)$$

It is well-known that the condition number of $\mathbf{A}_{j_0,s}$ increases quadratically with the matrix size. To remedy this, we use a diagonal matrix for preconditioning

$$\mathbf{A}_{j_0,s}^{prec} = \mathbf{D}_{j_0,s}^{-1} \mathbf{A}_{j_0,s} \mathbf{D}_{j_0,s}^{-1}, \quad \mathbf{D}_{j_0,s} = \text{diag} \left(\langle \psi'_{j,k}, \psi'_{j,k} \rangle^{1/2} \right)_{\psi_{j,k} \in \Psi_{j_0,s}^{comp}}. \quad (31)$$

j	$\Psi_{4,j}$				$\mathbf{A}_{4,j}^{prec}$			
	CF	CFcoarse	CFort	Primbs	CF	CFcoarse	CFort	Primbs
1	7.85	9.52	7.98	22.56	50.57	15.23	15.23	51.95
2	9.38	10.69	8.19	40.06	51.41	15.78	15.78	113.80
3	10.87	11.73	9.42	54.55	51.72	15.95	15.94	129.24
4	12.06	12.66	10.91	65.24	51.83	16.15	16.15	134.88
5	13.00	13.45	12.12	73.58	51.91	16.24	16.24	137.08
6	13.77	14.12	13.10	80.18	51.93	16.30	16.30	137.99
7	14.40	14.68	13.90	85.47	51.94	16.31	16.31	138.39

Table 1: The condition of multiscale wavelet bases and condition numbers of stiffness matrices for various constructions of wavelet bases.

It is known that the condition number of the stiffness matrix $\mathbf{A}_{j_0,s}^{prec}$ equals to the square of the condition number of the basis $\Psi_{j_0,s}^{comp}$ with respect to the H^1 -seminorm. Condition numbers of the resulting matrices are also listed in Table 1.

6 Numerical example

Now, we compare the quantitative behaviour of the adaptive wavelet method with bases constructed in this paper and bases from [5, 12]. We consider the Poisson equation

$$-u'' = f, \quad \text{in } (0, 1), \quad u(0) = u(1) = 0, \quad (32)$$

where the solution u is given by

$$u(x) := x(1-x) + \frac{(e^{50x} - 1)(e^{50} - e^{50x})}{(e^{50} - 1)^2}. \quad (33)$$

Note that the solution exhibits a large derivative near the point 1. We solve the problem by the adaptive wavelet method similar to the method proposed in [8]. We use wavelets up to the scale $|\lambda| \leq 11$. The convergence history for several wavelet bases is shown in Figure 3. In our experiments, the convergence rate is better for basis decomposed on coarser scales and basis with orthogonalization [5] than for wavelet bases from [5, 12]. Furthermore, there is also significant difference in the number of iterations needed to resolve the problem with desired accuracy. The adaptive method with CFort basis needs almost similar number of iterations as the method with CFcoarse basis, however, the price of one iteration for CFcoarse basis is higher. As can be seen in Figure 3 the adaptive method with CFcoarse basis requires the least degrees of freedom.

Acknowledgement

The research of the first author has been supported by the Ministry of Education, Youth and Sports of the Czech Republic through the Research Centers LC06024. The second author has been supported by the grant GP201/09/P641 of the Grant Agency of the Czech Republic.

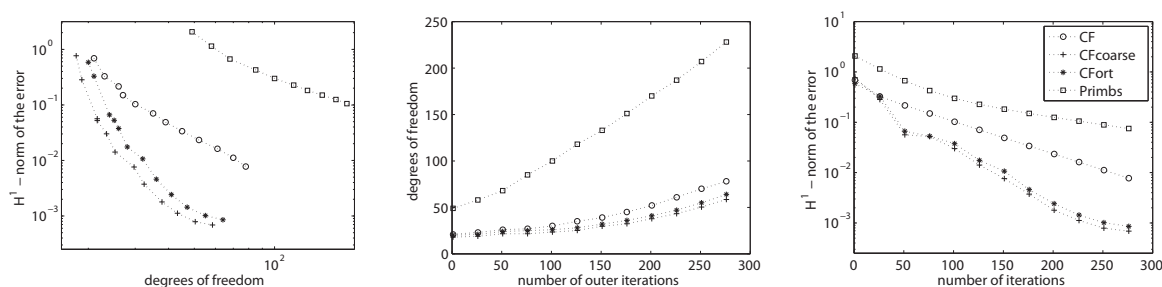


Figure 3: The convergence history for adaptive wavelet scheme with various wavelet bases.

References

- [1] BERTOLUZZA, S., FALLETTA, S.: *Building wavelets on $]0,1[$ at large scales*. J. Fourier Anal. and Appl., Vol. 9, No. 3, pp. 261-288, 2003.
- [2] CHUI, C. K., QUAK, E.: *Wavelets on a bounded interval*. In: Numerical Methods of Approximation Theory (Braess, D., Schumaker, L.L., eds.), Birkhäuser, pp. 53-75, 1992.
- [3] CARNICER, J. M., DAHMEN, W., PEÑA, J. M.: *Local decompositions of refinable spaces*. Appl. Comp. Harm. Anal., Vol. 3, pp. 127-153, 1996.
- [4] ČERNÁ, D.: *Biorthogonal wavelets*. Ph.D. thesis, Charles University, Prague, 2008.
- [5] ČERNÁ, D., FINĚK, V.: *Adaptive wavelet-frame method with stable boundary adapted bases*. In: ICNAAM 2008 (Simos T.E. et al., eds.), AIP Conference Proceedings, Vol. 1048, American Institute of Physics, New York, pp. 130-133, 2008.
- [6] ČERNÁ, D., FINĚK, V.: *Optimized construction of biorthogonal spline-wavelets*. In: ICNAAM 2008 (Simos T.E. et al., eds.), AIP Conference Proceedings 1048, American Institute of Physics, New York, pp. 134-137, 2008.
- [7] COHEN, A., DAUBECHIES, I., FEAUVEAU, J. C.: *Biorthogonal bases of compactly supported wavelets*. Comm. Pure and Appl. Math., Vol. 45, pp. 485-560, 1992.
- [8] COHEN, A., DAHMEN, W., DEVORE, R.: *Adaptive wavelet methods II - beyond the elliptic case*. Found. Math., Vol. 2, pp. 203-245, 2002.
- [9] DAHMEN, W.: *Stability of multiscale transformations*. J. Fourier Anal. Appl., Vol. 4, pp. 341-362, 1996.
- [10] DAHMEN, W., KUNOTH, A., URBAN, K.: *Biorthogonal spline wavelets on the interval - stability and moment conditions*. Appl. Comp. Harm. Anal., Vol. 6, pp. 132-196, 1999.
- [11] DAHMEN, W., KUNOTH, A., URBAN, K.: *Wavelets in numerical analysis and their quantitative properties*. In: Surface fitting and multiresolution methods (Le Méhauté, A., Rabut, C., Schumaker, L., eds.), Vol. 2, pp. 93-130, 1997.
- [12] PRIMBS, M.: *Stabile biorthogonale Spline-Waveletbasen auf dem Intervall*. Dissertation, Universität Duisburg-Essen, 2006.

Current address

Dana Černá, Mgr. Ph.D.

Department of Mathematics and Didactics of Mathematics, Technical University in Liberec,
Studentská 2, Liberec, 46117, Czech Republic, dana.cerna@tul.cz

Václav Finěk, RNDr. Ph.D.

Department of Mathematics and Didactics of Mathematics, Technical University in Liberec,
Studentská 2, Liberec, 46117, Czech Republic, vaclav.finek@tul.cz

USING PROGRAM *MATHEMATICA* FOR SOLUTION OF BEAM ON ELASTIC FOUNDATION

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] and [2].

2 Theoretical background

Let us consider a prismatical beam with length 2ℓ , 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 at that section. Under such conditions the reaction per unit length of the bar can be represented 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 not only frames but also foundation, which is a pretty complicated problem. Complicated is physical description of these problems, because complicated is description of elastic foundation, which has influence of a lot of factors of foundation.

In studying the deflection curve of the beam we use the differential equation

$$EI \frac{d^4 w}{dx^4} + q = q_0, \quad (1)$$

where EI is bending stiffness, q denotes the reaction of intensity of the load acting on the beam and q_0 is the uniform load, which is the intensity of weight of beam. Hence

$$q = k w. \quad (2)$$

and equation (1) becomes

$$EI \frac{d^4 w}{dx^4} + k w = q_0, \quad (3)$$

Let us consider a finite length of beam with the load F acting at the point $x = 0$ and uniform load q_0 , which represents the weight of beam, in the Fig. 1.

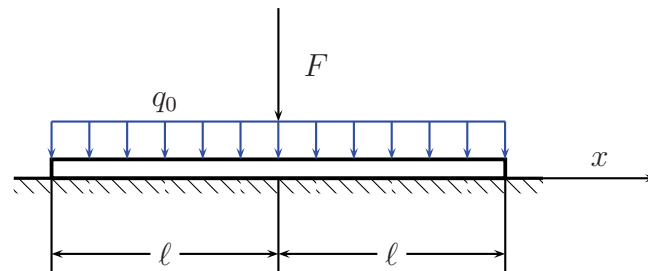


Figure 1: Beam on elastic foundation

For solution of problem in Fig. 1 we used method of superposition, which our problem divides into solution of two problems in Fig. 2.

Because we have axis of symmetry, we can solve only one half of beam, for example in the Fig. 3.

Solution of differential equation (3) is in the form

$$w(x) = e^{-\beta x}(C_1 \cos \beta x + C_2 \sin \beta x) + e^{\beta x}(C_3 \cos \beta x + C_4 \sin \beta x) + w^*, \quad (4)$$

where C_1 , C_2 , C_3 and C_4 are integration constants, w^* is particular solution of differential equation and

$$\beta = \sqrt[4]{\frac{k}{4EI}}.$$

Integration constants for beam in Fig. 3 are found for the following boundary conditions

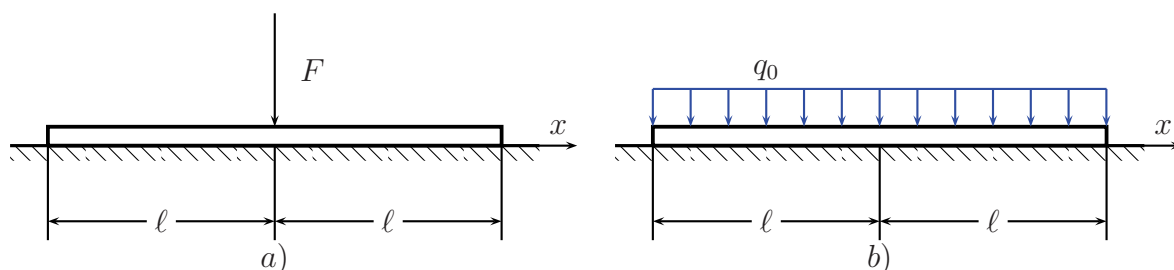


Figure 2: Method of superposition

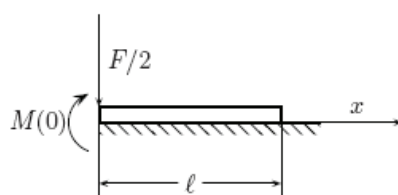


Figure 3: Model of beam on elastic foundation

1. $x = 0$ $w' = 0$,
2. $x = 0$ $w''' = -\frac{F}{2EI}$,
3. $x = \ell$ $w''' = 0$,
4. $x = \ell$ $w'' = 0$

and for beam in Fig. 2b) (x is start in middle of beam) are follow

1. $x = -\ell$ $w'' = 0$,
2. $x = -\ell$ $w''' = 0$,
3. $x = +\ell$ $w'' = 0$,
4. $x = +\ell$ $w''' = 0$.

3 Solution using program *Mathematica*

Finding of integration constant for given problem of analytical solution is very lengthy. Using program *Mathematica* is possible find integration constant in (4) following way for case in the Fig. 3

$$w[x_-] := e^{-\beta x}(C1 * \text{Cos}[\beta x] + C2 * \text{Sin}[\beta x]) + e^{\beta x}(C3 * \text{Cos}[\beta x] + C4 * \text{Sin}[\beta x])$$

$$\text{Solve}[\{w'[0] == 0, w'''[0] == \frac{2 F \beta^4}{k}, w'''[\frac{L}{2}] == 0, w''[\frac{L}{2}] == 0\}, \{C1, C2, C3, C4\}] // \text{Simplify}$$

Result of solution is following

$$\begin{aligned} C_1 &= \bar{B} e^{L\beta} (2 + e^{L\beta} + \cos L\beta + \sin L\beta), \\ C_2 &= \bar{B} e^{L\beta} (e^{L\beta} - \cos L\beta + \sin L\beta), \\ C_3 &= \bar{B} [1 + e^{L\beta} (2 + \cos L\beta - \sin L\beta)], \\ C_4 &= \bar{B} [-1 + e^{L\beta} (\cos L\beta + \sin L\beta)], \end{aligned} \quad (5)$$

where

$$\bar{B} = \frac{F\beta}{2k(e^{2L\beta} + 2e^{L\beta} \sin L\beta - 1)}. \quad (6)$$

Bending moment in location x is defined by [1, 2] as

$$\begin{aligned} M_B(x) &= -EI w''(x) \\ M_B(x) &= -EI \{2e^{-\beta x} \beta^2 [(-C_2 + C_4 e^{2\beta x}) \cos \beta x + (C_1 - C_3 e^{2\beta x}) \sin \beta x]\} \end{aligned} \quad (7)$$

and transversal force in location x is

$$\begin{aligned} V(x) &= \frac{dM_o(x)}{dx} = -EI w'''(x) = \\ &= -2e^{\beta x} \beta^3 EI \{[C_1 + C_2 + e^{2\beta x} (C_4 - C_3)] \cos \beta x - [C_1 - C_2 + e^{2\beta x} (C_3 + C_4)] \sin \beta x\} \end{aligned} \quad (8)$$

Result of solution for case in the Fig. 2 b) is following

$$\begin{aligned} C_1 &= 0, \\ C_2 &= 0, \\ C_3 &= 0, \\ C_4 &= 0, \end{aligned} \quad (9)$$

and

$$w^* = \frac{q_0}{k}. \quad (10)$$

Bending moment in location x is defined by [1, 2] as

$$M_B(x) = -EJ w''(x) = 0 \quad (11)$$

and transversal force in location x is

$$V(x) = \frac{dM_o(x)}{dx} = -EJ w'''(x) = 0 \quad (12)$$

Finally solution of our problems is sum of result for both case in the Fig. 2.

Above was find solution for differential equation (3), when we known, solution of this equation is in the form (4), where integration constant and particular part of solution differential equation was solved from boundary condition. Advantage of program *Mathematica* is follow, user have not know the shape of solution differential equation, the program calculate the shape of result using boundary condition by command DSolve, for example for case in Fig. 3

DSolve[{ $w''''[x] + 4\beta^4 w[x] == 0$, $w'[0] == 0$, $w'''[0] == \frac{2F\beta^4}{k}$, $w'''[L/2] == 0$, $w''[L/2] == 0$ }, w, x]

4 Example of solution of real beam on elastic foundation

For the beam on elastic foundation in the Fig. 1, let us consider following parameters: force $F = 10^5$ N, Young's modulus of beam $E = 2 \cdot 10^5$ MPa, moment of inertia of cross-section area $J = 1,06666 \cdot 10^{-3} \text{ m}^4$, length of beam $L = 1,6$ m modulus of elastic foundation $k = 20 \cdot 10^6 \text{ Nm}^{-2}$ and uniform load $q_0 = 6130 \text{ N/m}$.

For given parameters we have after calculation following integration constant $C_2 = 0,000791175$, $C_3 = 0,000315724$, $C_4 = 0,000187003$ and particular solution of differential equation $w^* = q_0/k = 3,065 \cdot 10^{-4} \text{ m}$ and put this result to equation (4) the function of deflection is

$$w(x) = e^{-0,391271x}(0,0012939 \cos 0,391271x + 0,000791175 \sin 0,391271x) + \quad (13)$$

$$+ e^{0,391271x}(0,000315724 \cos 0,391271x + 0,000187003 \sin 0,391271x) + 0,0003065.$$

Using command DSolve we get same function of deflection. Maximum of this function is in the location of external force F (at position $x = 0$), where value of deflection is $w(0) = 0.00191613 \text{ m}$, graphically the deflection in range $x \in \langle 0, L \rangle$ is the Fig. 4.

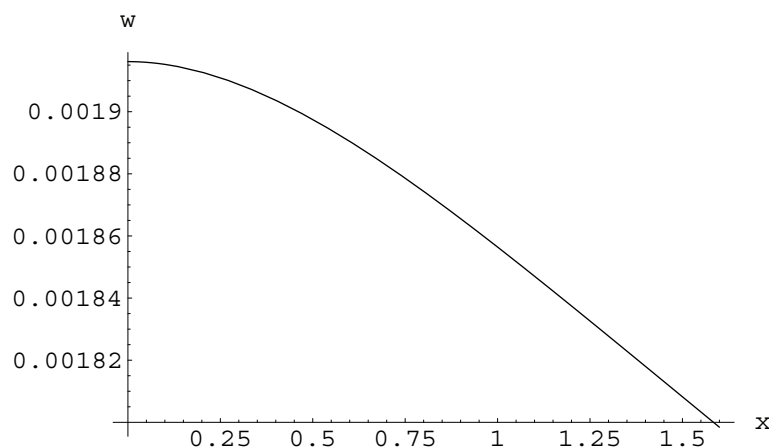


Figure 4: Graphically representation of deflection function in the range $x \in \langle 0, L \rangle$

Application of integration constant to equation (7) and (8) we get function of bending

moment and transversal load (addition from (11) and (12) is zero)

$$M_B(x) = -EI \{e^{-0,391271x} [(-0,000242247 + 0,0000572578e^{0.782542x}) \cos 0,391271x \quad (14) \\ + (0,000396175 - 0,0000966702e^{0.782542x}) \sin 0,391271x]\}$$

$$V(x) = -EI \{e^{-0,391271x} [(-0,000249796 + 0,0000154209e^{0.782542x}) \cos 0,391271x \quad (15) \\ + (0,0000602276 - 0,0000602276e^{0.782542x}) \sin 0,391271x]\}$$

Graphical representation of bending moment function is in the Fig. 5 and function of transversal load is in the Fig. 6.

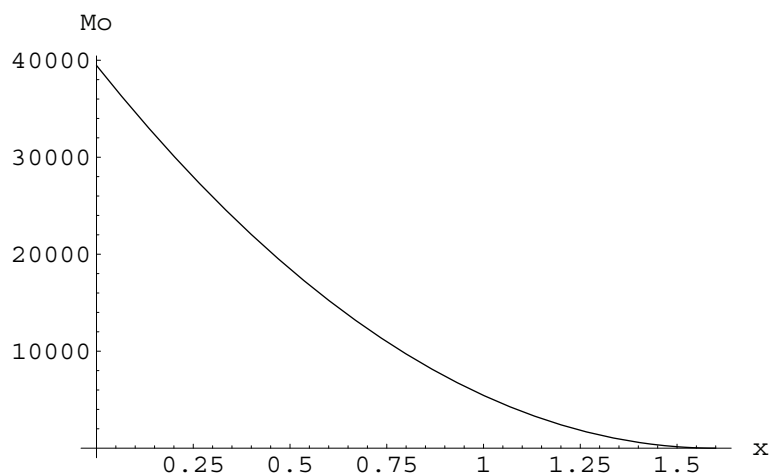


Figure 5: Function of bending moment in the range $x \in \langle 0, L \rangle$

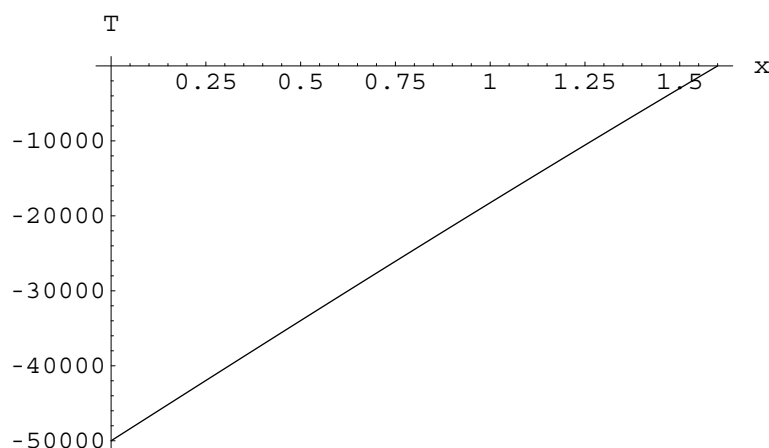


Figure 6: Function of transversal load in the range $x \in \langle 0, L \rangle$

5 Conclusion

Solution of beam on the elastic foundation is statically indeterminate problem of mechanics. In this paper is solution of beam load by external force and uniform load, which is equal to weight of beam. 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. Last chapter of this paper present some for given parameter, which is graphically present in Fig. 4, Fig. 5 and Fig. 6.

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.

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 .

LARGE SERIES-PARALLEL STRUCTURES AND METHODS USABLE FOR APPROXIMATE THEIR MTTF IN REPARABLE SYSTEMS

KOVÁČOVÁ Monika, (SK), JANČO Roland, (SK)

Abstract. Reliability and availability analysis of repairable systems is generally performed using stochastic processes, including Markov, semi-Markov and semi-regenerative processes. It is not easy to approximate these quantities for large series-parallel structures. In this contribution we will focus on such methods for approximate these expressions and we will compare them on numerical way. We will show three methods for the possible approximation on the same serial-parallel structure with brief mathematical background.

Key words: Reliability, availability, large series-parallel structures, active redundancy, one repair crew, repair priority, no future failures at system down.

Mathematics Subject Classification: 68M15, 60K20

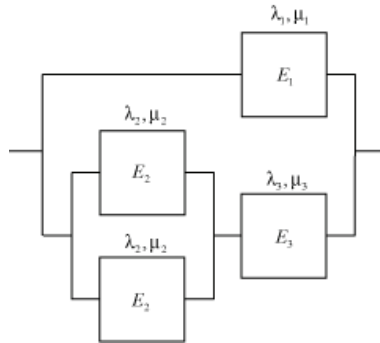
1 Introduction

Reliability and availability analysis of repairable systems is generally performed using stochastic processes, including Markov, semi-Markov and semi-regenerative processes. It is not easy to approximate these quantities for large series-parallel structures. In this contribution we will focus on such methods for approximate these expressions. We will show the possible approximation for several methods applied on the same structure.

The main problem in approximating process is the reliability and availability calculation of large series-parallel systems rapidly becomes time consuming, even if constant failure rate λ_i and the repair rate μ_i is assumed for each element E_i of the reliability block diagram and only mean time to failure $MTTF_{so}$ or steady-state availability $PA_s = AA_s$ is required. **This is because the large number of states involved to the approximation process.** Let us consider the reliability block diagram with n elements. That diagram can reach

$$1 + \sum_{i=1}^n \prod_{k=n-i+1}^n k = 1 + \sum_{i=0}^n \frac{1}{i!} \approx e \cdot n!$$

states, consider all possible repair strategies. In this contribution we will see in detail on system with 4 elements described on the next picture.



This system can reach more than 50 states if the assumption of no future failure at system down were dropped. The situation requires using of approximate expressions, use of these expressions become thus very important. 2^n states holds for non-repairable systems or for system with totally independent elements. We will explain later how to sort large repairable systems. Sometimes also the assumptions of one repair crew and no future failure at system down are applied to the restricted rules. The assumption provided $\lambda_i \ll \mu_i$ holds for each element E_i is also important in approximation theory for system reliability and availability.

Here are some concrete restrictions which allow classify series-parallel structures to several classes and hence to simplify the computation severity.

1 Totally independent elements

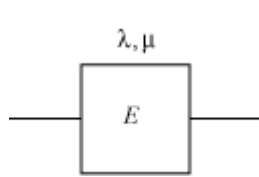
If each element of the reliability block diagram operates independently from every other, series-parallel structures can be reduced to one-item structures, which are themselves successively integrated into further series-parallel structured up to system level. This situation can be obtained in active redundancy, independent elements or one repair crew for each element cases. To each of the one-item structure obtained, the mean time to failure $MTTF_{so}$ and steady-state availability PA_s , calculated for underlying series-parallel structure, are used to calculate an equivalent $MTTR_s$ from $PA_s = \frac{MTTF_s}{MTTF_s + MTTR_s}$ using $MTTF_s = MTTF_{so}$.

To simplify calculations constant failure rate $\lambda_s = 1 / MTTF_{so}$ and constant repair rate $\mu_s = 1 / MTTR_s$ are assumed for each of the one-item structures obtained.

The following basic structures for investigation of large series-parallel systems by *assuming totally independent elements* (each element operates and is repaired independently of every other element), *constant failure and repair rates* (λ, μ) , *active redundancy*, *one repair crew for element*, *ideal*

failure recognition, Markov process with $\lambda_v = 0$, and $MTTF_{SO} = \frac{1}{k \lambda \binom{n}{k}} \left(\frac{\mu}{\lambda} \right)^{n-k}$ (n -repair crew,

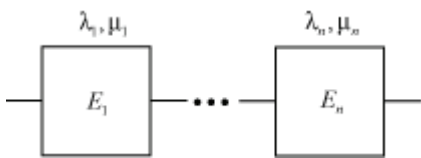
active redundancy, $\frac{\lambda}{\mu} \ll 1$), and $PA_S = 1 - \frac{k}{n-k+1} \binom{n}{k} \left(\frac{\lambda}{\mu} \right)^{n-k+1}$ are used to simplify the notation, approximations valid for $\lambda_i \ll \mu_i$, $PA_S = AA_S$ = asymptotic and steady-state point and average availability (denoted A).



$$\lambda_S = \lambda$$

$$\mu_S = \mu$$

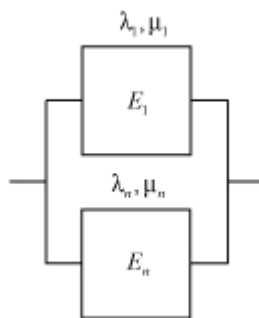
$$PA_S = \frac{1}{1 + \lambda_S / \mu_S} \approx 1 - \frac{\lambda_S}{\mu_S} \Rightarrow \mu_S = \frac{\lambda_S PA_S}{1 - PA_S} \approx \frac{\lambda_S}{1 - PA_S}$$



$$PA_S = PA_1 \dots PA_n \approx 1 - \left(\frac{\lambda_1}{\mu_1} + \dots + \frac{\lambda_n}{\mu_n} \right)$$

$$\lambda_S = \lambda_1 + \dots + \lambda_n \Rightarrow \mu_S = \frac{\lambda_S}{1 - PA_S} \approx \frac{\lambda_1 + \dots + \lambda_n}{\lambda_1 / \mu_1 + \dots + \lambda_n / \mu_n}$$

1-out-of-2
(active)



$$PA_S = PA_1 + PA_2 - PA_1 PA_2 \approx 1 - \frac{\lambda_1 \lambda_2}{\mu_1 \mu_2}$$

$$\frac{1}{\lambda_S} \equiv MTTF_{SO} \approx \frac{\mu_1 \mu_2}{\lambda_1 \lambda_2 (\mu_1 + \mu_2)} \Rightarrow$$

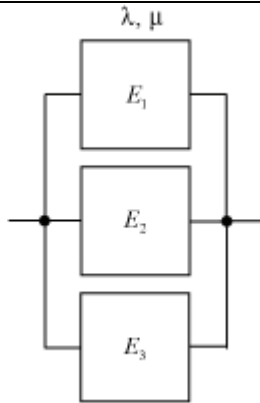
$$\mu_S \approx \frac{\lambda_S}{1 - PA_S} = \mu_1 + \mu_2$$

2-out-of-3 active
 $E_1 = E_2 = E_3 = E$

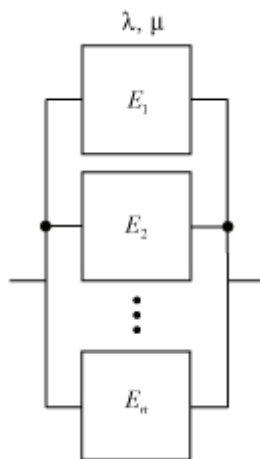
$$PA_S = 3PA^2 - 2PA^3 \approx 1 - \frac{3(\lambda / \mu)^2}{1 + 3\lambda / \mu} \approx 1 - 3 \left(\frac{\lambda}{\mu} \right)^2$$

$$1 / \lambda_S \equiv MTTF_{SO} = \frac{5\lambda + \mu}{6\lambda^2} \approx \frac{\mu}{6\lambda^2} \Rightarrow$$

$$\mu_S \approx \frac{\lambda_S}{1 - PA_S} \approx 2\mu$$



k -out-of- n active
 $(E_1 = \dots E_n = E)$



$$PA_S = 1 - \frac{k}{n-k+1} \binom{n}{k} \left(\frac{\lambda}{\mu} \right)^{n-k+1}$$

$$\frac{1}{\lambda_S} \equiv MTTF_{SO} \approx \frac{1}{k \lambda \binom{n}{k}} \left(\frac{\mu}{\lambda} \right)^{n-k} \Rightarrow$$

$$\mu_S \approx \frac{\lambda_S}{1 - PA_S} \approx (n-k+1) \mu$$

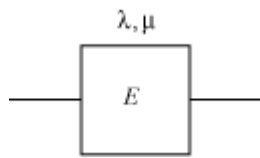
2 Macro structures

Macro –structure is a series, parallel, or simple series-parallel structure which is considered as a one-item structure for calculations at higher levels (integration into further macro structures up to system level). This system satisfies the assumptions

- *Continuous operation.* Each element of the system is in operating or reserved state, when not under repair or waiting for repair.
- *No further failures at system down.* At system down, the system is repaired (restored) according to a given maintenance strategy to an up state at system level from which operation is continued, failures during a repair at system down are not considered.
- *Only one repair crew.* At system level only one repair crew is available, repair is performed according to a stated strategy, e.g. first-in/first-out
- *Redundancy.* Redundant elements are repaired without interruption of operation at system levels, failure of redundant parts is immediately recognized.

- *States.* Each element in the reliability block diagram has only two states (good or failed), after repair it is as-good-as new.
- *Independence.* Failure-free and repair times of each element are stochastically independent, > 0 , and continuous random variables with finite mean ($MTTF, MTTR$) and variance. Failure-free time is in this case used as a synonym for failure-free operating time.
- *Support.* Preventive maintenance is neglected, fault coverage, switching and logistic support are ideal. It means repair time = restoration time = down time.

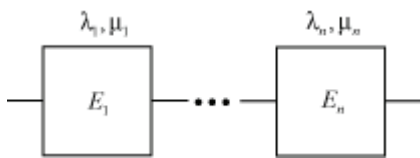
In these cases the procedure is similar to that of point 1 above. The following basic macrostructures for the investigation of large series-parallel systems by successive building of macrostructures bottom up to system level are used to simplify notation, approximations are valid in case $\lambda_i \ll \mu_i$.



$$\lambda_S \equiv \lambda$$

$$\mu_S \equiv \mu$$

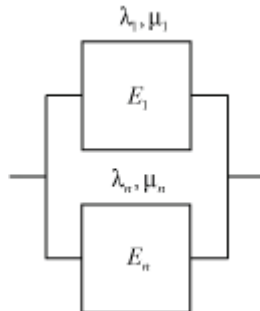
$$PA_S = \frac{1}{1 + \lambda_S / \mu_S} \approx 1 - \frac{\lambda_S}{\mu_S} \Rightarrow \mu_S = \frac{\lambda_S PA_S}{1 - PA_S} \approx \frac{\lambda_S}{1 - PA_S}$$



$$PA_S \approx 1 - \left(\frac{\lambda_1}{\mu_1} + \dots + \frac{\lambda_n}{\mu_n} \right)$$

$$\lambda_S = \lambda_1 + \dots + \lambda_n \Rightarrow \mu_S = \frac{\lambda_S}{1 - PA_S} \approx \frac{\lambda_1 + \dots + \lambda_n}{\lambda_1 / \mu_1 + \dots + \lambda_n / \mu_n}$$

1-out-of-2
(active)



$$PA_S \approx 1 - \frac{\lambda_1 \lambda_2}{\mu_1 \mu_2} (\mu_1^2 + \mu_2^2)$$

$$\frac{1}{\lambda_S} \equiv MTTF_{SO} \approx \frac{\mu_1 \mu_2}{\lambda_1 \lambda_2 (\mu_1 + \mu_2)} \Rightarrow$$

$$\mu_S \approx \frac{\lambda_S}{1 - PA_S} = \mu_1 \mu_2 \frac{\mu_1 + \mu_2}{\mu_1^2 + \mu_2^2} \quad (= \mu \text{ for } \mu_1 = \mu_2)$$

1-out-of-2 (active)
 $E_1 = E_2 = E$
repairing priority on E_v

$$PA_S \approx 1 - \frac{\lambda_v}{\mu_v} - \frac{2(\lambda / \mu)^2}{1 + 2\lambda / \mu}$$

	$\frac{1}{\lambda_S} \equiv MTTF_{SO} = \frac{1}{\lambda_v + \frac{2\lambda^2}{\mu + 3\lambda + \lambda_v}} \approx \frac{1}{\lambda_v + 2\lambda^2 / \mu} \Rightarrow$ $\mu_S \approx \frac{\lambda_S}{1 - PA_S} \approx \mu_v \quad \text{for } \mu_v \approx \mu$
<p>2-out-of-3 active $E_1 = E_2 = E_3 = E$ repair on E_v</p>	$PA_S \approx 1 - \frac{\lambda_v}{\mu_v} - \frac{6(\lambda / \mu)^2}{1 + 3\lambda / \mu}$ $1 / \lambda_S \equiv MTTF_{SO} \approx \frac{1}{\lambda_v + 6\lambda^2 / \mu} \Rightarrow$ $\mu_S \approx \frac{\lambda_S}{1 - PA_S} \approx \mu_v \frac{\lambda_v + 6\lambda^2 / \mu}{\lambda_v + \frac{6\lambda^2 / \mu}{1 + 3\lambda / \mu} \cdot \frac{\mu_v}{\mu}} \approx \mu_v \quad \text{for } \mu_v \approx \mu$
<p>k-out-of-n active $(E_1 = \dots E_n = E)$ repair on E_v</p>	$PA_S \approx 1 - \frac{\lambda_v}{\mu_v} - \frac{n!}{(k-1)!} \left(\frac{\lambda}{\mu} \right)^{n-k+1}$ $\frac{1}{\lambda_S} \equiv MTTF_{SO} \approx \frac{1}{\lambda_v + \lambda \frac{n!}{(k-1)!} \left(\frac{\lambda}{\mu} \right)^{n-k}} \Rightarrow$ $\mu_S \approx \frac{\lambda_S}{1 - PA_S} \approx \mu_v \quad \text{for } \mu_v \approx \mu$

3 One repair crew and no future failures at system down

In this case the following assumptions are applied in many practical applications.

- *No further failures at system down.* At system down, the system is repaired (restored) according to a given maintenance strategy to an up state at system level from which operation is continued, failures during a repair at system down are not considered.
- *Only one repair crew.* At system level only one repair crew is available, repair is performed according to a stated strategy, e.g. first-in/first-out

No future failures at system down means that failures during a repair at system level are neglected. This assumption has no influence on the reliability function at system level and its influence on the availability is limited, if $\lambda_i \ll \mu_i$ can be assumed for each elements E_i .

4 Cutting states

Removing the states with more than k failures from the diagram of transition probabilities in time interval $(t, t + \delta]$, or the state transition **diagram produces in general an important reduction of the state diagram**. The choice of k , often $k = 2$, is based on the required precision.

An upper bound of the error for the asymptotic and steady-state value of the point and average availability $PA_s = AA_s$ has been given in [4]. In this work the upper bound is estimated based on the mapping of states with k failures at the system level in the state Z_k of a birth and dead process and

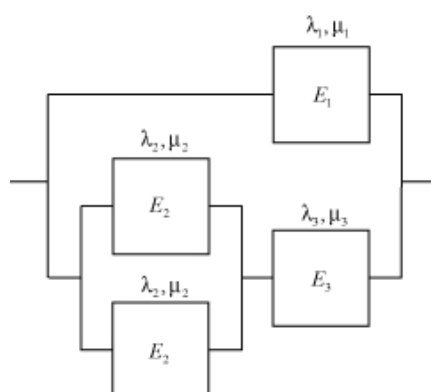
using $P_j \geq \sum_{i=j+1}^n P_i$, $j = 0, \dots, n-1$ valid for $2(\lambda_1 + \dots + \lambda_n) < \min(\mu_1, \dots, \mu_n)$

Combination of the above method is also possible. In many cases, series elements must be grouped before any analysis. Considering that the steady-state probability for states with more than one failure decreasing rapidly as the number of failures increases. All methods given above yield good approximate expressions for $MTTF_{so}$ and PA_s in practical applications.

However, in case we want to estimate the unavailability $1 - PA_s$, method 1 can deliver lower values, for instance a factor 2 with an order of magnitude $(\lambda / \mu)^2$ for 1-out-of-2 active redundancy. In general there is very difficult to compare the above mention methods. Numerical investigations show a close convergence of a result given by different methods for practical example with extremely low value μ / λ . Comparison will be given in the next section of this article.

5 Comparison for above mentioned method

In this section we will illustrate how the method 1 and the method 2 work. We will consider the same practical example with the following reliability block diagram.

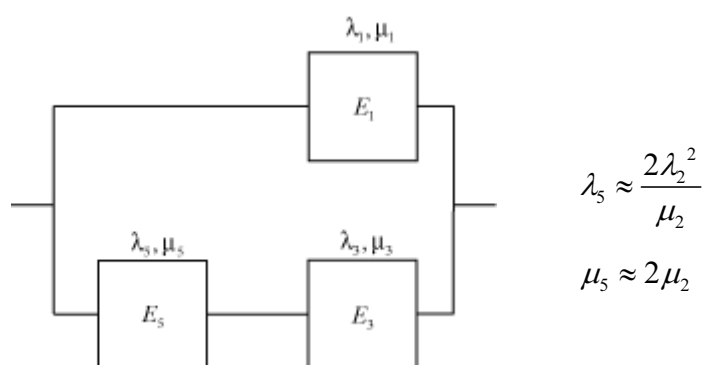


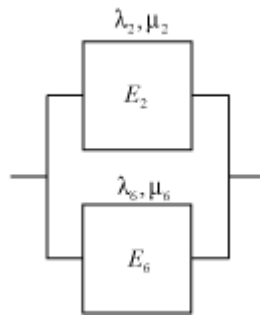
We will assume the system is new at time $t = 0$, active redundancy, constant failure rates $\lambda_1, \lambda_2, \lambda_3$, constant repair rates μ_1, μ_2, μ_3 , repair priority E_1, E_3, E_2 . Except for some series elements, the reliability block diagram describe an uninterruptible power supply (UPS) used for instance to buffer electrical power network failures in computer systems.

Although the system is limited to 4 elements, the stochastic process describing that system would contain more than 50 states if the assumption of no future failure at system down were dropped. Assuming these assumptions the state space will be reduced to 12 states. In the following, the mean time to failure ($MTTF_{SO}$) and the asymptotic and steady-state point and average availability ($PA_S = AA_S$) of that system is described using the method 1, 2 and 3. In numerical experiments both methods deliver the good approximation.

Method 1

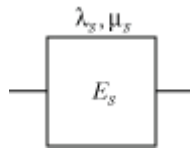
Using the section 1 we have





$$\lambda_6 \approx \lambda_3 + \lambda_5$$

$$\mu_6 \approx \frac{\lambda_5 + \lambda_3}{\lambda_5 / \mu_5 + \lambda_3 / \mu_3}$$



$$\lambda_8 \approx \frac{\lambda_1 \lambda_6 (\mu_1 + \mu_6)}{\mu_1 \mu_6}$$

$$\mu_8 \approx \mu_1 + \mu_6$$

From these equations it follows that

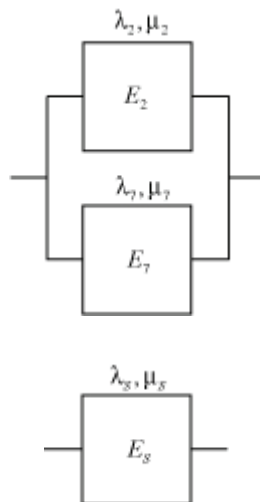
$$1 / MTTF_{SO} \equiv \lambda_8 \approx \lambda_1 \left[\frac{\lambda_3}{\mu_1} + \frac{2\lambda_2^2}{\mu_1 \mu_2} + \frac{\lambda_3}{\mu_3} + \left(\frac{\lambda_2}{\mu_2} \right)^2 \right]$$

and

$$PA_S = 1 - \frac{\lambda_8}{\mu_8} \approx 1 - \frac{\lambda_1}{\mu_1} \left[\frac{\lambda_3}{\mu_3} + \left(\frac{\lambda_2}{\mu_2} \right)^2 \right]$$

Method 2

Using the section 2 we have



$$\lambda_7 \approx \lambda_3 + \frac{2\lambda_2^2}{\mu_2}$$

$$\mu_7 \approx \frac{\mu_3 (2\lambda_2^2 + \mu_2 \lambda_3) (1 + 2\lambda_2 / \mu_2)}{\mu_2 \lambda_3 + 2\lambda_2 \lambda_3 + 2\lambda_2^2 \mu_3 / \mu_2}$$

$$\lambda_8 \approx \frac{\lambda_1 \lambda_7 (\mu_1 + \mu_7)}{\mu_1 \mu_7}$$

$$\mu_8 \approx \mu_1 \mu_7 \frac{\mu_1 + \mu_7}{\mu_1^2 + \mu_7^2}$$

From the previous equations it follows that

$$1 / MTTF_{SO} \equiv \lambda_s \approx \lambda_1 \left(\frac{2\lambda_2^2 + \mu_2\lambda_3}{\mu_1\mu_2} + \frac{\mu_2\lambda_3 + 2\lambda_2\lambda_3 + 2\mu_3\lambda_2^2 / \mu_2}{\mu_2\mu_3(1 + 2\lambda_2 / \mu_2)} \right)$$

and

$$PA_s \approx 1 - \frac{\lambda_s}{\mu_s} \approx 1 - \frac{2\lambda_2^2 + \mu_2\lambda_3}{\mu_2} \left(\frac{\lambda_1}{\mu_1^2} + \frac{\lambda_1(\mu_2\lambda_3 + 2\lambda_2\lambda_3 + 2\mu_3\lambda_2^2 / \mu_2)^2}{(2\lambda_2^2 + \mu_2\lambda_3)^2(1 + 2\lambda_2 / \mu_2)^2\mu_3^2} \right)$$

$$\approx 1 - \frac{\lambda_1}{\mu_1} \left(\frac{\lambda_3}{\mu_3} + \frac{2\lambda_2^2}{\mu_2\mu_3} \right) \left(\frac{\mu_3}{\mu_1} + \frac{(\mu_2\lambda_3 + 2\lambda_2\lambda_3 + 2\mu_3\lambda_2^2 / \mu_2)^2\mu_1 / \mu_3}{\frac{\lambda_1(\mu_2\lambda_3 + 2\lambda_2\lambda_3 + 2\mu_3\lambda_2^2 / \mu_2)^2}{(2\lambda_2^2 + \mu_2\lambda_3)^2(1 + 2\lambda_2 / \mu_2)^2}} \right)$$

Method 3

We will show only partially using of this method due to lack of place. Using the previous explanation and the relations from the section 1 we have the following system of algebraic equations for the mean time failures ($M_i = MTTF_{Si}$).

$$\begin{aligned} \rho_0 M_0 &= 1 + \lambda_1 M_1 + 2\lambda_2 M_2 + \lambda_3 M_3 & \rho_1 M_1 &= 1 + \mu_1 M_0 + 2\lambda_2 M_7 \\ \rho_2 M_2 &= 1 + \mu_2 M_0 + \lambda_3 M_4 + \lambda_2 M_6 + \lambda_1 M_7 & \rho_3 M_3 &= 1 + \mu_3 M_0 + 2\lambda_2 M_4 \\ \rho_4 M_4 &= 1 + \mu_3 M_2 + \lambda_2 M_5 & \rho_5 M_5 &= 1 + \mu_3 M_6 \\ \rho_6 M_6 &= 1 + \mu_2 M_2 + \lambda_3 M_5 & \rho_7 M_7 &= 1 + \mu_1 M_2 \end{aligned}$$

where

$$\begin{aligned} \rho_0 &= \lambda_1 + 2\lambda_2 + \lambda_3 & \rho_1 &= \mu_1 + 2\lambda_2 + \lambda_3 & \rho_2 &= \mu_2 + \lambda_1 + \lambda_2 + \lambda_3 \\ \rho_3 &= \mu_3 + 2\lambda_2 + \lambda_1 & \rho_4 &= \mu_3 + \lambda_2 + \lambda_1 & \rho_5 &= \mu_3 + \lambda_1 \\ \rho_6 &= \mu_2 + \lambda_3 + \lambda_1 & \rho_7 &= \mu_1 + \lambda_3 + \lambda_2 & \rho_8 &= \mu_1 \\ \rho_9 &= \mu_1 & \rho_{10} &= \mu_1 & \rho_{11} &= \mu_1 \end{aligned}$$

From the last two groups of equations it follows that

$$\frac{1}{\lambda_s} \equiv MTTF_{s0} = \frac{a_5 + a_6(a_8 + a_9 a_{10}) + a_7 a_{10}}{1 - a_6 a_{12} - a_{11}(a_7 + a_6 a_9)}$$

where

$$\begin{aligned}
 a_1 &= \frac{1}{\rho_4} + \frac{\lambda_2}{\rho_4 \rho_5} \left(1 + \mu_3 \frac{\lambda_3 + \rho_5}{\rho_5 \rho_6 - \lambda_3 \mu_3} \right) & a_2 &= \frac{\lambda_2 \mu_2 \mu_3}{\rho_4 (\rho_5 \rho_6 - \lambda_3 \mu_3)} + \frac{\mu_3}{\rho_4} \\
 a_3 &= \frac{1}{\rho_3} (1 + 2\lambda_2 a_1) & a_4 &= \frac{2\lambda_2}{\rho_3} a_2 & a_5 &= \frac{1 + \lambda_3 a_3}{\rho_0 - \lambda_3 \mu_3 / \rho_3} \\
 a_6 &= \frac{\lambda_1}{\rho_0 - \lambda_3 \mu_3 / \rho_3} & a_7 &= \frac{2\lambda_2 + \lambda_3 a_4}{\rho_0 - \lambda_3 \mu_3 / \rho_3} & a_8 &= \frac{1 + 2\lambda_2 / \rho_7}{\rho_1} \\
 a_9 &= \frac{2\lambda_2 \mu_1}{\rho_1 \rho_7} & a_{10} &= \frac{1 + \lambda_3 a_1 + \frac{\lambda_2 \lambda_3 + \lambda_2 \rho_5}{\rho_5 \rho_6 - \lambda_3 \mu_3} + \frac{\lambda_1}{\rho_7}}{\rho_2 - \lambda_3 a_2 - \frac{\lambda_2 \mu_2 \rho_5}{\rho_5 \rho_6 - \lambda_3 \mu_3} - \frac{\lambda_1 \mu_1}{\rho_7}} \\
 a_{11} &= \frac{\mu_2}{\rho_2 - \lambda_3 a_2 - \frac{\lambda_2 \mu_2 \rho_5}{\rho_5 \rho_6 - \lambda_3 \mu_3} - \frac{\lambda_1 \mu_1}{\rho_7}} & a_{12} &= \frac{\mu_1}{\rho_1}
 \end{aligned}$$

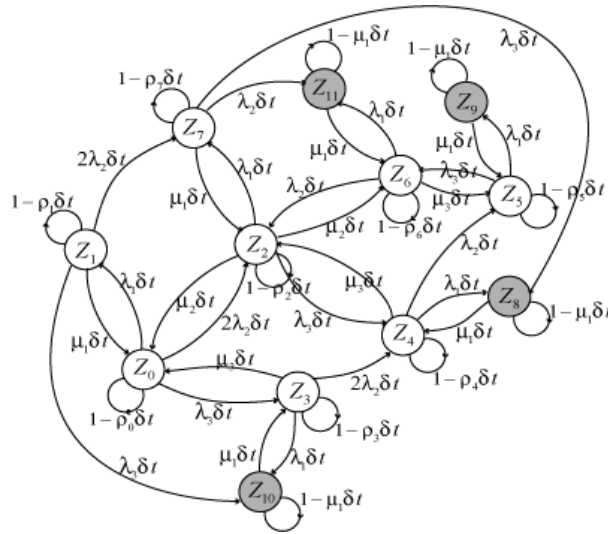
The same problem we can solve also for the asymptotic and steady state value for the point availability and average availability. These expressions were not described in the section 1, so we will not compute these quantities. The general principle for computed them is very similar to method 2.

We will see that there is large system in the method 2 and there is need to solve them in a symbolic way. It is not easy task, but we recommend for solving large symbolic system the use Computer Algebra Systems such as Mathematica or Maple. There is also possible to use numerical approximation methods for solving non-linear system such as robust Nelder Mead, or Newton method, or Fixed point method. These could be applied only for concrete value problem applications.

The following figure shows the situation computed in the previous in a simple graphical way. We note that the problem was solved under assumptions: active redundancy, one repair crew, repair priority in the sequences E_1, E_3, E_2 , no future failures at system down, ideal failure recognition, Markov process $\rho_i = \sum_j \rho_{ij}$.

The possible states in our practical examples were (they can help to better understand the principle of creation the previous equations and expressions).

$$\begin{array}{llllll}
 \rho_{01} = \lambda_1 & \rho_{02} = 2\lambda_2 & \rho_{03} = \lambda_3 & \rho_{10} = \mu_1 & \rho_{17} = 2\lambda_2 & \rho_{110} = \lambda_3 & \rho_{20} = \mu_2 \\
 \rho_{24} = \lambda_3 & \rho_{26} = \lambda_2 & \rho_{27} = \lambda_1 & \rho_{30} = \mu_3 & \rho_{34} = 2\lambda_2 & \rho_{310} = \lambda_1 & \rho_{42} = \mu_3 \\
 \rho_{45} = \lambda_2 & \rho_{48} = \lambda_1 & \rho_{56} = \mu_3 & \rho_{59} = \lambda_1 & \rho_{62} = \mu_2 & \rho_{65} = \lambda_3 & \rho_{611} = \lambda_1 \\
 \rho_{72} = \mu_1 & \rho_{78} = \lambda_3 & \rho_{711} = \lambda_2 & \rho_{84} = \mu_1 & \rho_{95} = \mu_1 & \rho_{103} = \mu_1 & \rho_{116} = \mu_1
 \end{array}$$



6 Conclusion

An analytical comparison of the two previous methods is very difficult. During the numerical comparing it is easy verify that for $\lambda_1 = 1/100$, $\lambda_2 = 1/1000$, $\lambda_3 = 1/10000$ and $\mu_1 = 1$, $\mu_2 = 1/5$, $\mu_3 = 1/5$ method 1 gives $MTTF_S = 1.575 \cdot 10^5$, method 2 gives $MTTF_S = 1.528 \cdot 10^5$ and the method 3 gives $MTTF_S = 1.589 \cdot 10^5$. The methods were compared also for other λ_i, μ_i values such that $\lambda_i \ll \mu_i$ and they both give a good approximation for practical examples (see next table with several comparisons).

The results obtained with method 1 give higher value in $MTTF_S$, because of the assumption that each element has it own repair crew (totally independent elements). The main disadvantage of the method 3 is its computation severity, not only during the computation process but also in model creating process. In general can be show that all methods under each own assumption gives very similar approximation of result.

λ_1	1/100	1/100	1/1 000	1/1 000
λ_2	1/1 000	1/1 000	1/10 000	1/10 000
λ_3	1/10 000	1/10 000	1/100 000	1/100 000
μ_1	1	1/5	1	1/5
μ_2	1/5	1/5	1/5	1/5
μ_3	1/5	1/5	1/5	1/5
$MTTF_{SO}$ (method 1)	$1.575 \cdot 10^5$	$9.302 \cdot 10^4$	$1.657 \cdot 10^7$	$9.926 \cdot 10^6$
$MTTF_{SO}$ (method 2)	$1.528 \cdot 10^5$	$9.136 \cdot 10^4$	$1.652 \cdot 10^7$	$9.906 \cdot 10^6$
$MTTF_{SO}$ (method 3)	$1.589 \cdot 10^5$	$9.332 \cdot 10^4$	$1.658 \cdot 10^7$	$9.927 \cdot 10^6$

References

- [1] ARUNKUMAR S., Lee S.H.: Enumeration of all minimal cut-sets for a node pair in a graph, IEEE trans. Rel. 28(1987) 1, pp. 51-55
- [2] BIROLINI Alesandro. *Reliability Engineering, Theory and Practice*, Springer, 2007 5th. ed., pp. 593, ISBN: 978-3-540-49388-4
- [3] Bellcore SR-TSY -001171, Methods and procedures for System Reliability Analysis, 1989.
- [4] BERNET R.: Modellierung reparierbarer Systeme durch Markoff und Semiregenerative Prognose, PhD. Thesis 1992, ETH Zurich
- [5] MEEKER William Q., Escobar Luis A. Statistical Methods for Reliability Data (Wiley Series in Probability and Statistics), John Wiley & Sons, 1998, pp.712, ISBN: 0471143286
- [6] RIGDON Steven E., Basu Asit P.. Statistical Methods for the Reliability of Repairable Systems, Wiley-Interscience, 2000, pp. 224, ISBN: 0471349410
- [7] SHELDON M. Ross. *Stochastic Processes*, John Wiley & Sons, 1995, 2nd. ed., pp. 510, ISBN: 0471120626
- [8] VENKATARAMA Krishnan. *Probability and Random Processes*, John Wiley & Sons, 2006, 1st ed., pp. 723, ISBN: 0471703540

Current address

Roland Jančo, MSc. , PhD. ING-PAED IGIP

Institute of Applied Mechanics and Mechatronics, Section of Strength of Material, Faculty of Mechanical Engineering, Slovak University of Technology Bratislava, Nam. slobody 17, 812 31 Bratislava, Slovak Republic,
e-mail: roland.janco@stuba.sk .

Monika Kováčová, Mgr., PhD.

Institute of natural sciences, humanities and social sciences, Faculty of Mechanical Engineering, Slovak University of Technology Bratislava, Nam. slobody 17, 812 31 Bratislava, Slovenská Republika,
e-mail: monika.kovacova@stuba.sk

EVALUATION OF MEASURED DATA FROM RESEARCH OF PARAMETERS IMPACT ON FINAL BRIQUETTES DENSITY

KRIŽAN Peter, (SK), ŠOOŠ Ľubomír, (SK), MATÚŠ Miloš, (SK),
SVÁTEK Michal, (SK), VUKELIĆ, Djordje (SRB)

Abstract. This paper describes shortly evaluation of experiment for research of parameter impact on final briquettes quality. On our department we also give due attention to research of compacting process. One of the goals of the research is to detect the impact of some parameters on final briquettes quality. After theoretical analyze of lonely parameters impact we know that most distinguished impact on briquette quality have pressing temperature, material moisture, compacting pressure and fraction largeness. Contribution is describing process of data evaluation from detecting of signification and lonely parameters impact amount until mathematical model designing. Contribution is describing the design of mathematical model which describes impact of these parameters and their mutual interactions. Described experiment and this evaluation is valid only for compacting of pine sawdust.

Key words. briquetting, compacting, mathematical model, briquettes density

Mathematics Subject Classification: Primary 97M50, 62K15.

1 Introduction

Biomass forms present huge amounts of unutilized waste. Biofuels production is convenient way how to energetically utilize the waste. Before the waste becomes biofuel it has to be modified. Very interesting possibility is to compact the modified waste into the solid biofuels. This is performed by compacting technologies, e.g. briquetting or pelleting. The final products of briquetting and pelleting are briquettes with various shapes and sizes. The feature of these both technologies is pressing of material (waste) under high pressure. It is very important to produce briquettes with Standards given quality. Briquettes quality is evaluate mainly by briquette density. During the pressing process there are many parameters which influencing the final briquettes quality – density. On our department we made some analyses and experiment to detect the impact of these parameters.

2 Parameters which influencing final briquettes quality and compacting process [3]

We did many theoretical analyses about impact of parameters at compacting process. Therefore we divided analyzed parameters into the three basic groups because they don't have the same nature. First group of parameters related with type of pressing material, second group related with pressing technology and to the third group belong all constructional parameters:

1.) pressed material

- type,
- fraction largeness,
- moisture,
- temperature.

2.) parameters of pressing:

- pressing way,
- temperature in pressing chamber,
- compacting pressure in pressing chamber,
- pressing speed,
- holding time.

3.) constructional parameters:

- dimension, length and shape of pressing chamber,
- type, dimension and shape of pressing tool,
- material, treatment (surface roughness) and surface modification of pressing chamber and tool,
- counter pressure effecting on compacted briquette,
- length of cooling channel.

As you can see the final briquettes quality is influenced by many parameters. We decided that only four parameters have the most significant effect. These parameters are pressing temperature, compacting pressure, fraction largeness and material moisture.

Compacting pressure is the most important factor having influence particularly upon the strength of briquettes. The strength of briquettes increases with increasing pressure within its strength limit and tendency to absorb atmospheric humidity during the long term storage is decreasing.

Pressing temperature belongs together with compacting pressure to the most considerable parameters, what means - that it has significant effect on the quality and strength of briquettes. It determines the lignin excretion by cellular structures within the wood. Lignin is released under certain pressing temperature, which has to be unconditionally reached to assure undergoing process. Fraction largeness has also effect on compacting process. The bigger input fraction causes the higher power output required for compacting. Despite of that the briquette has lower homogeneity and strength. With increasing fraction largeness, the strength of binding is decreasing what is resulting in fast briquette disintegration in the process of burning (briquette is burning faster, which represents its disadvantage).

Material moisture has influence on lignin plastification. Water has positive role in growing tree because it is essential life condition for existence of every plant organism. Presence of water in tree that has been cut is undesirable. All recently known compressing technologies enable to compact material having relative moisture lower than 18%. The humidity around 12 % appears to be the value optimal for compacting, because if the moisture is very small or vice-versa very high, particles of material are not compact and briquettes may disintegrate.

3 Experiment – design, evaluation, results

In the experiment we focused on detecting the impact of these parameters. Therefore we designed experimental pressing stand which allow us to make changes (constructional and also technological) as we need according to experimental plan. Pressing stand has been designed and constructed and

we were able to measure the effects of compacting pressure, pressing temperature, material moisture and fraction largeness on quality of briquettes (see Figure 1).

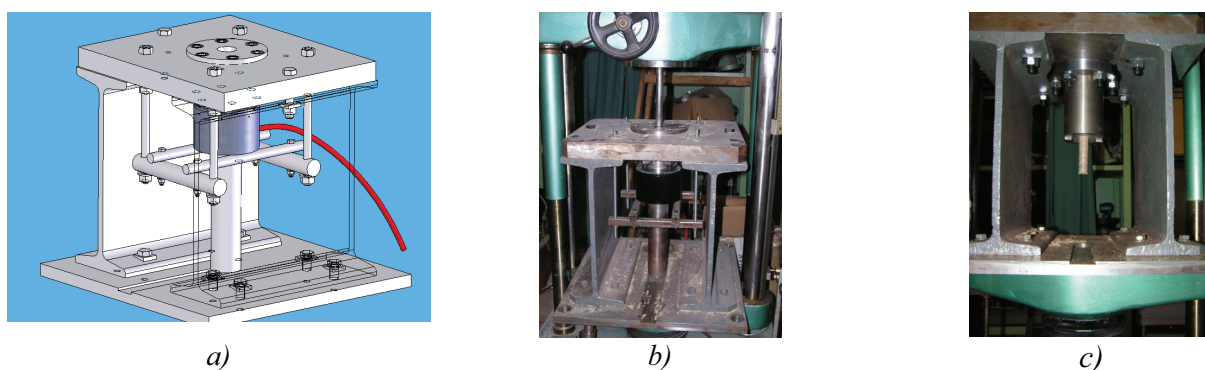


Figure 1 Experimental pressing stand [4]
a) 3D model; b) produced pressing stand; c. pressing stage of a briquette

In the first step we choose the type of pressed material as pine sawdust. We had to make measurements only with one type of material because every type of material has own material properties and nature. We designed also levels of measured parameters (see Table 1). These levels come from our analyses, further experience and pressing stand possibilities. In the Table 2 (below) you can see the design of experiment with calculated average values of briquettes density.

Table 1 Levels of measured parameters [1], [2]

Pressure p (MPa)	Temperature T (°C)	Largeness L (mm)	Moisture w_r (%)
95 - 159	85 - 115	1 - 4	8 - 12

Table 2 Design of experiment with calculated average values of briquettes density [1]

i	p (MPa)	T (°C)	w (%)	L (mm)	$\bar{\rho}_j$ (kg.dm ⁻³)	$s^2(\rho)$ (kg.dm ⁻³)
1	95	85	8	1	1,13919	0,001258
2	159	85	8	1	1,156558	0,001384
3	95	115	8	4	1,167468	0,002513
4	159	115	8	4	1,200171	0,001332
5	95	85	12	1	0,79969	0,00727
6	159	85	12	1	1,007282	0,009572
7	95	115	12	4	1,127726	0,007584
8	159	115	12	4	1,135111	0,006393
9	95	85	8	4	1,088649	0,001384
10	159	85	8	4	1,081203	0,002974
11	95	115	8	1	1,190808	0,003207
12	159	115	8	1	1,235596	0,001062
13	95	85	12	4	0,699423	0,006349
14	159	85	12	4	0,923658	0,006146
15	95	115	12	1	1,173846	0,004268
16	159	115	12	1	1,236333	0,001377

We realized experiment by form of full factorial experiment 2^4 according to Table 1. Goal of the experiment was to follow up the briquettes quality in dependence with pressing temperature, compacting pressure, fraction largeness and material moisture. Briquettes quality was evaluated by briquettes density as is given in EU Standards about solid biofuels. According to EU Standards briquette have very good quality if density is from 1 to 1,4 kg/dm³. In every setting according to Table 2 we pressed 7 briquettes. We measured briquette's dimension, length and weight. These measured values were the base for density calculation. Briquettes density values were processed by various mathematical and statistical methods (e.g. Bartlett's Test, ANOVA, etc...) with the help of software Stathgraphic S Plus. In the following Table 3 you can see result from ANOVA test which say us that how significant is in this process temperature, moisture, pressure and fraction.

Table 3 Result from ANOVA test [1]

Variability source	Square Sum	Degrees of freedom	Average square	Ratio F	Value P	Signification
A – pressure „p“	0,0210523	1	0,0210523	3,40	0,0923	3.
B – temperature „T“	0,13857	1	0,13857	22,37	0,0006	1.
C – moisture „w _r “	0,0709649	1	0,0709649	11,46	0,0061	2.
D – fraction „L“	0,0107137	1	0,0107137	1,73	0,2152	4.
Residual	0,0681389	11	0,00619445			
	0,30944	15				

For closely determination of parameters impact and also impact of their mutual interaction we used method of parameters effect [5]. Results you can see on the following Figure 2.

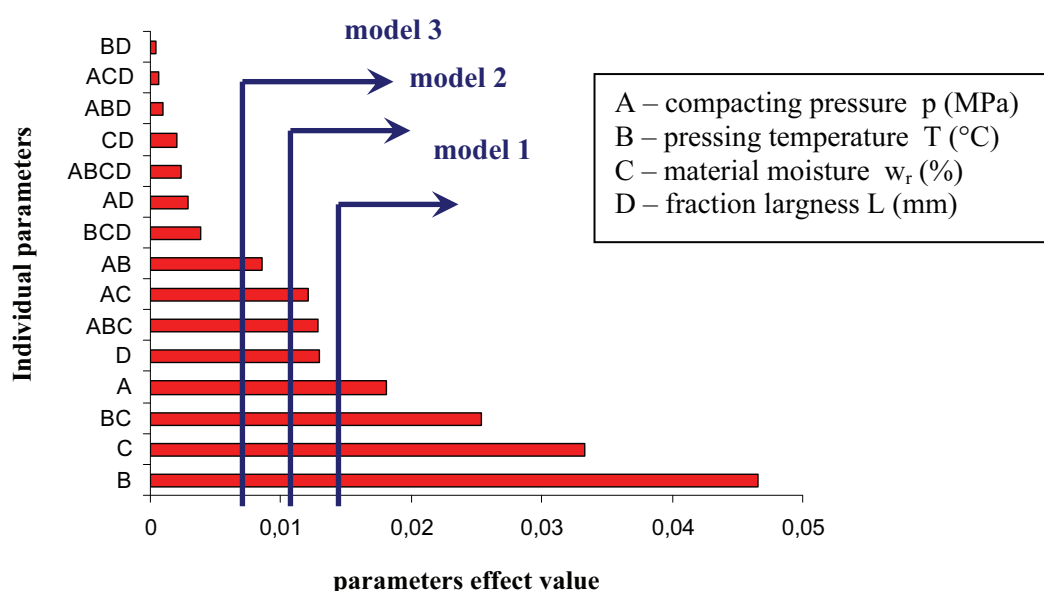


Figure 2 Individual parameters impact – Pareto's effects diagram [1], [2]

Method of parameters effect determined that the biggest impact have pressing temperature and material moisture. Also their interaction has very significant impact on final briquette density. Now we could design the model which will describe the influencing of examined parameters. But which model is the correct? Model with which parameters and interaction is the model which most exact

describe the examined area? On the figure 2 you can see also example how was executed selection of variables for mathematical model. For this selection we used three widely known criterions: index of multilaunching determination (R^2), AIC criterion and Root mean squared error criterion (RMSE). For this selection we used software JMP 8, which comes from SAS software.

Table 4 The best models calculated by JMP 8 according to criterions [1]

Model	Parameters	Number of parameters	R^2	AICc	RMSE
model 1	B,C,BC,A	4	87,9%	-30,1238	0,058448
model 2	B,C,BC,A,D,ABC,AC	7	96,8%	-24,8346	0,035139
model 3	B,C,BC,A,D,ABC,AC,AB	8	99,3%	-32,7996	0,017764
model 4	C,AB,AC,BC,ABCD	5	94,3%	-35,5702	0,041982
model 5	B,C,AB,AC,BC, ABCD	6	96,1%	-33,239	0,036413
model 6	B, C, BC, A, D	5	91,3%	-28,8272	0,05183

The designed mathematical model has to describe the compacting process and therefore we choose model that including all of parameters (pressure, temperature, moisture and fraction). It is very important that model contains all 4 parameters. The calculations showed that also very significant are mutual interactions. Selection of final model design went over from objective examination and comparison of individual values of criterions. Therefore we choose “model 3”.

The next step was calculation or estimation of regression parameters values. For this estimation we used also software JMP 8 as well as for the calculation and design the final form of the mathematical model (see the following formula).

$$\rho = e^{\left(\begin{array}{l} 4,98371 - 0,0261781.p - 0,0410292.T - 0,620594.w_r - 0,015446.L + 0,000228845.p.T + \\ 0,0031851.p.w_r + 0,00528717.T.w_r - 0,0000273004.p.T.w_r \end{array} \right)} \quad (\text{kg.dm}^{-3}) \quad (1)$$

This designed mathematical model is displayed already with estimated regression parameters values. Model is valid only for compacting of pine sawdust and only in levels of parameters according to Table 1. With this model we obtained tool for effective and quickly prediction of final briquettes density values, pressing temperature values, compacting pressure values, material moisture values and fraction largeness values. Following dependencies were created by designed mathematical model of single axis pressing of pine sawdust.

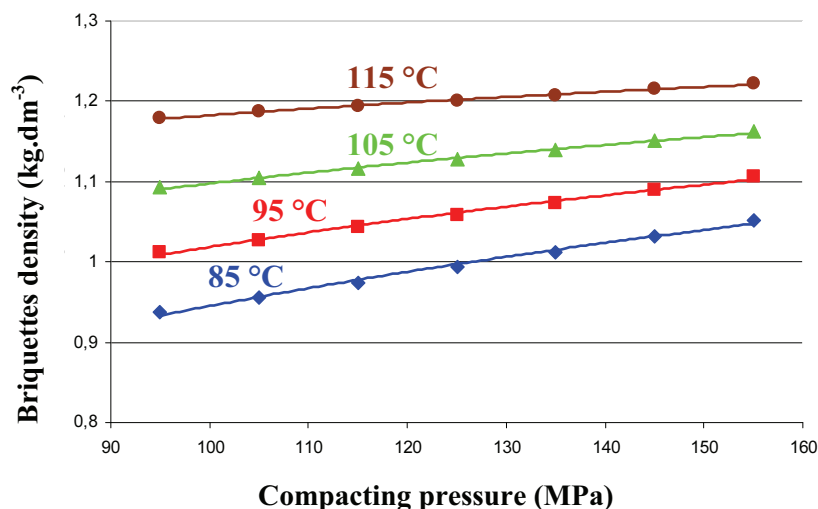


Figure 3 Dependence of briquettes density from compacting pressure by various pressing temperatures for pine sawdust ($w_r = 10\%$; $L = 2\text{ mm}$) [1]

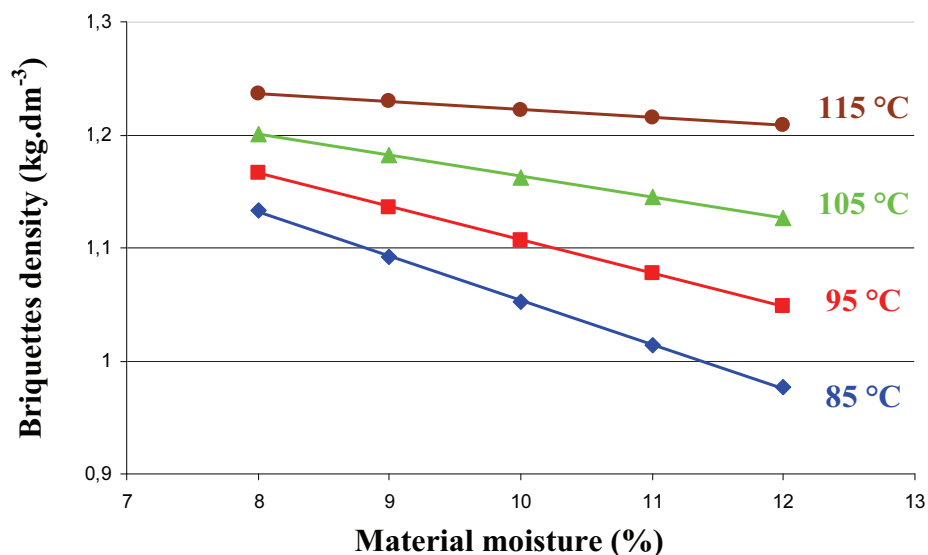


Figure 4 Dependence of briquettes density from material moisture by various pressing temperatures for pine sawdust ($p = 155\text{ MPa}$; $L = 2\text{ mm}$) [1]

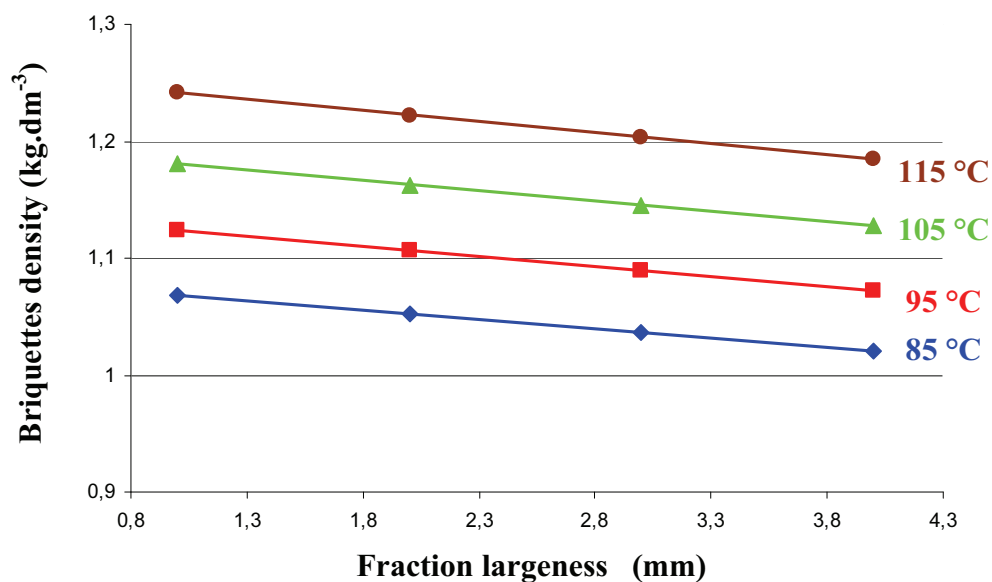


Figure 5 Dependence of briquettes density from fraction largeness by various pressing temperatures for pine sawdust ($w_r = 10\%$; $p = 155\text{ MPa}$) [1]

Dependencies showed on Figures 3, 4 and 5 proved that monitored parameters have significant impact on final briquettes quality at compacting process. On Figure 3 you can see that briquettes density is increasing with increase of pressing temperature and compacting pressure. By lower temperatures we need higher compacting pressure for pressing briquette with same quality, and vice

versa. But the higher temperature is better than higher pressure from lignin plastification point of view. Therefore we recommend for engineers of compacting machines using heating equipment by compacting machine construction.

On Figure 4 you can see that briquettes density is increasing with increase of pressing temperature and decrease of material moisture. Higher temperature and lower moisture is better also for lignin plastification and its binding with cellular structures of material at compacting.

If we decrease the fraction largeness we can help with binding. Smaller particles are able to bind more strongly. This is showed on Figure 5. Briquettes density is increasing with increase of pressing temperature and decrease of fraction largeness. With these parameters we influence not only briquettes density, but also briquettes stability, hardness and absorption of atmospheric moisture. These are very important by transport, storage and handling with briquettes. Following Figure 6 you can see samples – pine briquettes. It is very clearly displayed that with decreasing of fraction largeness is increasing visible briquettes quality.

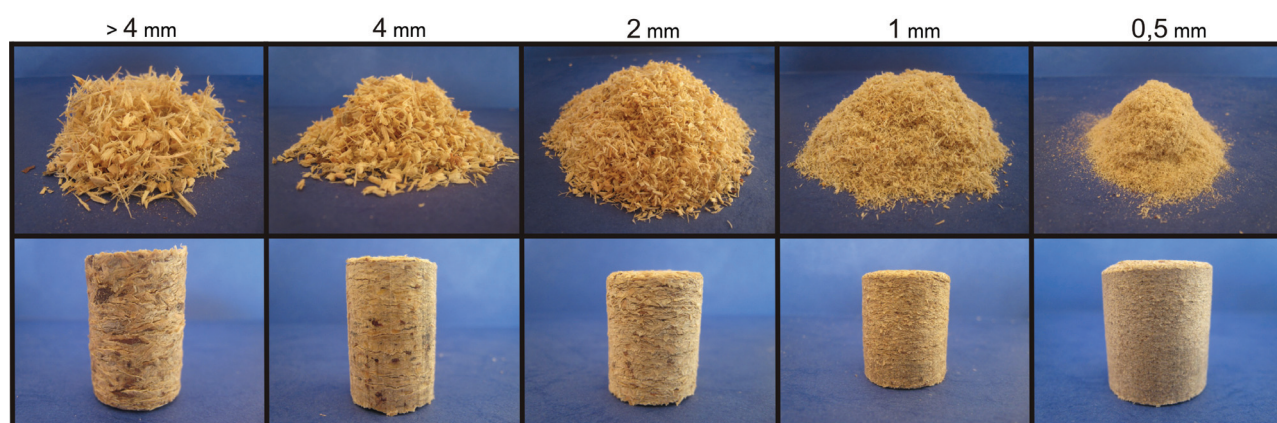


Figure 6 Samples – pine briquettes showed changing briquettes quality by decreasing of fraction largeness

4 Conclusion

The main aim of the experiment was to detect and identify the effect rate of monitored parameters on the final briquettes quality evaluated by briquettes density. By the individual steps we discovered that the most significant effect on briquettes quality has pressing temperature and then material moisture and mutual interaction of these two parameters. The results our hypothesis that compacting pressure, which may seem to be a parameter having the biggest effect on the final briquettes quality, is minor in analyze of effects on briquettes quality. With usage of mathematical and statistical tools we were able to design mathematical model of single axis pressing of pine sawdust. Our following steps will head for another experiment and we would like to expand the interval of validity of every monitored parameter. Then we can repeat the experiment with other materials. Also the dimensional analysis will be the one of our aims for future.

Acknowledgement

„This contribution was created by realization of project „Development of progressive biomass compacting technology and production of prototype and high-productive tools“(ITMS Project code: 26240220017), on

base of Operational Programme Research and Development support financing by European Regional Development Fund.”

References

- [1.] KRIŽAN, P.: *Process of wood waste pressing and conception of presses construction*, Dissertation work, FME SUT in Bratislava, IMSETQM, Bratislava, July 2009, p.150, (in Slovak)
- [2.] SVÁTEK, M.; KRIŽAN, P.; ŠOOŠ, L.; KUREKOVÁ, E.: *Evaluation of parameter impact on final briquettes quality*, In: Measurement 2009: Proceedings. 7th International Conference on Measurement. Smolenice, Slovak Republic, 20.-23.5.2009. - Bratislava: Institute of Measurement Science Slovak Academy of Sciences, 2009. - ISBN 978-80-969672-1-6. - pp. 219-222
- [3.] HOLM, J.; HENRIKSEN, U.; HUSTAD, J.; SORENSEN, L.: *Toward an understanding of controlling parameters in softwood and hardwood pellets production*. Published on web 09/09/2006, American Chemical Society
- [4.] KRIŽAN, P.; VUKELIČ, Dj.: *Properties of some types of materials at pressing*, In: TOP 2009. Proceedings. 15th International Conference on Engineering of Environment Protection, Častá-Papiernička, Slovak Republic, 17.-19.06.2009, FME, SUT in Bratislava, ISBN 978-80-227-3096-9, pp. 289-296 (in Slovak)
- [5.] JAROŠOVÁ, E.: *Process variance models for robust design*. Journal of Applied Mathematics, 1 (1): 251-259, 2008.

Current address

Križan Peter, MSc.

SUT, Faculty of Mechanical Engineering in Bratislava, +421257296537,
e-mail: peter.krizan@stuba.sk

Šooš Ľubomír, Prof. MSc., PhD.

SUT, Faculty of Mechanical Engineering in Bratislava,
e-mail: lubomir.soos@stuba.sk

Matúš Miloš, MSc.

SUT, Faculty of Mechanical Engineering in Bratislava,
e-mail: milos.matus@stuba.sk

Svátek Michal, MSc.

SUT, Faculty of Mechanical Engineering in Bratislava,
e-mail: Michal.Svatek@Officedepot.com

Vukelić Djordje, MSc.

University of Novi Sad, Faculty of Technical Sciences,
e-mail: vukelic@uns.ac.rs

K -DOMINATION SETS ON DOUBLE LINEAR HEXAGONAL CHAINS

MAJSTOROVIĆ Snježana, (HR)

Abstract. A hexagonal chain is a catacondensed hexagonal system in which every hexagon is adjacent to at most two hexagons. Double linear hexagonal chain is consisted of 2 condensed linear hexagonal chains. For any graph G by $V(G)$ and $E(G)$ we denote the vertex-set and the edge-set of G , respectively. For graph G subset D of the vertex-set of G is called k -dominating set, $k \geq 1$, if for every vertex $v \in V(G) \setminus D$, there exists at least one vertex $w \in D$, such that $d(v, w) \leq k$. The k -domination number $\gamma_k(G)$ is the cardinality of the smallest k -dominating set. The 1-domination set (number) is also called domination set (number). In this paper I determine minimal k -dominating sets for double linear hexagonal chain B_{2h} of length h and give exact results for its k -domination number.

Key words and phrases. k -dominating set, k -domination number, double linear hexagonal chain.

Mathematics Subject Classification. 05 C69, 92E10.

1 Introduction

Hexagonal systems are geometric objects obtained by arranging mutually congruent regular hexagons in the plane. They are of considerable importance in theoretical chemistry because they are natural graph representation of benzenoid hydrocarbons [15]. Each vertex in hexagonal system is either of degree two or of degree three. Vertex shared by three hexagons is called an internal vertex of the respective hexagonal system. We call hexagonal system catacondensed if it does not possess internal vertices, otherwise we call it pericondensed.

A hexagonal chain is a catacondensed hexagonal system in which every hexagon is adjacent to at most two hexagons. Linear hexagonal chain is hexagonal chain which is a graph representation of linear polyacene.

Double hexagonal chain is a chain consisted of 2 condensed identical hexagonal chains. It can be considered as benzenoid constructed by successive fusions of successive naphthalenes along a zig-zag sequence of triples of edges as appear on opposite sides of each naphthalene unit. Double linear hexagonal chain is consisted of 2 condensed linear hexagonal chains. Such chain will be denoted with B_{2h} , where h is the number of hexagons in corresponding linear hexagonal chain. See Figure 1.

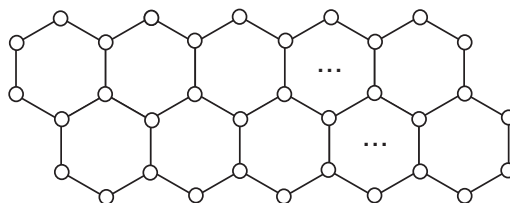


Figure 1: Double linear hexagonal chain of finite length.

Since chemical structures are conveniently represented by graphs, where atoms correspond to vertices and chemical bounds correspond to edges, many physical and chemical properties of molecules are well correlated with graph theoretical invariants [14]. Ren and Zhang [12] studied k -matchings and k -independent sets on double hexagonal chains and obtained some extremal results. Hosoya index and MerrifieldSimmons index were also investigated [11].

Very important theoretical invariant is k -domination number [3]. For any graph G by $V(G)$ and $E(G)$ we denote the vertex-set and the edge-set of G , respectively. For graph G subset D of the vertex-set of G is called k -dominating set, $k \geq 1$, if for every vertex $v \in V(G) \setminus D$, there exists at least one vertex $w \in D$, such that $d(v, w) \leq k$. The k -domination number $\gamma_k(G)$ is the cardinality of the smallest k -dominating set. A set S perfectly k -dominates G if for each vertex $v \in G$ there is exactly one vertex $u \in S$, such that $d(u, v) \leq k$. The concept of a k -dominating set in a graph was introduced in 1975 in a paper by Meir and Moon [10]. They studied k -packing and k -covering (k -dominating) sets, and established many relations between the k -packing number and the k -domination number of a tree. Later, in 1976, Slater [13] dealt with graph application in communication networks, and he considered a problem of finding a minimum k -dominating set in a graph, called it k -basis of a graph.

The concept of distance domination in graphs has many applications. For example, if we consider a graph G associated with the road grid of a city where the vertices of G correspond to the street intersections and where two vertices are adjacent if and only if the corresponding street intersections are a block apart, then we may use a minimum k -dominating set in G to locate a minimum number of facilities (such as police stations, hospitals, banks) so that every intersection is within k city blocks of a facility. Other examples can be found in [7] and [8]. One of present authors studied the problem of determining k -dominating number on Cartesian products of two paths [5], and on linear benzenoids and infinite hexagonal grid [6]. Other varieties of domination, such as total domination, were investigated on linear and double linear hexagonal chains. [17, 9]

For two distinct vertices u and v in G , the distance $d(u, v)$ between u and v is the length of a shortest path between u and v .

The open k -neighborhood $N_k(v)$ of $v \in V(G)$ is the set of vertices in $V(G) \setminus \{v\}$ at distance

at most k from v .

In this paper I determine minimal k -domination sets on $B_{2,h}$ and calculate exact values for $\gamma_k(B_{2,h})$.

Before I give main results, I will need the following propositions:

Proposition 1 Let P_n be a path and C_n be a cycle with n vertices. Then

$$\gamma_k(P_n) = \gamma_k(C_n) = \left\lceil \frac{n}{2k+1} \right\rceil.$$

Proposition 2 For $k \geq 1$, let \mathcal{D} be a k -dominating set of a graph G . Then \mathcal{D} is a minimal k -dominating set of G if and only if each $d \in \mathcal{D}$ has at least one of the following properties:

- (i) There exists a vertex $v \in V(G) \setminus \mathcal{D}$ such that $N_k(v) \cap \mathcal{D} = \{d\}$;
- (ii) The vertex d is at distance at least $k+1$ from every other vertex of \mathcal{D} in G .

2 Domination number on double hexagonal chain

Domination numbers for isomorphic graphs are equal. Therefore, we shall represent $B_{2,h}$ with the following figure and introduce the following coordinates:

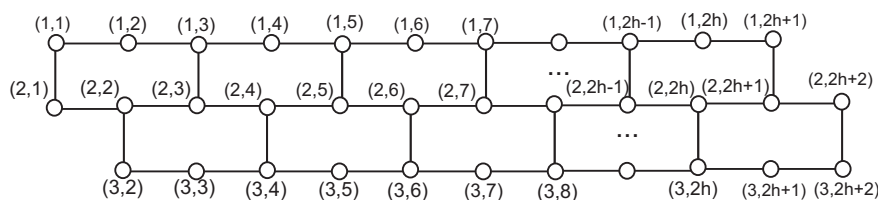


Figure 2: Coordinate system for double hexagonal chain of length h .

Theorem 1 Let B_{2h} be a double hexagonal chain. Then

$$\gamma(B_{2h}) = \begin{cases} 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 2, & h = 0, 1(\text{mod } 5) \\ 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 3, & h = 2, 3(\text{mod } 5) \\ 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 4, & h = 4(\text{mod } 5). \end{cases}$$

Proof:

For $h = 1$ we have dominating set $T_1 = \{(1, 1), (2, 3), (3, 3)\}$.

If $h = 2$ then $T_2 = T_1 \cup \{(1, 5), (3, 6)\}$ is dominating set for $B_{2,2}$.

For $h = 3$ we have $T_3 = \{(1, 2), (1, 5), (2, 2), (2, 7), (3, 4), (3, 7)\}$, while for $h = 4$ the dominating set is $T_4 = T_3 \cup \{(1, 9), (3, 10)\}$.

Finally, for $h = 5$ we have

$$T_5 = \{(1, 1), (1, 4), (1, 7), (1, 11), (2, 4), (2, 9), (3, 2), (3, 6), (3, 9), (3, 12)\}.$$

Sets T_i , $i = 1, \dots, 5$ are the smallest among all dominating sets for considered chains, since each vertex from B_{2h} , $h = 1, \dots, 5$, is dominated with exactly one vertex from T_i , where $i = h$.

In the sequel we proof theorem for general cases for h .

For $h = 1(\text{mod } 5)$ we consider the set $D = T_1 \cup D_1$, where

$$D_1 = \{(1, 5 + 10p), (1, 8 + 10p), (1, 11 + 10p), (2, 8 + 10p), (2, 13 + 10p), (3, 6 + 10p), \\ (3, 10 + 10p), (3, 13 + 10p) : p = 0, 1, \dots, \left\lfloor \frac{h}{5} \right\rfloor - 1\}.$$

D is dominating set for B_{2h} so $\gamma(B_{2h}) \leq |D| = 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 2$. See Figure 3.

Proof of minimality is obvious from Figure 3 because each vertex in B_{2h} is dominated with D exactly once. We conclude $\gamma(B_{2h}) = |D| = 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 2$.

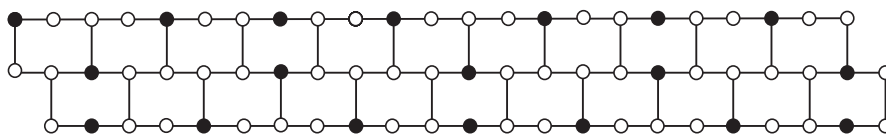


Figure 3: Dominating vertices for $B_{2,11}$.

For $h = 2(\text{mod } 5)$ the dominating set is $D = T_2 \cup D_2$, where

$$D_2 = \{(1, 8 + 10p), (1, 11 + 10p), (1, 15 + 10p), (2, 8 + 10p), (2, 13 + 10p), (3, 10 + 10p), \\ (3, 13 + 10p), (3, 16 + 10p) : p = 0, 1, \dots, \left\lfloor \frac{h}{5} \right\rfloor - 1\}.$$

Because each vertex from B_{2h} is dominated with D once, we conclude $\gamma(B_{2h}) = |D| = 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 3$. See Figure 4.

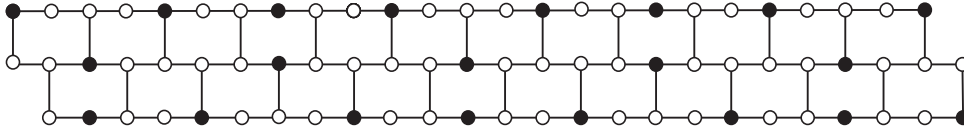


Figure 4: Dominating vertices for $B_{2,12}$.

For $h = 3(mod 5)$ we have dominating set $D = T_3 \cup D_3$, where

$$D_3 = \{(1, 9 + 10p), (1, 12 + 10p), (1, 15 + 10p), (2, 12 + 10p), (2, 17 + 10p), (3, 10 + 10p), (3, 14 + 10p), (3, 17 + 10p) : p = 0, 1, \dots, \left\lfloor \frac{h}{5} \right\rfloor - 1\},$$

$$\text{so } \gamma(B_{2h}) = |D| = 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 3.$$

For $h = 4(mod 5)$ the dominating set is $D = T_4 \cup D_4$, where

$$D_4 = \{(1, 11 + 10p), (1, 15 + 10p), (1, 18 + 10p), (2, 13 + 10p), (2, 18 + 10p), (3, 13 + 10p), (3, 16 + 10p), (3, 20 + 10p) : p = 0, 1, \dots, \left\lfloor \frac{h}{5} \right\rfloor - 1\}.$$

$$\text{We conclude } \gamma(B_{2h}) = |D| = 8 + 8 \left\lfloor \frac{h}{5} \right\rfloor = 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 4.$$

For $h = 0(mod 5)$ we have $D = T_5 \cup D_5$, where

$$D_5 = \{(1, 14 + 10p), (1, 17 + 10p), (1, 21 + 10p), (2, 14 + 10p), (2, 19 + 10p), (3, 16 + 10p), (3, 19 + 10p), (3, 22 + 10p) : p = 0, 1, \dots, \left\lfloor \frac{h}{5} \right\rfloor - 1\}.$$

$$\text{We conclude } \gamma(B_{2h}) = |D| = 10 + 8 \left\lfloor \frac{h}{5} \right\rfloor = 3 \left\lfloor \frac{h}{5} \right\rfloor + h + 5.$$

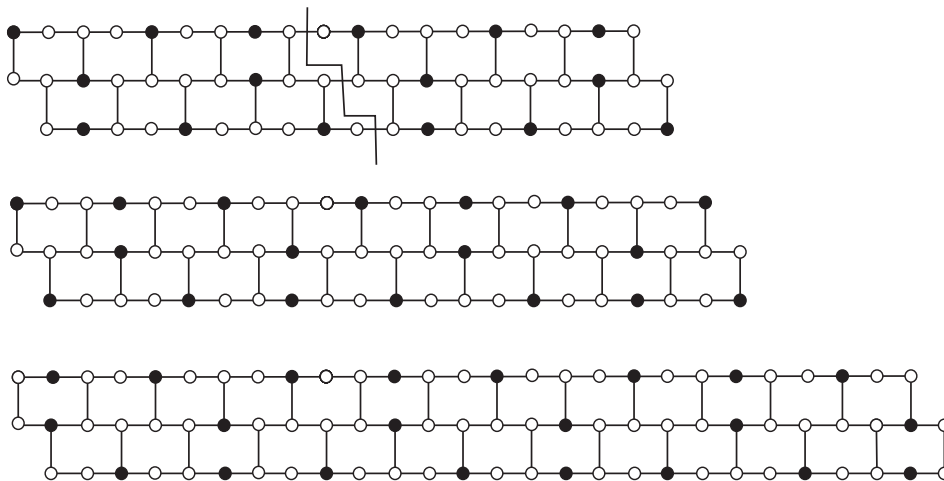


Figure 5: Dominating vertices for $B_{2,9}$, $B_{2,10}$ and $B_{2,13}$. Minimality is obvious.

Remark 1 Minimal dominating sets from Theorem 1 are not unique. They were chosen so that every vertex in B_{2h} is dominated with the corresponding minimal dominating set exactly once, that is, every vertex in B_{2h} is perfectly dominated with D .

Corollary 1 Let D_h be the minimal dominating set for B_{2h} as in Theorem 1. Then for $h = 4(mod 5)$

$$D_{1+5m} \subset D_{2+5m} \subset D_{4+5m}, \quad m = 0, 1, \dots, \left\lfloor \frac{h}{5} \right\rfloor.$$

Corollary 2 For B_{2h} following relationships between domination numbers hold:

$$\begin{aligned} \gamma(B_{2(2+5m)}) &= \gamma(B_{2(1+5m)}) + 2 \\ \gamma(B_{2(3+5m)}) &= \gamma(B_{2(1+5m)}) + 3 \\ \gamma(B_{2(4+5m)}) &= \gamma(B_{2(2+5m)}) + 3 = \gamma(B_{2(1+5m)}) + 5 \\ \gamma(B_{2(5+5m)}) &= \gamma(B_{2(4+5m)}) + 2 = \gamma(B_{2(2+5m)}) + 5 \end{aligned}$$

where $m \in \mathbb{N}_0$.

3 2-domination on double hexagonal chains

Theorem 2 Let B_{2h} be a double hexagonal chain. Then

$$\gamma_2(B_{2h}) = h + 1.$$

Proof:

For h odd, we consider the set

$$D = \left\{ (1, 1 + 4p), (3, 4 + 4p) : p = 0, 1, \dots, \left\lfloor \frac{h}{2} \right\rfloor \right\}.$$

Set D is one 2-dominating set for $B_{2,h}$.

If h even, then we have 2-dominating set

$$D' = D \setminus \{(3, 4 + 2h)\}.$$

In both cases the cardinality of 2-dominating set is equal to $h + 1$, so it follows that $\gamma_2(B_{2h}) \leq h + 1$.

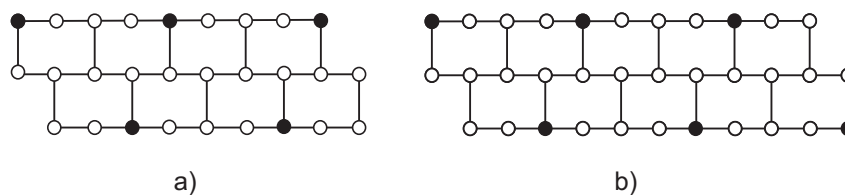


Figure 7: Dominating vertices for a) $B_{2,4}$ and b) $B_{2,5}$.

Proof of minimality. For h even or h odd, the distance between every 2-dominating vertices is at least 3.

We have

$$\begin{aligned} d((1, 1 + 4p), (1, 1 + 4(p + 1))) &= d((3, 4 + 4p), (3, 4 + 4(p + 1))) = 4, \\ d((1, 1 + 4p), (3, 4 + 4p)) &= 5, \quad d((3, 4 + 4p), (1, 1 + 4(p + 1))) = 3. \end{aligned}$$

We conclude that minimality of 2-dominating sets D and D' follows directly from Proposition 2 (2). Therefore, $\gamma_2(B_{2h}) = h + 1$.

Remark 1 Sets D and D' satisfy condition (1) of Proposition 2 as well.

4 $(K \geq 3)$ -domination on double hexagonal chains

Theorem 3 Let B_{2h} be a double hexagonal chain and let $k \geq 3$. Then

$$\gamma_k(B_{2h}) = \left\lceil \frac{2h + 2}{2k - 1} \right\rceil.$$

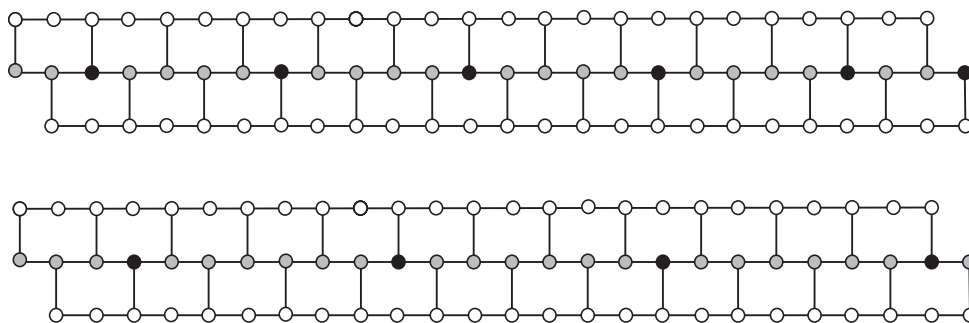
Proof: Let $t = \left\lceil \frac{2h + 2}{2k - 1} \right\rceil$. We consider the set

$$D = \{(2, (2k - 1)i + k) : i = 0, 1, \dots, t - 1\}.$$

If $(2k - 1)t - k + 1 \leq 2h + 2$, then the set D is one k -dominating set for B_{2h} . Otherwise, the k -dominating set for B_{2h} is

$$D' = (D \setminus \{(2, (2k - 1)t - k + 1)\}) \cup \{(2, 2h + 2)\}.$$

It follows that $\gamma_3(B_{2h}) \leq |D| = |D'| = t$.


Figure 8: K -dominating vertices for $B_{2,12}$ and $k = 3, 4$.

Proof of minimality: The distance between every two k -dominating vertices in D is equal to $2k - 1$, and in D' this is true for vertices $(2, (2k - 1)i + k)$, where $i = 0, 1, \dots, t - 2$. For the distance between vertex $(2, 2h + 2)$ and vertex for which $i = t - 2$ we have

$$d((2, (2k - 1)t - 3k + 2), (2, 2h + 2)) > k - 1.$$

For every k -dominating vertex $(2, (2k - 1)i + k)$ from D , we choose vertex $(2, (2k - 1)i + k - 1) \in B_{2h} \setminus D$. We have

$$d((2, (2k - 1)i + k - 1), (2, (2k - 1)i + k)) = 1$$

$$d((2, (2k - 1)i + k - 1), w) \geq 2k, \quad w \neq (2, (2k - 1)i + k).$$

We conclude that

$$N_k((2, (2k - 1)i + k - 1)) \cap D = \{(2, (2k - 1)i + k)\}, \quad i = 0, 1, \dots, t - 1.$$

For D' and $i = 0, 1, \dots, t - 2$ we choose the same vertices as for D , and for $(2, 2h + 2)$ we choose vertex $(3, 2h + 2)$, since

$$d((2, 2h + 2), (3, 2h + 2)) = 1 \quad \text{and} \quad d((3, 2h + 2), (2, (2k - 1)t - 3k + 2)) \geq k + 1.$$

It follows that

$$N_k((3, 2h + 2)) \cap D' = \{(2, 2h + 2)\}.$$

From the above results we conclude that for every $w \in D$ ($w' \in D'$) there exists $v \in B_{2h} \setminus D$ ($v' \in B_{2h} \setminus D'$) so that $N_k(v) \cap D = \{w\}$ ($N_k(v') \cap D' = \{w'\}$). Now, minimality of k -dominating sets D and D' follows directly from Proposition 2 (1).

Corollary 3 Let B_{2h} be a double hexagonal chain of length h and let $k \geq 3$. Then

$$\gamma_k(B_{2h}) = \gamma_{k-1}(P_{2h+2}).$$

Proof. Follows directly from Theorem 3. In Figure 8 all vertices of considered path are represented with grey circles together with black circles as k -dominating vertices.

References

- [1] M. EL-ZAHAR, C.M. PAREEK, Domination number of products of graphs, *Ars Combin.* 31 (1991) 223–227.
- [2] R.J FAUDREE, R.H. SCHELP, The domination number for the product of graphs, *Congr. Numer.* 79 (1990) 29–33.
- [3] T.W. HAYNES, S.T. HEDETNIEMI, P.J. SLATER, Fundamentals of domination in graphs, *Marcel Dekker Inc.* New York (1998)
- [4] C.F.A. DE JAENISCH, Traité des Applications de l'Analyse Mathématique au Jeu des Échecs, *St. Pétersbourg*, 1862.

- [5] A. KLOBUČAR, Domination numbers of cardinal products, *Math. Slovaca* 49 (1999) 241–250.
- [6] A. KLOBUČAR, Domination numbers of cardinal products $P_6 \times P_n$, *Math. Communications* 4 (1999) 241–250.
- [7] D. LICHTENSTEIN, *Planar satisfiability and its uses*, SIAM J. Comput. 11 (1982) 329–343.
- [8] B.C. TANSEL, R.L. FRANCIS, T.J. LOWE, *Location on networks: a survey. I. The p-center and p-median problems*, Management Sci. 29 (1983) 282–297.
- [9] S. MAJSTOROVIĆ, A. KLOBUČAR, Upper bound for total domination number on linear and double hexagonal chain, *Int. J. Chem. Model.* 2009.
- [10] A. MEIR, J.W. MOON, *Relations between packing and covering numbers of a tree*, Pacific J. Math. 61 (1975) 225–233.
- [11] H. REN, F. ZHANG, Double hexagonal chains with maximal Hosoya index and minimal Merrifield-Simmons index, *Journal of Mathematical Chemistry* 42 (2007)
- [12] H. REN, F. ZHANG, Extremal double hexagonal chains with respect to k-matchings and k-independent sets *Discrete Applied Mathematics* 155 (2007)
- [13] P.J. SLATER, *R-domination in graphs*, J. Assoc. Comput. Mach. 23 (1976) 446–450.
- [14] N. TRINAJSTIĆ, Chemical Graph Theory, *CRC Press, Boca Raton*, (1983), 2nd revised ed., 1992
- [15] H. WIENER, Structural Determination of Paraffin Boiling Points, *J. Amer. Chem. Soc.* 69 (1947) 17–20
- [16] V.G. VIZING, The Cartesian product of graphs, *Vychisl. Sistemy* 9 (1963) 30–43.
- [17] D. VUKIČEVIĆ, A. KLOBUČAR, *K*-dominating sets on Linear Benzenoids and on the Infinite Hexagonal Grid, *Croatica Chemica Acta* 2007.

Current address

Snježana Majstorović

Department of Mathematics, University of Osijek,
Trg Ljudevita Gaja 6, HR-31000 Osijek, Croatia,
e-mail: smajstor@mathos.hr

INFLUENCE OF STRUCTURAL PARAMETERS IN COMPACTING PROCESS ON QUALITY OF BIOMASS PRESSINGS

MATÚŠ Miloš, (SK), KRIŽAN Peter, (SK)

Abstract: The contribution deals with the compacting of material, mainly biomass material. The biomass compacting is very complicated process, because the biomass is “live” material and there are many factors influencing to this process. Every kind of material has different contributions to achieve a high-grade pressing. The contribution is focused on influencing parameters which is possible to control by structural changes on tools of production machine. There is also described their influence on the quality of pressings

Key words: biomass, biofuel, pressing, compacting, briquetting, pelleting, quality of pressing

1 Quality of biomass pressing

The quality of pressing made of biomass is defined by standards. On the present time there is no united European standard for solid biofuel. Therefore the European countries have used own national standards for biomass pressing quality definition. The most important and completed standards are comparison in table 1. The mentioned standards are important in Euro-region because owning countries are the biggest producers of solid biofuel and the biggest consumers at the same time. That means if a producer from other country want to export solid biofuel to mentioned countries, the solid biofuel has to be up to national standard of imported country.

Mentioned standards determining quality parameters of biomass pressings are strict and it is possible to achieve their limits mostly just by compacting pure wood material. Quality indexes of biomass pressings are possible to divide on thermo-chemical indexes and physical-mechanical indexes. Thermo-chemical indexes define content of particular chemical elements in pressings, ash content, humidity and heat value. Physical-mechanical indexes define geometrical parameters of pressings, pressings density, abrasion and their toughness.

Thermo-chemical quality indexes result from properties of compacted material. It is possible to decrease the humidity for quality improvement, but other parameters are bound to compacted material (chemical elements bounding to biomass, ash content, heat value, etc.). Properties of

compacted material (mainly the material toughness) have also the influence on physical-mechanical quality indexes, but these indexes are possible to manipulate by used technology. The most important physical-mechanical quality indexes, which we can positive control by suitable choice of technology, is density and abrasion. Minimum value of these two indexes is defined by mention standards.

Table 1: Compare some pressing parameters of European countries standards

PARAMETER	DIN 51 731 	Ö-Norm M 7135 	Certification DINplus 	SS 18 71 20 
Diameter	Od 4 do 10 mm	Od 4do 10 mm	Not defined	< 25 mm
Length	< 50 mm	< 5 x d	< 5 x d	< 5 x d
Density	> 1,0-1,4 kg/dm ³	> 1,12 kg/dm ³	> 1,12 kg/dm ³	Not defined
Humidity	< 12 %	< 10%	< 10%	< 10 %
Powder density	Not defined	Not defined	Not defined	> 500 kg/m ³
Abrasion	Not defined	< 2,3 %	< 2,3 %	Not defined
Ash content	< 1,5 %	< 0,5 %	< 0,5 %	< 1,5 %
Heat value	17,5 - 19,5 MJ/kg	> 18 MJ / kg	> 18 MJ / kg	> 16,9 MJ / kg
Sulphur content	< 0,08 %	< 0,04 %	< 0,04 %	< 0,08 %
Nitrogen content	< 0,3 %	< 0 ,3 %	< 0 ,3 %	Not defined
Chlorine content	< 0,03 %	< 0,02 %	< 0,02 %	< 0,03%
Arsenic content	< 0,8 mg / kg	Not defined	< 0,8 mg / kg	Not defined
Lead content	< 10 mg / kg	Not defined	< 10 mg / kg	Not defined
Cadmium content	< 0,5 mg / kg	Not defined	< 0,5 mg / kg	Not defined
Chrome content	< 8 mg / kg	Not defined	< 8 mg / kg	Not defined
Copper content	< 5 mg / kg	Not defined	< 5 mg / kg	Not defined
Silver content	< 0,05 mg / kg	Not defined	< 0,05 mg / kg	Not defined
Zinc content	< 100 mg / kg	Not defined	< 100 mg / kg	Not defined

Density is the most important quality index of pressings. It is important in light of manipulation – pressings have to be compact without rifts and it is not allowed to spall small pieces of material. By raising of pressings density the longer time of burning is achieved, what is their most considerable attribute concerning their primary function as fuel. Higher density also favourably influences long-lasting volume and shape stability as well as decreases pressings ability to absorb the air humidity. Abrasion is in common measured just for pellets, i.e. for the smallest size class. It relates with request to prevent dust particles creation in process of automated manipulation this kind of fuel, eventually to prevent dust particles explosion. Briquettes abrasion measurement is not used frequently, but there are also the standards for its determination. Value of abrasion is depending on the reached pressing density.

2 Structural parameters in compacting process

On base of in-depth study of parameters influencing on final pressing quality we are able to divide these parameters to three classes:

- Material parameters,
- Technological parameters,
- Structural parameters.
- The major structural parameters influencing on pressings quality are:
- Diameter of pressing chamber,
- Length of pressing chamber,
- Material of tools and their roughness (friction factor between tools and compacted material),
- Conicalness of surface in pressing chamber,
- Type of pressing tool,
- Length of cooling canal.

3 Signification and influence of geometry pressing chamber and friction factor

After the finished analysis we can state that there exist only few mathematical models describing the compacting process, describing mainly the influence of structural parameters. Despite of this state, on base of the theory of uniaxial compacting in the closed chamber, we are able to analyse an influence of the pressing chamber length change and an influence of friction factor change between compacted material and the pressing chamber. The friction factor depends on material of the chamber and compacted material (and its state – humidity, temperature, etc.) The diameter of the pressing chamber with its length cooperate has significant influence on pressings properties as fuel but also on tools wear. For slow burn-up of pressings, the ratio of surface to volume has to be minimal. The same ratio is important for pressing tools (die, pressing chamber, piston, worm, roll, etc.) – small ratio causes small tools wear. Therefore it is desirable to search optimal dimensions of the pressing from different aspects. The diameter of the pressing chamber is generally strictly given on base of shape and size requests of final pressings. These requests are determined by external impulses (market, usage, etc.).

4 Pressure conditions in the cylindrical pressing chamber by uniaxial compacting

Pressure conditions in the cylindrical pressing chamber by uniaxial compacting, when a back pressure is caused by a plug, are shown on figure 1. Pressure conditions in the chamber between the piston and the plug are possible to explain on the element of compacted material dx , while its weight is ignored during the compacting process.

Equation of forces balance may be derived from pressure conditions:

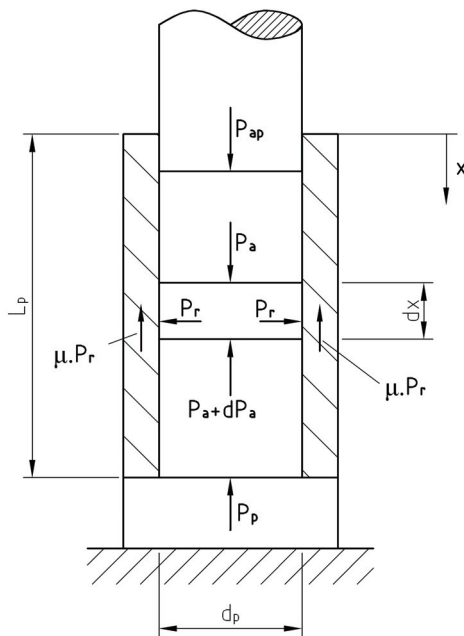
$$[p_a - (p_a + dp_a)] \frac{\pi d^2}{4} - \mu p_r \pi d_p dx = 0 \quad (1)$$

Compacting bulk materials usually goes from pressure anisotropy (higher pressures are in normal direction). The main stress ratio (radial σ_r / axial σ_a) is marked the residual pressure coefficient or horizontal pressure ratio λ . For dispersive materials λ has values from 0 to 1.

$$\lambda = \frac{\sigma_r}{\sigma_a} = \frac{p_r}{p_a} \Rightarrow p_r = \lambda p_a \quad (2)$$

After substitution the equation (2) into the equation (1) and for its boundary conditions $x = 0$ $p_a = p_{ap}$, $x = L$, $p_a = p_p$ we can get final relation:

$$p_{ap} = p_p \cdot e^{\frac{4 \cdot \lambda \cdot \mu \cdot L_p}{d_p}} \quad [\text{Pa}] \quad (3)$$



- p_{ap} – axial pressure of piston [MPa]
- p_p – back pressure in the pressing chamber [MPa]
- p_r – radial pressure [MPa]
- p_a – axial pressure on the plug [MPa]
- d_p – diameter of the pressing chamber [mm]
- μ – friction factor [-]
- L_p – length of the pressing chamber [mm]

Figure 1 Pressure conditions in pressing chamber by uniaxial compaction

This equation (3) tells about relation between piston pressure and back pressure affecting the compacting material. The equation is also possible to use for compacting in open pressing chamber where the back pressure is caused by surface friction drag between the extruded material through pressing chamber and the surface of the chamber. We can also use this equation to determine size of the necessary back pressure and to determine the surface friction drag of extruded column already compacted. We know to provide the necessary back pressure for open chambers by right combination of friction factor and pressing chamber length.

The back pressure affecting the compacting material exponentially grows by increasing the friction factor. The same principle is done after positive change of the pressing chamber length. The back pressure exponentially decreases by decreasing of chamber diameter. The friction factor, the pressing chamber length and its diameter have huge influence on final pressings quality.

5 Influence of surface conicalness in pressing chamber

Extruding material through conicalness chamber use multiaxis compacting and therefore it increases the pressings quality like higher density and higher mechanical properties. But the tools wear is also higher. On this base a mathematical analysis, experimental verification and consequential optimization of geometry is needed.

For analysis of influence of pressing chamber surface conicalness change we use the theory of forward extrusion as a basic technology of metal volume moulding. Force and pressure distributing in this technology is very close to theory of biomass compacting. The theory of forward extrusion and definition of pressure condition in pressing chamber is shown for the simplest example of extrusion where the chamber consists of three parts (figure 2). This type of pressing chamber is often used in the structure of compacting machines.

Cylindrical part is incoming reservoir for the compacting process. Material is filled into this part and then starts to be compacted by pressing piston. In conical (reductive) part the main compacting of material is being done. The pressure and conical chamber cause the multiaxis compacting effect. The holding time while the pressing is under the pressure is necessary for elimination pressing expansion. Calibration part gives final shape to pressing and provides the holding time under the pressure and temperature what is suitable for preclusion of pressing disintegration after leaving the conical part. The size changes of pressing outgoing from the calibration part are very small.

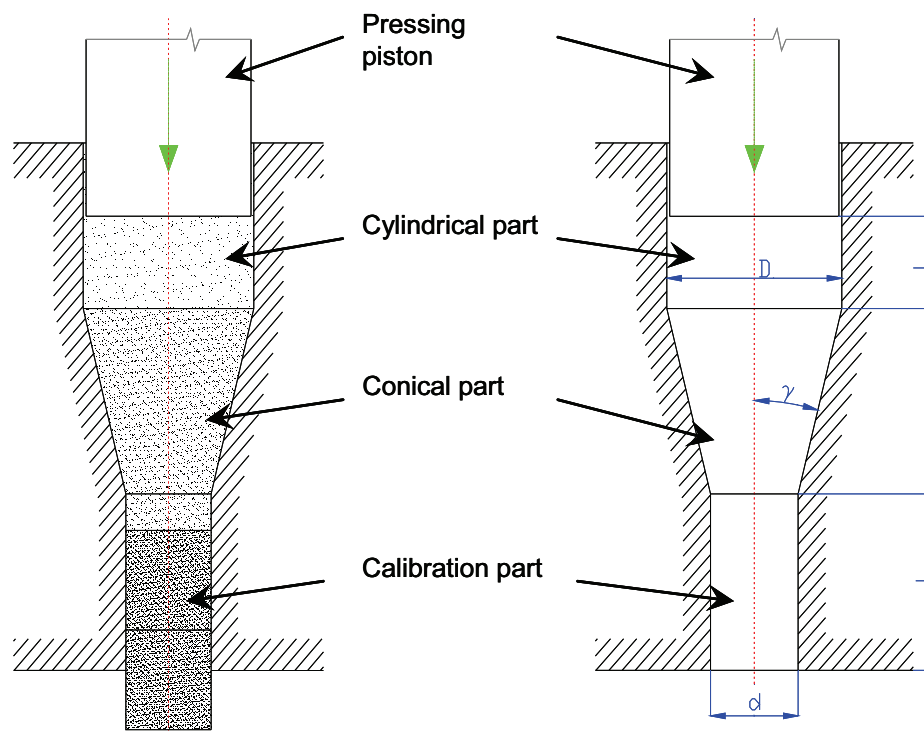


Figure 2 Main parts of conical pressing chamber

The final equation for calculation of compacting pressure is an addition of pressures in particular chamber parts. The equation for needful pressure in calibration part of chamber is

$$p_4 = \sigma_{k4} \frac{4 \cdot f_4 \cdot l}{d} \quad [\text{MPa}] \quad (6)$$

where: f_4 – friction factor for the calibration part [-]
 σ_{k4} – back pressure affecting the compacting material [MPa]
 d – diameter of calibration part [mm]
 l – length of calibration part [mm]

The equation (7) is valid for pressure distribution in conical part of chamber. However this equation is valid for pressing chamber with small angle γ ($\gamma \leq 30^\circ$)

$$p_3 = \sigma_{k3} \left(1 + \frac{f_3 + 0.5}{2\gamma} \right) \cdot \ln \frac{S}{s} + p_4 \quad [\text{MPa}] \quad (7)$$

where: f_4 – friction factor for the calibration part [-]
 σ_{k3} – back pressure affecting the compacting material [MPa]
 γ – conicalness of surface [$^\circ$]
 S – cross-section area of cylindrical part [mm^2]
 s – cross-section area of calibration part [mm^2]

The pressure on the pressing piston is

$$p_2 = p_3 + \sigma_{k3} \cdot \frac{2L}{D} \quad [\text{MPa}] \quad (8)$$

where: L – distance between the piston and the bottom end of the pressing chamber [mm]
 D – diameter of cylindrical part of pressing chamber [mm]

The final equation (9) for compacting pressure in pressing chamber on figure 2 we get by substitution the equations (6) and (7) to the equation (8)

$$p_3 = \sigma_{k3} \left(1 + \frac{f_3 + 0.5}{2\gamma} \right) \cdot \ln \frac{S}{s} + \sigma_{k3} \frac{2L}{D} + \sigma_{k4} \frac{4f_4 \cdot l}{d} \quad [\text{MPa}] \quad (9)$$

This simple process allows designing optimal shape of pressing chamber. It also makes the possibility to calculate necessary compacting pressure in the compacting process. It is very important to deal with pressing chamber optimization, because the shape of chamber has the significant influence on the final pressing quality. All mentioned structural parameters have very significant influence on the final pressing quality and therefore in the close future it will be necessary to make important experiments in this field for determinate their exact influence. I hope that our theoretical states will be confirmed by result of experiments. Subsequently we want to verify the results on the real production machines.

6 Influence of type of pressing tool on pressings quality

There are two technologies for compacting biomass for energetic purposes – briquetting and pelleting. The principle of pelleting is the same for all types pelleting machines. Material is extruding through the holes in pelleting die by the tool rolling up the die. Because of small dimensions of pellets it is not important deal with pellet inner defects in dependence on the type of pelleting machine. The situation in briquetting is different. Pressings are bigger and their inner defects are more considerable what influence the quality, i.e. toughness and mechanical properties of pressings. We can divide the briquetting technology to three principles. Every principle has own specific tool and different influence on pressing quality.

6.1 Pressing piston of hydraulic press

Biomass on hydraulic press is compacted in closed chamber and whole volume of biomass, what is needed for one briquette creation, is compacted at once – with one stroke of piston. By this principle it is possible to made different shapes of pressings. If the briquette length is growing, the homogeneity in whole briquette (figure 3) volume is getting down and the structure defects are making. The rifts are made in the briquette (figure 4) and mechanical toughness gets worse.



Figure 3 Briquettes made on hydraulic press



Figure 4 Structure of briquette - defects

6.2 Pressing piston of mechanical press

Creation of briquette on mechanical presses with pressing piston happens in closed chamber, where is material compacted and extruded through pressing die by piston. Biomass is compacted to the cylindrical endless briquette, which is cut to suitable length (figure 5). There is created slim sheet of briquette by every stroke of piston. Successive sheets are break through themselves by special shaped end of piston, so there creates joint by shaped contact. Defects (rifts) are made on the border line between sheets (figure 6). It decreases the briquette quality – mainly its mechanical toughness.



Figure 5 Briquettes made by piston on mechanical press



Figure 6 Structure of briquette - defects

6.3 Pressing spiral worm

Biomass briquetting by spiral worm achieves the best quality of briquettes. This principle can create different shapes of briquettes (cylinder, n-angle parallelepiped, with hole or without)(figure 7). Material is compacted continually and structural defects do not create. Material is compacted to the cylindrical endless briquette, which is cut to suitable length. Continual compacting of material insures high grade of mechanical pressing quality indexes (figure 8).



Figure 7 Briquettes made by spiral worm



Figure 8 Structure of briquette – no defects

Every principle of briquetting using other type of pressing tool has advantages and disadvantages. But the pressing quality is not always determining for briquetting principle selection. Producers have to consider also the production costs and capital costs on the unit of compacted material to be competitive on the biofuel market. Therefore we do research work in this field to increase the pressing quality. But the quality must not raise at the expense of costs if biofuel has to be successfully using and gradually gets competitive to fossil fuel.

7 Influence of cooling canal length on pressing quality

Length of cooling chamber has also influence on the final quality of biomass pressings. The basis is coming from the biomass compacting process where this material is compacted with high pressure (cca 120 MPa) and high temperature (90 to 120 °C) and extruded through the pressing chamber. The lignin contained in cells of biomass plastifies, is squeezed out of cells and envelops biomass particulates to create binding material between them. After the pressing extrusion out of chamber the pressing is hot and plastic because the lignin is not solid. It is necessary to create conditions for lignin to cool down under the pressure not to come to pressing destruction. This task is insured by the cooling canal that holds the pressing shape under the low pressure (much lower as

in the pressing chamber) until the pressing gets cold and the lignin gets solid. After that is possible to achieve high-grade pressing. The cooling canal as a separately standing equipment is used in briquetting technology. In pelleting the cooling canal is integrated into pelleting die and creates continuation of hole which is material extruded – behind the pressing part of hole. The size of cross-section area of the cooling canal is given by the size of cross-section area of the pressing – it is little bigger like the cross-section area of the pressing chamber to decrease the friction and possibility of small pressing expansion after extruding from pressing chamber. The length of cooling canal depends on cool down intensity (temperature of the environs, the size of the cross-section area of the pressing).

Conclusion

Compacting is a technology on which influence many factors. The most variables is coming to the compacting process with the compacted material, kind, structure, chemical composition, mechanical conditions, humidity. For every kind of material is needed to search and configure suitable technological parameters for achievement the required pressing quality. There is also necessary to search and configure structural parameters of production machines. The right choice of structural parameters is base condition for creating pressings with high quality. Therefore it needs to deal with the deep analysis of these parameters, experiments and subsequently verification of modification design. It needs to repeat whole research for every kind of material, but the database will be filled what contributes for progression in pressings production not only made of biomass and helps decrease production costs and capital costs.

Acknowledgment

„This publication/contribution was created by realization of project „Development of progressive biomass compacting technology and production of prototype and high-productive tools“ (ITMS Project code: 26240220017), on base of Operational Programme Research and Development support financing by European Regional Development Fund.”

References

- [1] Ö-Norm M 7135:2000 Presslinge aus naturbelassenem Holz oder naturbelassener Rinde. Pellets und Briketts; Anforderungen und Prüfbestimmungen.
- [2] DIN 51731:1996 Prüfung fester Brennstoffe. Presslinge aus naturbelassenem Holz. Anforderung und Prüfung.
- [3] KRIŽAN, P.: Research of structural parameters in compacting process. In.: Proceedings of the Conference Mechanical Engineering 2008, Bratislava, ISBN 978-80-227-2987-1, str. IV – 26.
- [4] HORRIGHS, W. Determining the dimensions of extrusion presses with parallel-wall die channel for the compaction and conveying of bulk solids. In *Aufbereitungs-Technik* : Magazine. Duisburg, 1985, no. 12.
- [5] STOROŽEV, M.V.; POPOV, J.A.: Theory of metal shaping. Alfa Bratislava, SNTL Praha, 1978, 63-560-78, str. 488.

Current address

Miloš MATÚŠ, MSc.

Slovak University of Technology in Bratislava, Faculty of Mechanical Engineering, Institute of Manufacturing Systéme, Environmental Technologies and Quality Management, Nam. Slobody 17, 812 31 Bratislava, Slovak Republic. Phone: +421 257 296 573, e-mail: milos.matus@stuba.sk.

Peter Križan, MSc.

Slovak University of Technology in Bratislava, Faculty of Mechanical Engineering, Institute of Manufacturing Systéme, Environmental Technologies and Quality Management, Nam. Slobody 17, 812 31 Bratislava, Slovak Republic. Phone: +421 257 296 537, e-mail: peter.krizan@stuba.sk.

SOLID SOLUTION HARDENING IN CADMIUM SINGLE CRYSTALS

NAVRÁTIL Vladislav, (CZ), NOVOTNÁ Jiřina, (CZ)

Abstract. Basinski [1] and Butt and Feltham [2] using single crystals based on copper, silver, magnesium and polycrystalline brasses found interesting relation between Critical Resolved Shear Stress (CRSS) and Activation Volume. In the present work similar correlations were obtained for monocrystalline cadmium – zinc alloys in very wide temperature interval (1,5K – 380K). The results are shown to comply with a relatively simple theory of solid – solution hardening [3,4], which was developed further so as to account for the “anomalous” temperature – dependence of the CRSS and of the Activation Area observed below $T \approx \frac{1}{4}T_D$, where T_D is the Debye temperature of the alloy.

Key words. Single crystals, Plastic Deformation, CRSS, Activation Volume, Temperature Anomaly.

Mathematics Subject Classification: Primary 74C99; Secondary 74D05.

1 Introduction

Hardening induced in crystalline materials by dispersed solute atoms is referred to as solid – solution hardening. Theories of alloy hardening invoke the resistance of solute atoms, to the movement of dislocations. Such movement is possible only if the shear stress acting on the dislocations is high enough to overcome this resistance.

Theories of solid – solution hardening can be divided into two groups, according to the type of interaction between dislocations and solute atoms – if the interaction is individual or collective. According to first one, the solute atoms are isolated and are surmounted by the dislocations (Friedel [5], Fleischer [6]). In the second one the dislocation is assumed to overcome less localised stresses due to Groups of obstacles (Mott and Nabarro [7], Feltham [8], Labusch [9], Nabarro [10])

According to [8] thermally activated transition of dislocations between consecutive equilibrium positions was considered to involve break – away of segment – lengths comprising numerous alloy atoms in a manner analogous to the double – kink mode of propagation of dislocations in pure

metals. As many solute atoms are involved in the process, the latter may be regarded “geometrically” as a limiting case of that envisaged by Nabarro [10] for rather more dilute alloys. Further scope for investigating the validity of the theory [2] arose with the publication by Basinski [1] of a paper on solid – solution hardening at 4K – 400K of Cu, Ag and Mg – based alloy crystals. Similarly in Butt and Feltham studied polycrystalline α - brasses with various contents of zinc and various grain size [2].

The aim of present work was to examine Cd single crystals with various content of Zn in light of Feltham’s theory [2]

2 Experimental

The experimental work reported in this article considers the mechanical properties of Cd – Zn alloys. No previous work has been reported on mechanical properties of this alloy, although interesting results can be expected. Zinc is namely slightly soluble in cadmium and forms with it various types of alloys.

The influence of Zn solute atoms on mechanical properties of Cd – Zn alloy may be described especially by measuring the critical resolved shear stress. Results obtained are then compared with those mentioned in literature and discussed.

The measurements of CRSS in the temperature interval 77 – 340 K were performed using the experimental equipment of the Department of General Physics of the Faculty of Science, Masaryk University in Brno

The temperature of deformation was determined using the resistivity measured by means of platinum resistor (ZPA Jinonice) placed near sample. The temperatures 202 – 240 K and 77 – 150 K were achieved by immersing the samples into Dewar flask with a cooling bath. For the temperature of 77 K the cooling bath was liquid nitrogen, for 77 – 150 K petrolether cooled by liquid nitrogen and for 202 – 240 K ethanol cooled by solid CO₂. The temperature of 340 K was achieved by means of a water bath and higher temperatures by immersing the samples into a heated oil bath, temperature stabilized.

CRSS τ_0 at the very low temperatures (1.5 – 80 K) were measured in the Laboratory of mechanical properties of metals FTINT of Ukraine. (Physico – Technical Institute of the Low Temperature).

The temperature of 4.2 K was achieved by means of liquid helium, temperatures 1.5 – 4.2 by means of liquid helium whose vapors were exhausted with the help of a vacuum pump. In both cases the sample was immersed into the Dewar flask with liquid helium which was immersed into another flask with liquid nitrogen. The temperatures 4.2 – 80 K were achieved in such a way that the sample were placed in helium vapors and a heating element was placed in liquid helium. The greater the intensity of electrical current flowing in the heating element, the more intense was the evaporation of helium and therefore the lower the temperature of helium vapors was achieved.

3 Temperature dependence of the critical resolved shear stress

The Critical Resolved Shear Stress (CRSS) of Cd – Zn alloys was measured at very low temperatures (1.5 – 80 K). The main problem that we solved was the existence, and/or the shape of temperature dependence of the CRSS anomaly. To eliminate the dependence of CRSS on the history of the sample, the method of one sample was used (Yoshizawa and Kamada [11]). The substance of this method is as follows: the sample is loaded up to the time, when the first creep step

occurs (the first major plastic deformation). But this creep step must be interrupted the sample unloaded and the temperature changed (or we can repeat the measurement at the same temperature to make sure about the reproducibility of the measurements). At that new (or the same) temperature we proceed like in the previous case. In this way we “divide” the elongation belonging to one creep step into 5 – 10 sections and obtain a correspond number of CRSS values.

In the course of that procedure the relative elongation of the sample increases all the time (and thus the CRSS increases slightly as well), but that the increment of CRSS of most samples was so short that even in the return to the initial temperature no increase in the CRSS was observed. The creep of Cd – Zn alloys with a low concentration of the solute, corresponding to CRSS, is sufficiently great, so that all measurements of the CRSS are usually covered by the segments of the first great creep step. Only for the most alloyed samples a slight hysteresis of $\tau_0(T)$ dependence in the return to the initial temperature was observed.

In this way 12 samples of various solute concentrations were measured in the temperature interval 1.5 – 80 K. The results of these measurements are summarized in Fig. 1. From this figure it can be seen that the CRSS increases monotonously with decreasing temperature only for alloys with a low concentration of solute atoms (max. 0.053 at.% Zn). For more alloyed samples (0.053 – 0.746 at.% Zn) we found that in the very low temperature region the CRSS is independent on temperature. The temperature at which the monotony is affected is an undefined function of solute concentration. But it is always lower than ~ 12 K, as is evident from Fig.1.

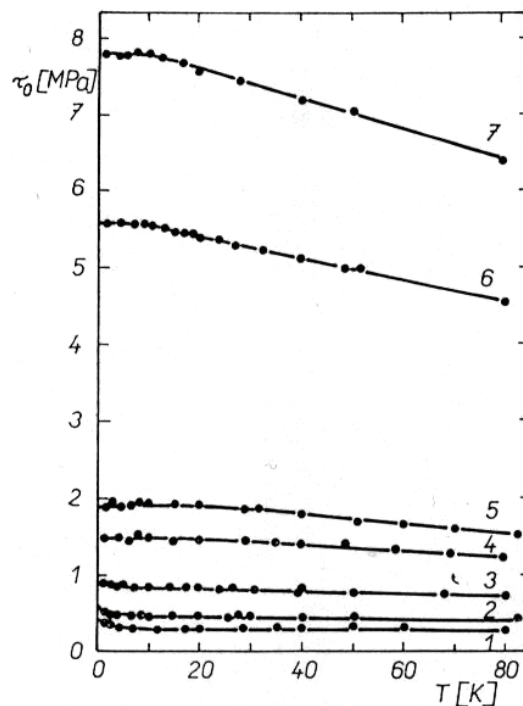


Fig.1. The temperature dependence of the CRSS (Cd – Zn alloy single crystals)
 1 – Cd+0.0022 at.% Zn, 2 – Cd+0.013 at.% Zn, 3 – Cd+0.0533 at.% Zn,
 4 – Cd+0.215 at.% Zn, 5 – Cd+0.348 at.% Zn, 6 – Cd+0.567 at.% Zn,
 7 – Cd+0.645 at.% Zn.

4 Concentration dependence of CRSS

The results of the concentration dependence of the CRSS measurements show that the CRSS value for Cd – Zn alloys, derived by extrapolation to 0 K satisfies in the concentration region 0.0022 – 0.25 at.%Zn the dependence

$$\tau_0 = \tau_0(Cd) + S \cdot c^{\frac{1}{2}} \quad (1)$$

where $\tau_0(Cd) = 0.24$ MPa and $S = 78$ MPa

This statement is evident from Fig.2, where the dependences $\tau_0\left(c^{\frac{1}{2}}\right)$, $\tau_0\left(c^{\frac{2}{3}}\right)$, and $\tau_0(c)$ are plotted. From this figure it follows that only dependence $\tau_0\left(c^{\frac{2}{3}}\right)$ is linear and it means true.

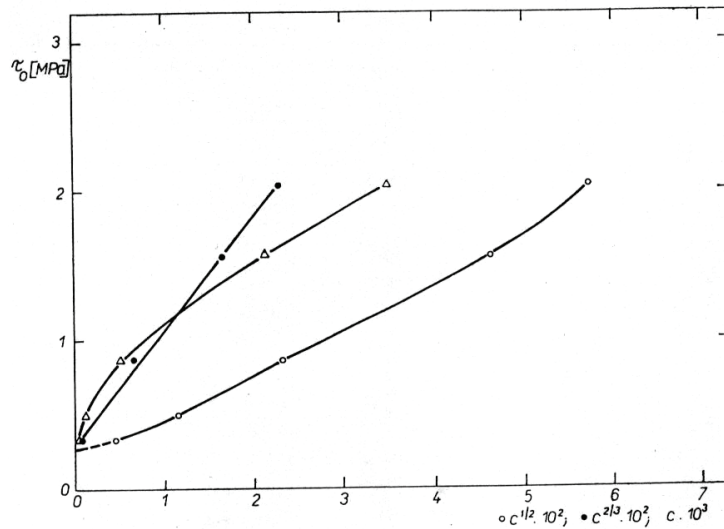


Fig.2. The concentration dependence of the CRSS.

5 Discussion

5.1 Temperature dependence of the CRSS

The measurements of the temperature dependence of the CRSS by the method of one sample showed that $\tau_0(T)$ dependence is of double character as a function of solute atoms concentration:

- The CRSS increased monotonously with decreasing temperature in the concentration interval 0.0022 – 0.053 at.%Zn.
- In the concentration region 0.215 – 0.746 at.%Zn $\tau_0(T)$ dependence is similar to that in a) only for the temperatures $T > 12$ K. For temperatures $T < 12$ K the CRSS is nor temperature dependent.

The change of dependence type a) into dependence type b) can be considered as an anomaly of $\tau_0(T)$ dependence, which is dependent on the concentration of solute atoms. A relatively small number of authors engaged in the occurrence of the $\tau_0(T)$ dependence up to this time. The digest of the results is presented in [10,12].

Present theories of the $\tau_0(T)$ dependence either do not consider the solute atoms influence, or when they consider it, they do not give an explicit answer to the question of the shape $\tau_0(T)$ dependence for various concentrations of solute atoms.

The influence of the concentration of solute atoms on the anomaly of $\tau_0(T)$ dependence are considered especially in the dynamical and quasidynamical theories [10,12]. Because for Cd – Zn single crystals the anomaly of temperature dependence of the CRSS is very concentration sensitive, we consider that for temperature $T < 12$ K one of both mechanisms is accepted. But it is clear that this argument is insufficient. To determine the most probable dislocation mechanism active at low temperatures, more comparative measurements should be made both on Cd – Zn alloys, and on other cadmium alloys. It is also necessary to improve hitherto theories of temperature dependence of the CRSS.

5.2 Concentration dependence of the CRSS.

The CRSS of pure cadmium and cadmium alloy single crystals was mostly investigated by Lukáč et al. [13,14,15]. They determined the $\tau_0(c)$ dependence in the form

$$\tau_0 = \tau_0(Cd) + S \cdot c^{\frac{2}{3}} \quad (2)$$

where $\tau_0(Cd) = 0,20$ MPa and $S = 50$ MPa.

To summarize our results of concentration dependence of the CRSS (1) with those of Lukáč et al, we can conclude that they are in good relation with Labusch's theory of CRSS.

References

- [1] BASINSKI, Z.S., FOXAL, R.A., PASCUAL, R.: Scripta Met., Vol. 6, pp. 807, 1972.
- [2] BUTT, M.Z., FELTHAM, P.: Acta Metallurgica, Vol. 26, pp. 167 – 173, 1978.
- [3] FELTHAM, P.: J. Appl. Phys. (J. Phys. D), Vol.1, pp. 303, 1968.
- [4] LEHMAN, G.: J. Appl. Phys. (J. Phys. D), Vol. 2, pp.126, 1969.
- [5] FRIEDEL, J.: *Dislocations*, Pergamon Press, London 1964.
- [6] FLEISCHER, R.L.: Acta Met., Vol.11, pp. 203, 1963.
- [7] NABARRO, F.R.N.: Phil. Mag., Vol. 35, pp. 613, 1977.
- [8] FELTHAM, P.: Mater. Sci. Engn., Vol. 11, pp. 118, 1973.
- [9] LABUSCH, R.: Acta Met., Vol. 20, pp. 917, 1972.
- [10] CAILLARD, D., MARTIN, J. L.: *Thermally activated mechanisms in crystal Plasticity*. Pergamon 2003.
- [11] KAMADA, R., YOSHIKAWA, I.: J. Phys. Soc. Japan, Vol. 31, pp. 1056, 1971.
- [12] STARTSEV, V.I., ILJICHEV, V.,J., PUSTOVALOV, V.V.: *Plasticity and Strength of Metals and Alloys at Low Temperatures*. Metallurgia Moskva 1975.
- [13] LUKÁČ, P., TROJANOVÁ, Z.: Z. Metallkunde, Vol.58, pp.57, 1967.

- [14] LUKÁČ, P.: Phys. Stat. Sol. Vol.19, pp K47, 1967.
[15] LUKÁČ, P.: Journ. of Sci. Ind. Res., Vol. 32, pp.569, 1973.

Current address

Navrátil Vladislav, prof.,RNDr.,CSc.

Department of Physics, Faculty of Education, Masaryk University

Poříčí 7, 603 00 Brno, Czech Republic

Tel: +420 549495753

e-mail: navratil@ped.muni.cz

Novotná Jiřina, PhDr., PhD.

Department of Mathematics, Faculty of Education, Masaryk University

Poříčí 31, 603 00 Brno, Czech Republic

Tel: +420 549491663

e-mail: novotna@ped.muni.cz

COMPUTATION OF BOUNDARIES AND SUMS OF REDUCED HARMONIC SERIES FORMED BY ZERO AND ANOTHER DIGIT

POTŮČEK Radovan, (CZ)

Abstract. This contribution deals with harmonic series and especially with one of the simplest reduced harmonic series – the reduced harmonic series formed by zero and another digit. The lower boundaries and the upper one's of sums of all these nine reduced harmonic series are derived. The values of these boundaries are then given in more precise way. Furthermore sums of these nine infinite series are computed. This contribution as an example of scientific computations in CAS Mathcad can be an inspiration for teachers of mathematics whose are teaching the topic Infinite series or as a subject matter for work with talented students.

Key words and phrases. harmonic series, geometric series, reduced harmonic series, lower and upper boundaries.

Mathematics Subject Classification. Primary 40A05; Secondary 65B10.

1 Harmonic series and reduced harmonic series

Let us recall the basic terms and notions. The *harmonic series* is the sum of reciprocals of all natural numbers (except zero), so this is the series

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \cdots . \quad (1)$$

The divergence of the series (1) can be proved in many ways. A simple proof can be done by using the integral test of convergence. We can also use the comparison test and show that

$$\sum_{n=1}^{2^m} \frac{1}{n} > 1 + \frac{m}{2}$$

for every positive integer m . To do so we prove that the sequence $\{s_n\} = \{s_{2^m}\}$ of partial sums is increasing, although very very slowly (e.g. $s_n > 100$ for $n > 2^{198} \doteq 4 \cdot 10^{59}$). In the paper [1] twenty various proofs of the divergence for the series (1) are stated.

The *reduced harmonic series* is defined as the subseries of the harmonic series, which arises by omitting some its terms. As an example of the reduced harmonic series we can take the series formed by reciprocals of primes and number one

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \frac{1}{11} + \frac{1}{13} + \dots \quad (2)$$

This reduced harmonic series is still divergent, although from the first 100, 1000, 10000, 100000, ... terms of the harmonic series 74, 831, 8770, 90407, ... terms with composite denominators are omitted. The proof of divergence of the series (2) (see e.g. book [2]) was made first time by Leonhard Euler in 1737.

An interesting example of reduced harmonic series are so-called Kempner's series K_a . The Kempner series is a modification of the harmonic series, formed by omitting all terms whose denominator expressed in base 10 contains a digit a . That is, it is the sum of fractions $1/n$ where n takes only values whose decimal expansion has no digit a . The series K_9 with omitted the 9 digit was first studied by A. J. Kempner in 1914 in the paper [3]. This series is interesting because of the counter-intuitive result that unlike the harmonic series it converges. Kempner showed this value was less than 80. The upper bound of 80 is very crude, and F. Irwin showed in 1916 in the paper [4] by a slightly finer analysis of the bounds that the value of this Kempner series is between 22.4 and 23.3. T. Schmelzer and R. Baillie in their paper [5] showed that up to 20 decimals; the actual sum is 22.92067661926415034816.

Kempner's proof of convergence the series

$$K_9 = \frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{8} + \frac{1}{10} + \frac{1}{11} + \dots + \frac{1}{18} + \frac{1}{20} + \frac{1}{21} + \dots + \frac{1}{88} + \dots + \frac{1}{100} + \frac{1}{101} + \dots + \frac{1}{888} + \dots,$$

is very simple. First we group the terms of the series K_9 according to the number of the digits in the denominator. We obtain the series in the form

$$\left(\frac{1}{1} + \dots + \frac{1}{8}\right) + \left(\frac{1}{10} + \dots + \frac{1}{88}\right) + \left(\frac{1}{100} + \dots + \frac{1}{888}\right) + \dots + \left(\frac{1}{10^{n-1}} + \dots + \frac{1}{88\dots 8}\right) + \dots,$$

where the last fraction contains in the denominator exactly n eights. Note that the number of fractions with exactly k digits, which have no one digit 9 in their denominators, is $8 \cdot 9^{k-1}$, because the digit from the set $\{1, 2, \dots, 8\}$ on the first position can be choosen by 8 ways and the digit from the set $\{0, 1, \dots, 8\}$ on the further $k - 1$ positions by 9 ways. Since all terms in every one parenthesis are less or equal to the first term, i.e. the terms in the first parenthesis are less or equal to $1/1$, the terms in the second parenthesis are less or equal to $1/10$ etc. as far as the terms of n -th parenthesis are less or equal to fraction $1/10^{n-1}$, so that the sum S_9 of the series K_9 can be bounded from above by the sum

$$S = 8 \cdot 1 + \frac{8 \cdot 9}{10} + \frac{8 \cdot 9^2}{10^2} + \dots + \frac{8 \cdot 9^{n-1}}{10^{n-1}} + \dots$$

Because this sum S is the sum of the convergent infinite geometric series with the first term $a_1 = 8$ and with the ratio $q = 9/10$, we can use the formula

$$S = \frac{a_1}{1 - q}. \quad (3)$$

So the sum S_9 is upper bounded by the number $S = 8/(1 - 9/10) = 8/(1/10) = 80$ and therefore the series K_9 converges. By the same way we can prove that the sum S_9 is bounded from below by the number 8. Finally we have $8 < S_9 < 80$. Analogously we can prove the convergence of the remaining nine Kempner series K_0, K_1, \dots, K_8 . The following table shows the approximative values of sums S_0, S_1, \dots, S_9 of all ten Kempner sums rounded to three decimal places, which were more precisely enumerated (rounded to twenty decimal places) and published in 1979 by the American mathematician and software expert Robert Baillie in his paper [6]:

a	0	1	2	3	4	5	6	7	8	9
S_a	23.103	16.177	19.257	20.570	21.327	21.835	22.206	22.493	22.726	22.921

2 Calculation of boundaries and sums for reduced harmonic series formed by one digit

Since this paper is a free continuation and generalization of the paper [7] dealing with calculation of boundaries and sums s_a for reduced harmonic series formed by the only one digit a , we now briefly state its main results. Let us consider for $a = 1, 2, \dots, 9$ the series of the form

$$\frac{1}{a} + \frac{1}{aa} + \frac{1}{aaa} + \frac{1}{aaaa} + \dots = \frac{1}{a} \left(1 + \frac{1}{11} + \frac{1}{111} + \frac{1}{1111} + \dots \right). \quad (4)$$

Using the formula for the sum of convergent geometric series it was proved that

$$s(1) = 1 + \frac{1}{11} + \frac{1}{111} + \frac{1}{1111} + \dots < 1 + \frac{1}{10} + \frac{1}{100} + \frac{1}{1000} + \dots = \frac{1}{1 - 1/10} = \frac{10}{9},$$

so $s_a < (1/a) \cdot (10/9) = 10/9a$ holds for $a = 1, 2, \dots, 9$. From inequalities

$$\frac{1}{11} > \frac{9}{100}, \quad \frac{1}{111} > \frac{9}{1000}, \quad \frac{1}{1111} > \frac{9}{10000}, \quad \dots,$$

we get

$$\begin{aligned} s_1 &= 1 + \frac{1}{11} + \frac{1}{111} + \frac{1}{1111} + \dots > 1 + \frac{9}{100} + \frac{9}{1000} + \frac{9}{10000} + \dots = \\ &= 1 + \frac{9}{100} \left(1 + \frac{1}{10} + \frac{1}{100} + \frac{1}{1000} + \dots \right) = 1 + \frac{9}{100} \cdot \frac{10}{9} = 1 + \frac{1}{10} = \frac{11}{10}, \end{aligned}$$

so $s_a > (1/a) \cdot (11/10) = 11/10a$ for $a = 1, 2, \dots, 9$. Finally, we obtain boundaries

$$\frac{11}{10a} < s_a < \frac{10}{9a}, \quad a = 1, 2, \dots, 9.$$

Since the n th term t_n of the series (4) with the n digits a in the denominator can be written in the form

$$t_n = \frac{1}{\underbrace{aa \dots a}_n} = \frac{1}{a \cdot \underbrace{11 \dots 1}_n} = \frac{1}{a \cdot (10^{n-1} + 10^{n-2} + \dots + 1)} = \frac{1}{a \cdot \sum_{i=0}^{n-1} 10^i},$$

then

$$s_a = \sum_{n=1}^{\infty} t_n = \frac{1}{a} \cdot \sum_{n=1}^{\infty} \frac{1}{\sum_{i=0}^{n-1} 10^i} = \frac{1}{a} \cdot \sum_{n=1}^{\infty} \left(1 / \sum_{i=0}^{n-1} 10^i \right).$$

For an evaluation of the sum s_a we use the approximation formula $s_a \doteq \frac{1}{a} \cdot \sum_{n=1}^m \left(1 / \sum_{i=0}^{n-1} 10^i \right)$ with the convenient upper limit m . Using the computer algebra system **Mathcad 13** we can compute the sum $\sum_{n=1}^m \left(1 / \sum_{i=0}^{n-1} 10^i \right)$ up to $m = 309$. For $m = 310$ we obtain the error message. So for $a = 1$ we get

$$s_1 \doteq \sum_{n=1}^{309} \left(1 / \sum_{i=0}^{n-1} 10^i \right) = 1.10091819083620068,$$

hence $s_a \doteq 1.10091819083620068/a$ for $a = 2, 3, \dots, 9$. In the following table ℓ_a denotes the lower boundary $11/10a$ and u_a the upper one $10/9a$ of the sum s_a for $a = 2, 3, \dots, 9$:

a	ℓ_a	s_a	u_a
1	1.100000	1.10091819083620068	1.111111
2	0.550000	0.55045909541810034	0.555556
3	0.366667	0.36697273027873356	0.370370
4	0.275000	0.27522954770905017	0.277778
5	0.220000	0.22018363816724013	0.222222
6	0.183333	0.18348636513936678	0.185119
7	0.157143	0.15727402726231438	0.158730
8	0.137500	0.13761477385452509	0.138889
9	0.122222	0.12232424342624452	0.123457

3 Reduced harmonic series formed by zero and digit a

The series of the form

$$R_{0,a} = \frac{1}{a} + \left(\frac{1}{a0} + \frac{1}{aa} \right) + \left(\frac{1}{a00} + \frac{1}{a0a} + \frac{1}{aa0} + \frac{1}{aaa} \right) + \dots, \quad (5)$$

where $a \in \{1, 2, \dots, 9\}$, is called the *reduced harmonic series formed by zero and digit a* . Since the first digit in denominators cannot be zero, the group of fractions with n digits in their denominators has exactly 2^{n-1} terms. Since the equality

$$\frac{1}{a} + \frac{1}{a0} + \frac{1}{aa} + \frac{1}{a00} + \dots + \frac{1}{aaa} + \dots = \frac{1}{a} \left(\frac{1}{1} + \frac{1}{10} + \frac{1}{11} + \frac{1}{100} + \dots + \frac{1}{111} + \dots \right)$$

holds, i.e.

$$R_{0,a} = \frac{1}{a} R_{0,1},$$

we shall primarily deal with the reduced harmonic series $R_{0,1}$ formed by 0 and 1. Furthermore obtained results will be extended for another reduced series $R_{0,a}$, where $a \in \{2, 3, \dots, 9\}$.

4 Upper boundaries $u_{0,a}$ for the sums $S_{0,a}$

Similarly as in the Section 1 by evaluation the upper boundary for the sum S_9 of the Kempner's series K_9 the sum $S_{0,1}$ of the series $R_{0,1}$ can be bounded from above:

$$\begin{aligned} \frac{1}{1} + \left(\frac{1}{10} + \frac{1}{11} \right) + \left(\frac{1}{100} + \dots + \frac{1}{111} \right) + \dots + \left(\frac{1}{10^n} + \dots + \underbrace{\frac{1}{11\dots1}}_{n+1} \right) + \dots < \\ < 1 + \frac{2}{10} + \frac{4}{100} + \dots + \frac{2^n}{10^n} + \dots \end{aligned} \quad (6)$$

Using (3), the upper sum can be evaluated as the number

$$u_{0,1} = 1 + \frac{2}{10} + \frac{4}{100} + \dots + \frac{2^n}{10^n} + \dots = 1 + \frac{1}{5} + \frac{1}{5^2} + \dots + \frac{1}{5^n} + \dots = \frac{1}{1 - 1/5} = \frac{5}{4},$$

so the increasing series $R_{0,1}$ is bounded from above, thus it is the convergent series. From this we have that the series $R_{0,a}$, where $a \in \{2, 3, \dots, 9\}$, are convergent, too, because the sums $S_{0,a}$ are bounded from above by the numbers

$$u_{0,a} = \frac{1}{a} \cdot \frac{5}{4} = \frac{5}{4a}.$$

5 Lower boundaries $\ell_{0,a}$ for the sums $S_{0,a}$

It is clear that the inequalities

$$\frac{1}{11} > \frac{9}{100}, \quad \frac{1}{111} > \frac{9}{1000}, \quad \dots, \quad \frac{1}{10^n + 10^{n-1} + \dots + 1} > \frac{9}{10^{n+1}}, \quad \dots$$

hold. Using the formula for the sum of the first $n+1$ terms of the geometric sequence with the first term a_1 and with the ratio q , $q \neq 1$,

$$s_{n+1} = a_1 \frac{q^{n+1} - 1}{q - 1}, \quad (7)$$

we can write the denominator on the left hand side of the last inequality in the form

$$1 + 10 + \dots + 10^n = \frac{10^{n+1} - 1}{9}.$$

Since all the terms of each parenthesis on the left hand side in the inequality (6), expressing the sum $S_{0,1}$, are less or equal of the last term of each parenthesis, we get

$$\begin{aligned} & \frac{1}{1} + \left(\frac{1}{10} + \frac{1}{11} \right) + \left(\frac{1}{100} + \dots + \frac{1}{111} \right) + \dots + \left(\frac{1}{10^n} + \dots + \underbrace{\frac{1}{11\dots 1}}_{n+1} \right) + \dots > \\ & > 1 + \frac{2}{11} + \frac{4}{111} + \dots + \frac{2^n}{10^n + \dots + 10 + 1} + \dots > 1 + \frac{2 \cdot 9}{100} + \frac{4 \cdot 9}{1000} + \dots + \frac{2^n \cdot 9}{10^{n+1}} + \dots \end{aligned}$$

The last expression, which can be using (3) written in the form

$$1 + \frac{2 \cdot 9}{100} \left(1 + \frac{1}{5} + \dots + \frac{1}{5^{n-1}} + \dots \right) = 1 + \frac{9}{50} \cdot \frac{1}{1 - 1/5} = 1 + \frac{9}{50} \cdot \frac{5}{4} = \frac{49}{40},$$

represents the lower boundary $\ell_{0,1}$ of the sum $S_{0,1}$ for the series $R_{0,1}$. Analogously the sums $S_{0,a}$ of the series $R_{0,a}$, where $a \in \{2, 3, \dots, 9\}$, are bounded from below by numbers

$$\ell_{0,a} = \frac{1}{a} \cdot \frac{49}{40} = \frac{49}{40a}.$$

Generally, we obtain inequalities

$$\frac{49}{40a} < S_{0,a} < \frac{50}{40a}$$

which represent the boundaries for the sums $S_{0,a}$ of the series $R_{0,a}$, where $a \in \{1, 2, \dots, 9\}$.

6 More accurate value $L_{0,1}$ of the lower boundary for the sum $S_{0,1}$

The last (the smallest) terms in the parentheses, by means of them the series $R_{0,1}$ was written in the relation (6), were used for evaluation of the lower boundary $\ell_{0,1}$ for the sum $S_{0,1}$. For obtaining a more accurate value of the lower boundary $\ell_{0,1}$ for the sum $S_{0,1}$ of the series $R_{0,1}$ we now use the second of two middle terms in these parentheses. This means that we use the $(2^{n-1} + 1)$ -st term in the n -th parenthesis with 2^n terms. If we denote the sum of the n -th parenthesis by b_n , i.e. if we denote

$$\begin{aligned} b_1 &= \frac{1}{10} + \frac{1}{11}, \quad b_2 = \frac{1}{100} + \frac{1}{101} + \frac{1}{110} + \frac{1}{111}, \\ b_3 &= \frac{1}{1000} + \frac{1}{1001} + \frac{1}{1010} + \frac{1}{1011} + \frac{1}{1100} + \frac{1}{1101} + \frac{1}{1110} + \frac{1}{1111}, \quad \dots, \end{aligned}$$

we get

$$\frac{2}{11} < b_1, \quad \frac{4}{110} < b_2, \quad \frac{8}{1100} < b_3, \quad \frac{16}{11000} < b_4, \quad \dots, \quad \frac{2^n}{11 \cdot 10^{n-1}} < b_n, \quad \dots$$

If we denote by $L_{0,1}$ the more accurate value of the lower boundary $\ell_{0,1}$ for the sum $S_{0,1}$ of the series

$$R_{0,1} = 1 + b_1 + b_2 + \cdots + b_n + \cdots = 1 + \sum_{n=1}^{\infty} b_n, \quad (8)$$

then

$$L_{0,1} = 1 + \sum_{n=1}^{\infty} \frac{2^n}{11 \cdot 10^{n-1}} = 1 + \sum_{n=1}^{\infty} \frac{2^n \cdot 10}{11 \cdot 10^n} = 1 + \frac{10}{11} \sum_{n=1}^{\infty} \frac{2^n}{10^n} = 1 + \frac{10}{11} \sum_{n=1}^{\infty} \frac{1}{5^n}.$$

The last expression can be simplified using (3), so we get

$$L_{0,1} = 1 + \frac{10}{11} \cdot \frac{1/5}{1 - 1/5} = 1 + \frac{10}{11} \cdot \frac{1}{4} = 1 + \frac{5}{22} = \frac{27}{22}.$$

Indeed, the number $L_{0,1} = 27/22$ is in comparison with the number $\ell_{0,1} = 49/40$ the more accurate lower boundary, since $27/22 > 49/40$.

7 More accurate value $U_{0,1}$ of the upper boundary for the sum $S_{0,1}$

The first (the greatest) terms in each parentheses

$$\frac{1}{1} + \left(\frac{1}{10} + \frac{1}{11} \right) + \left(\frac{1}{100} + \cdots + \frac{1}{111} \right) + \cdots + \left(\frac{1}{10^n} + \cdots + \underbrace{\frac{1}{11 \dots 1}}_{n+1} \right) + \cdots \quad (9)$$

were used for evaluation of the upper boundary $u_{0,1}$ for the sum $S_{0,1}$. To get the more accurate value of the upper boundary $u_{0,1}$ for the sum $S_{0,1}$ of the series $R_{0,1}$ we now use the first of two middle terms in these parentheses. In other words, we use the $(2^{n-1} - 1)$ -st term in the n -th parenthesis with 2^n terms. For the sum b_n of the n -th parenthesis

$$b_1 < \frac{2}{10}, \quad b_2 < \frac{4}{101}, \quad b_3 < \frac{8}{1011}, \quad \dots, \quad b_n < \frac{2^n}{10^n + 10^{n-2} + \cdots + 10 + 1}, \quad \dots$$

holds. The last mentioned fraction can be using (7) written in the form

$$\begin{aligned} \frac{2^n}{(10^n + 10^{n-1} + 10^{n-2} + \cdots + 1) - 10^{n-1}} &= \frac{2^n}{(10^{n+1} - 1)/(10 - 1) - 10^{n-1}} = \\ &= \frac{2^n}{(10^{n+1} - 1 - 9 \cdot 10^{n-1})/9} = \frac{9 \cdot 2^n}{10^{n+1} - 9 \cdot 10^{n-1} - 1}. \end{aligned}$$

If we denote by $U_{0,1}$ the more accurate value of the upper boundary $u_{0,1}$ for the sum $S_{0,1}$ of the series (8), then we get

$$U_{0,1} = 1 + \sum_{n=1}^{\infty} \frac{9 \cdot 2^n}{10^{n+1} - 9 \cdot 10^{n-1} - 1} = 1 + 9 \cdot \sum_{n=1}^{\infty} \frac{2^n}{10^{n+1} - 9 \cdot 10^{n-1} - 1}.$$

The last mentioned sum cannot be either simplified nor fully exactly evaluated. For its approximate calculation we hence use the sum $s_{0,1}(m) = \sum_{n=1}^m \frac{2^n}{10^{n+1} - 9 \cdot 10^{n-1} - 1}$ for the convenient upper limit m . As the computer algebra system **Mathcad 13** gives for $m = 23$ the value $s_{0,1}(23) = 0.02772166195402140$ and for $m \geq 24$ we obtain $s_{0,1}(m) = 0.02772166195402141$, then we get

$$U_{0,1} \doteq 1 + 9 \cdot s_{0,1}(24) = 1 + 9 \cdot 0.02772166195402141 = 1.2494949575861927.$$

8 More accurate values $L_{0,a}$, $U_{0,a}$ of the lower and upper boundaries for the sums $S_{0,a}$ and values of the sums $S_{0,a}$

In Section 6 we have obtained $L_{0,1} = 27/22 = 1.2\overline{27}$, so we get $1.2\overline{27} < S_{0,1} < 1.2494949575861927$. Considering the relations $L_{0,a} = L_{0,1}/a$, $U_{0,a} = U_{0,1}/a$, where $a \in \{2, 3, \dots, 9\}$, we get inequalities

$$1.22727272727273/a < S_{0,a} < 1.2494949575861927/a.$$

For computing the sum $S_{0,a} = S_{0,1}/a$ (expressed in 6 decimals) the first 9 parentheses in (8) were used in computation the sum $S_{0,1}$, i.e. this sum was approximated by $1 + 2^1 + \dots + 2^9 = 1023$ first terms. So we have obtained the approximate value $S_{0,1} = 1.238406$.

The lower and upper boundaries $\ell_{0,a}$, $u_{0,a}$, $L_{0,a}$, $U_{0,a}$ and the sums $S_{0,a}$ (expressed in 6 decimals) are presented in the following table:

a	$\ell_{0,a}$	$L_{0,a}$	$S_{0,a}$	$U_{0,a}$	$u_{0,a}$
1	1.225000	1.227273	1.238406	1.249495	1.250000
2	0.612500	0.613637	0.619203	0.624748	0.625000
3	0.408333	0.409091	0.412802	0.416498	0.416667
4	0.306250	0.306818	0.309602	0.312374	0.312500
5	0.245000	0.245455	0.247681	0.249899	0.250000
6	0.204167	0.204546	0.206401	0.208249	0.208333
7	0.175000	0.175325	0.176915	0.178499	0.178571
8	0.153125	0.153409	0.154801	0.156187	0.156250
9	0.136111	0.136364	0.137601	0.138833	0.138889

9 Conclusion

In this paper the boundaries $1.225/a < S_{0,a} < 1.25/a$ and $1.227273/a < S_{0,a} < 1.249495/a$ for the sums $S_{0,a} \doteq 1.238406/a$ of the series $R_{0,a}$ (see (5)), where $a \in \{1, 2, \dots, 9\}$, were derived. The sums presented in the table above can be used for evaluation the sums of special reduced harmonic series formed by 0 and exclusively one of some other digits, as for example the series

$$\frac{1}{1} + \frac{1}{4} + \frac{1}{10} + \frac{1}{11} + \frac{1}{40} + \frac{1}{44} + \frac{1}{100} + \frac{1}{101} + \frac{1}{110} + \dots = S_{0,1} + S_{0,4} \doteq 1.548008.$$

References

- [1] KIFOWIT, S. J., STAMPS, T. A.: The Harmonic Series Diverges Again and Again. *The AMATYC Review*. 2006, Vol. 27, No. 2, pp. 31–43. ISSN 0740-8404.
- [2] HARDY, G. H., Wright, E. M.: An Introduction to the Theory of Numbers, 4th Edition. Oxford University Press, London, 1975. ISBN 0-19-853310-7.
- [3] KEMPNER, A. J.: A Curious Convergent Series. *American Mathematical Monthly*. 1914, Vol. 21, No. 1, pp. 48–50. ISSN 0002-9890.
- [4] IRWIN, F.: A Convergent Series Derived from the Harmonic Series. *American Mathematical Monthly*, 23 (5), 149-152. ISSN 0002-9890.
- [5] SCHMELZER, T. and BAILLIE, R.: Summing a Curious, Slowly Convergent Series. *American Mathematical Monthly*, 115 (6), 525-540. ISSN 0002-9890.
- [6] BAILLIE, R.: Sums of Reciprocals of Integers Missing a Given Digit. *American Mathematical Monthly*. 1979, Vol. 86, pp. 372–374. ISSN 0002-9890.
- [7] POTŮČEK, R.: Výpočet ohraňování a součtů redukovaných harmonických řad tvořených jednou cifrou. *Proceedings of the 6th Žilina didactic conference DIDZA* (on CD-Rom), Žilinská univerzita, Slovak Republic, 2009, 7 pp. ISBN 978-80-554-0050-1 (in Czech).

Current address

RNDr. Radovan Potůček, Ph.D.

Department of Mathematics and Physics, University of Defence, Faculty of Military Technology, Kounicova 65, 662 10 Brno, Czech Republic, tel. 0042 973 443 056, Radovan.Potucek@unob.cz

MINIMALIST IMPLEMENTATION OF A GENETICALLY PROGRAMMING PROCESS WRITTEN IN ASSEMBLY LANGUAGE

SKORKOVSKÝ Petr, (CZ)

Abstract: If we try to focus to the field of genetically programming and today's usual methods how to implement it's algorithms in praxis, it is obvious that new kind of genetically programming languages must be developed. Very often the language itself uses complex syntactic expressions to describe algorithms themselves on one hand and then it uses difficult, time consuming processing of this so coded language on the other hand. Paper tries to introduce an implementation of genetically programming process highly efficient for time and for computing resources, written completely in assembly language. Currently the author is working on the coding of the program. As such it is not possible to show examples and results of the processing for now as the program is finished around 70%, however theoretical background of this program is completed and it is further discussed through the paper.

Key words: genetically programming, assembly language, minimalist implementation, efficient computing, cellular automaton

Mathematics Subject Classification: Computer science, Artificial intelligence 68T04.

1 Theoretical introduction

As one of the promising directions of artificial intelligence - the wide scientific discipline, there are evolutionary algorithms and as a subclass of them, are genetic algorithms. All of them have found inspiration in the nature, derived from the Darwin's theory of evolution.

As another subpart of the genetic algorithms field, is known the technology for automatic evolution of algorithms - genetically programming. To use this technology, algorithms are encoded with specialized programming language and thus are coded like a genetic code [1.]. Programs are then cyclically, genetically evaluated, to achieve with the use of a step by step process, better performance for solving of some specific, very difficult and complicated problems.

Usually when looking for a solution of a problem where good algorithms already exist and all used variables are well known also, it would be faster and more effective if a classical way (numeric calculation, other programming language) is used for the problem solving. It means, where classic

methods of problem solving are not effective enough the genetic programming could be more successful. Unfortunately, there is no warranty of success at all cases of course. In the Table 1, there are listed several typical problems which could be solved by a genetically programming process [1.] p.129.

	Definition of the problem to be solved	Algorithm to be found	Input data, parameters of the problem	Output data, solved results
1.	Induction of a sequence of consecutive numbers	Analytical rule	Index of a number from the sequence	The number's value of the sequence member
2.	Symbolical regression of a data set	Mathematical formula	Independent variables	Dependent variables
3.	Optimal control	Controlling strategy	State variables	Action magnitudes
4.	Identification and prediction	Mathematical system model	Independent variables	Dependent variables
5.	Classification	Decision tree	Attribute values	Classification into a class
6.	Learning of a specific individual behavior	Program describing a behavior	Data from sensors	Movements, actions of artificial limbs
7.	Deduction of collective behavior	Program describing behavior of one member from a collective	Information about links between one member and rest of the collective	Actions of one member inside the collective

Table 1: Typical problems which could be solved by a genetically programming process

2 Motivation

Usually for implementation of genetic algorithm and genetically programming, high programming languages like C, C++, Visual Basic, etc. are used. According to the main principle of genetic algorithms a lot of calculation cycles are repeated many times, while searching through huge number of generations to fit the target definition of solution. The computer microprocessor repeats in each of the generation cycle more program code than necessary and genetic algorithm could not be fast and effective enough. One algorithm written in a high programming language contains much more CPU instructions like jumps, subroutine calls, memory readings and writings and stack operations than the same algorithm written directly in the assembly language.

To achieve much higher speed and efficiency of the processing, programming in assembly language is to be used. The speed of the whole process can be significantly increased by limiting stack and memory operations, jumps and calls, while CPU registers must be used to hold values of variables instead of to leave them to be in the RAM. It is recommended to evaluate each part of the algorithm step by step and to increase the efficiency of the code several times again and again. The evolution of assembler programs consumes a lot of time and costs a lot of work, but the resulting speed and quality of the genetic algorithm is unbeatable by any other concurrent program written in a high programming language.

3 Program description

For coding of my implementation of genetic programming application I have used assembly language of 32-bit CPU x86 family instruction set. The program is compatible with Windows OS family and the user interface is programmed to use standard Win32 API functions. For compilation of the application's source code the Microsoft's MASM32 SDK V10 must be used.

Unfortunately the program itself is not finished yet, however about 70% of the code is completed and functional but currently the author can not provide results of the application execution. Program's algorithms, several used data types and different ideas coming from theoretical constructions are discussed through next parts of the paper.

3.1 Main program algorithm

My implementation of genetically programming algorithm which consists of several programming blocks (assembler subroutines), is described on the Figure 1 (inspired by [2.], p.27).

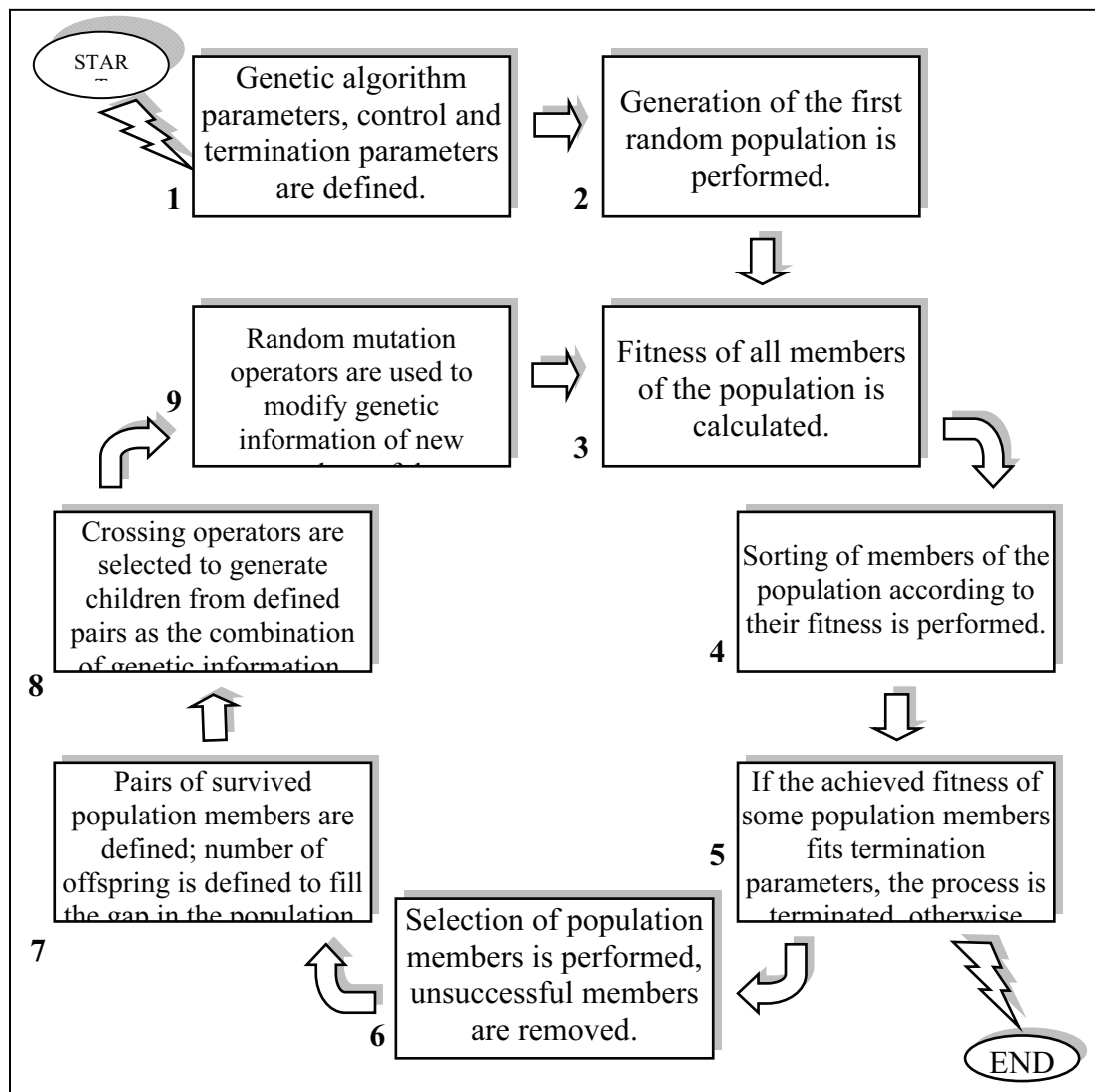


Figure 1: The main program's algorithm

3.1.1 Definition of genetic algorithm's parameters, control and termination parameters

Before the searching for an algorithm and before the evolution can be started, it is necessary to define and describe the problem which is to be solved. The quality (fitness) of an algorithm will be tested during the step 3 (see Figure 1 and its description in chapter 3.1.3). To perform the step 3 correctly, it is necessary to prepare a set of input-output patterns sequence, containing well known examples of correct algorithm behavior. Details about coding of such sequence of patterns can be found in the chapter 3.1.3.

As next there is a need to define the size of the population of algorithms to be used for the evolution. It is decided to keep the population size constant for all evolution cycles, so when a lot of algorithm candidates are removed (killed) during one evolution cycle, all missing places are filled by new children (genetically combinations of remaining algorithm candidates). The defined size of the population depends on many factors and it must be estimated according to previous experiences.

During the step 3, for the performance of one algorithm candidate, the so-called cellular processor of logical functions is used (details are described in chapter 3.1.3). The size of the cellular processor's vector must be defined; it means how many binary cells are used inside the cellular processor. Again, its size must be estimated according to previous experiences. It depends on the complexity of the solved problem, on the number of used input and output binary variables and on the number of possible cases and internal states of the target algorithm which may be encountered.

As last it is necessary to define the condition, when the evolution process shall be terminated. The description of reasons for terminating the evolution can be found in the chapter 3.1.5.

3.1.2 Generation of the first random population

To start the evolutionary process, the population is filled by algorithm candidates containing random binary values – random genetically information [1.], p.143. A big table of random values is used which is placed to the RAM and it is used for more purposes, like probability calculations during candidates selections, genetic code crossing operators selections, mutation operators selections etc. In the table, numbers in the range of 0 to 399 are used. As like described in the part of cellular processor description, each cell of the processor's vector is coded in 32 bits. Next four numbers from the random numbers table are taken; a random offset (it is calculated at the moment when the user clicked to start the evolution process) is added to each of them and a 32 bits random number is then calculated. All calculated 32 bit random numbers are used to fill all population members - cells of their vectors.

Each vector's cell contains binary information about two relative addresses to other two cells inside the vector and their size must be limited according to the size of the vector defined on the beginning. It is used to prevent the case that some random relative address is high while the vector's size is too small.

3.1.3 Calculation of fitness for all members of the population

One part of the algorithm which is dedicated for calculation of fitness for each of the population member contains the so called cellular processor of logical functions (See 3.2). There are several patterns of input data used to feed inputs of the cellular processor. After a number of

iteration cycles of the cellular processor, there are calculated new output data on outputs from the cellular processor. They are compared to expected output data patterns from the table of input-output patterns which were defined before the evolution started. For each of the bit of the output data which is at the state like expected, one point is added to the total score variable, corresponding to fitness. If the maximum of score points is achieved (all bits of output data correspond to expected data of all predefined test cases) target algorithm is found and the main genetically programming algorithm may be terminated. Termination conditions are further described at 3.1.5.

3.1.4 Sorting of members of the population according to their fitness

All members of the population – all algorithm candidates, have their fitness calculated from the previous step. There is a need to sort all candidates. Then these with a high calculated fitness will be first in the list with a low ordinal number and these with a low fitness values will be later in the list with higher ordinal numbers [1.] p.164. After the sorting is finished, the most successful algorithm will be the first one in the list and the worst will be the last one in the list.

3.1.5 Testing of termination conditions

To avoid the situation, that the evolution algorithm lasts forever and algorithm candidates are not converging to the target algorithm fast enough or not at all, termination conditions are defined. One condition for process termination can be the case, that certain number of algorithm candidates has already reached maximum fitness during the step 3 or at least very high fitness. As the second condition may be the limitation of maximum number of evolution cycles which were already performed. The evolution will not last forever and it can be terminated if the searching for algorithm fails to converge to candidate with a good or maximum fitness.

3.1.6 Selection of population members, removal of unsuccessful members

All algorithm candidates are sorted from the maximum to the minimum fitness. Now they will proceed to the probabilistic selection. A probability function and a random numbers generator are used to decide which candidate has a right to proceed to the next evolution cycle. The probability function itself has the shape in such way, that algorithm candidates with the highest fitness will have also best expectations to pass to a next stage while others with the worst fitness have almost no change to survive, however due to the non-zero probability; they still may succeed if they are “lucky”.

To generate a probability function a table of integer values stored in the RAM may be used, while table values can be prepared in advance to prevent calculations for each time when they would be needed.

Below you can see an example of exponential function which can be used as a probability function. On the figure Nr.2, there is on left side the probability function with $k = 2$ and on the right side the same function normalized to 1000 population members and $p(x)$ normalized to 400 as the maximum probability, stored in RAM as integer numbers.

$$p(x) = e^{-\frac{x}{k}} \quad (1)$$

x : ordinal number of the algorithm candidate starting with zero (max. fitness)

k : constant which must be selected according to previous experiences

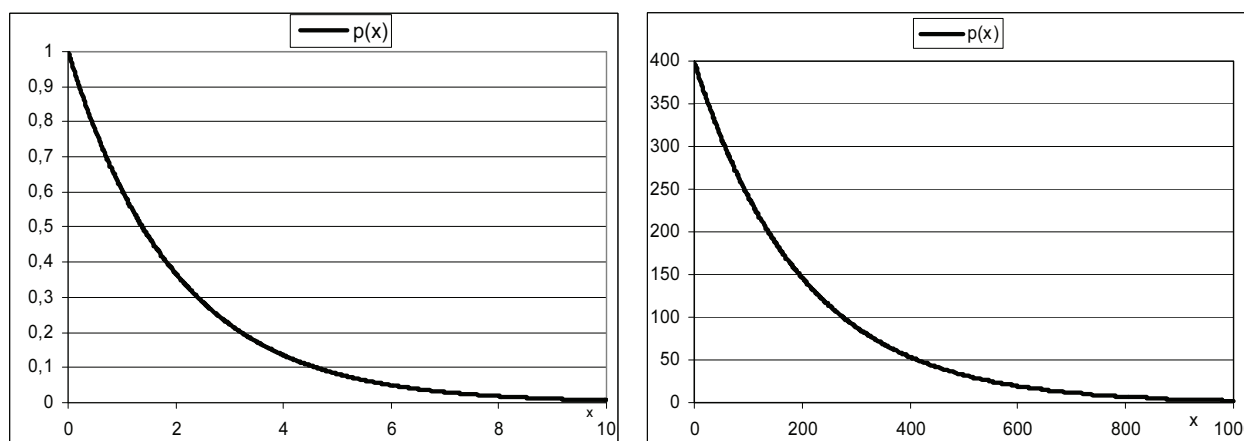


Figure 2: Left -probability function, Right -normalized to 1000 candidates and $p_{max} = 400$

3.1.7 Definition of pairs of survived population members and number of offspring

As a result of the selection, not passed candidates will be removed from the list. Because a population with constant number of members is used here, the population gap must be filled by offspring - children of most successful algorithm candidates. As a next step privileges to have children must be distributed between survived algorithm candidates. Parent pairs are formed from Even-Odd neighbored members in the list and quantity of children for each of the pair is determined according to their fitness. Pairs with highest fitness have also highest number of children.

The sum of all children can not exceed the population gap so the total could be variable depending on number of previously removed candidates. There is a table in the RAM describing how children are distributed between pairs. The strategy for distribution depends on the total missing members to fill the gap. As an example (other strategies could be used also) you can see below one strategy for children privileges distribution (g = number of missing members, the gap) for several gap widths:

$$\begin{aligned} g = 36 &= 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1, \\ g = 35 &= 8 + 7 + 6 + 5 + 4 + 3 + 2 \\ g = 34 &= 8 + 7 + 6 + 5 + 4 + 3 + 1 \\ g = 33 &= 8 + 7 + 6 + 5 + 4 + 3 \\ g = 32 &= 8 + 7 + 6 + 5 + 4 + 2 \\ g = 31 &= 8 + 7 + 6 + 5 + 4 + 1 \\ g = 30 &= 8 + 7 + 6 + 5 + 4 \\ &\dots \end{aligned} \quad (3.1.7.1)$$

3.1.8 Selection of crossing operators to generate children from defined pairs as the combination of genetic information

A table of crossing schemes is accessible at some RAM address and for each of the children one random number is used from the RND generator to select one crossing scheme which will be applied. In such crossing schemes, there are defined points of crossing and there could be one, two, or several crossing points. The resulting genetic information can be used for example from following mixtures:

$$\text{parent1} : \text{parent2} = 50\% : 50\%, 25\% : 75\%, 75\% : 25\%, 33\% : 67\%, 95\% : 5\%, \dots, \quad (2)$$

3.1.9 Random mutation operators to modify genetic information of new members of the population

A table of mutation schemes is accessible at some RAM address and for each of the children one random number is used from the RND generator to select one mutation scheme which will be applied. In such mutation schemes, there are defined algorithms of mutation and there could be one, two, or several mutation points. The resulting genetic information can be mutated for example with following algorithms: Shifting of blocks, rotation of blocks, mirroring of blocks, shuffling of blocks, inversion of some bits, addition of some offset to some bytes, etc.[2.].

3.2 Cellular processor of logical functions

Basic construction unit of the cellular processor of logical functions is one cell $B_{n,k}$, which in cooperation with other cells forms one-dimensional cellular automaton with the absolute cell address range $\langle 0, n_{\max} \rangle$. Each of the cell $B_{n,k}$ is coded in 32 bits and carries the binary information:

- 1 bit: the last valid value calculated during the current step “k” in

$$y_{n,k} \in \{(0)_2, (1)_2\}, \quad (3)$$

- 1 bit: the previous value valid in previous step “k-1” in

$$y_{n,k-1} \in \{(0)_2, (1)_2\}, \quad (4)$$

- 11 bits: the relative link from the first other cell

$$a_n = \langle -(n_{\max}+1)/2, +(n_{\max}+1)/2 - 1 \rangle, \quad (5)$$

- 11 bits: the relative link from the second other cell

$$b_n = \langle -(n_{\max}+1)/2, +(n_{\max}+1)/2 - 1 \rangle, \quad (6)$$

- 8 bits: the logical function F_n coded by the 8-bit table (described by one byte)

$$F_n = [f_{n,0}, f_{n,1}, f_{n,2}, f_{n,3}, f_{n,4}, f_{n,5}, f_{n,6}, f_{n,7}] = \langle (0)_{10}, (255)_{10} \rangle. \quad (7)$$

During each of the iteration step „k“, all new output values of all cells $B_{n,k}$ are calculated from previous output value of cell $B_{n+an,k}$ (address of the cell A is calculated from $n + a_n$), previous output value of the cell $B_{n+bn,k}$ (address of the cell B is calculated from $n + b_n$) and the previous output value of the cell $B_{n,k}$ itself. All these three binary values are used as an address (0 ... 7) to get a new output value ($f_{n,0} \dots f_{n,7}$) of the cell $B_{n,k}$ from the 8-bit table F_n .

3.3 Data types used in the program

In this part of the paper, there are listed several examples of structures which are used in the program; to each of the item belongs a commentary:

```
VECTHEAD STRUCT                                ; header of the vector in RAM and in a file
    vhLength      DD ? ; sizeof(VECTHEAD) : 42
    vhPTRData      DD ? ; pointer to the start of data
    vhSignature    DB 27 dup(0) ; signature of the file - 27 bytes
    vhBytesPerCell DB ? ; Number of bytes per one cell : 4
    vhNumberOfCells DD ? ; Contains size of the vector in cells
    vhPassed       DB ? ; The member passed the selection? TRUE /FALSE
    vhABlimit      DB ? ; for limitation of A and B cell's pointers
VECTHEAD ENDS
```

```
GENERATIONHEAD STRUCT ; header of one generation in RAM and in a file
    vhLength      DD ? ; sizeof(GENERATIONHEAD) : 43
    vhPTRData      DD ? ; pointer to the start of data
    vhSignature    DB 27 dup(0) ; signature of the file - 27 bytes
    vhNumbOfVect   DD ? ; Number of Vectors in the generation
    vhNumbOfCells  DD ? ; Number of cells in one vector
GENERATIONHEAD ENDS
```

```
DEFHEAD STRUCT ; header of the definition in RAM and in a file
    vhLength      DD ? ; sizeof(DEFHEAD) : 45
    vhPTRData      DD ? ; pointer to the start of data
    vhSignature    DB 27 dup(0) ; signature of the file - 27 bytes
    vhNumbOfCases  DW ? ; Number of all definition cases
    vhNbInBytes    DW ? ; Contains number of input bytes
    vhNbOutBytes   DW ? ; Contains number of output bytes
    vhMaxPoints    DD ? ; FFFFFFFFh/MaxFitness
DEFHEAD ENDS
```

```
BINDATAHEAD STRUCT ;header of the RND, Selection, Repr.rules, ... in RAM and in file
    vhLength      DD ? ; sizeof(BINDATAHEAD) : 53
    vhPTRData      DD ? ; pointer to the start of data
    vhSignature    DB 27 dup(0) ; signature of the file - 27 bytes
    vhNumbOfRndBytes DD ? ; Number of all bytes used for random numbers
    vhNumbOfSelectBytes DD ? ; Number of all bytes used for the selection
    vhNumbOfRepRulesBytes DD ? ; Number of all bytes used for the repr. rules
    vhRndLastPosition DD ? ; position of last used RND
    vhMaxRND       DW ? ; information about range of random numbers
BINDATAHEAD ENDS
```

3.4 User interface

Several screenshots are taken from the program as a demonstration:

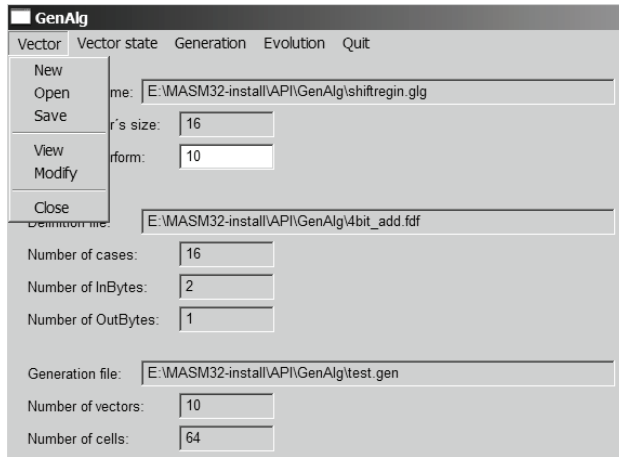


Figure 3: Vector files operations

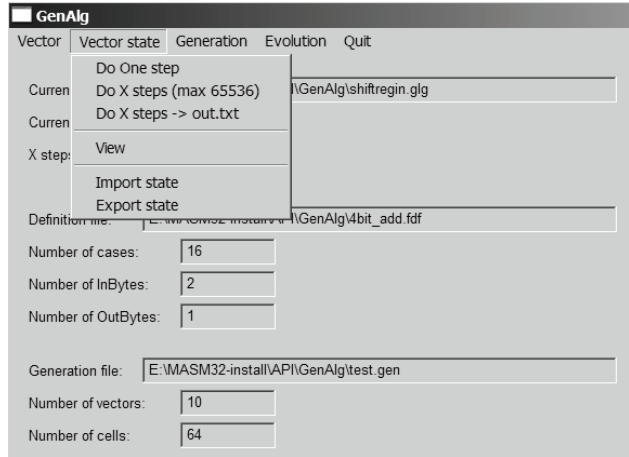


Figure 4: Vector states operations

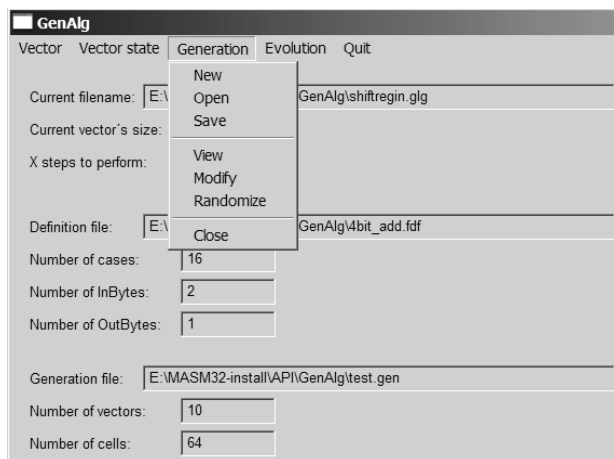


Figure 5: Generation files operations

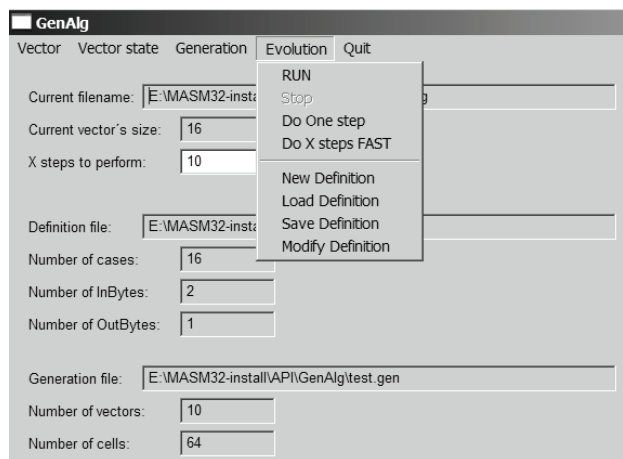


Figure 6: Evolution operations

4 Conclusion

The description of algorithms in this paper, used inside a win32 assembler program, goes not so deep into details as much more space would be needed to do it properly. There are just main ideas enlisted which may be used for development of a program to be capable to search automatically different kinds of algorithms. Thus is based on operation of the cellular processor of logical functions, linked to already known technologies for genetically programming. The main idea is to put all of these parts together and to design them to run fast and effective with use of the assembly language. The real performance statistics of the program cannot be provided at the

moment, the author is still working on the coding of the program and as soon as the coding will be finished, the program would be subject of planned thorough performance measurements.

References

- [1.] MAŘÍK V., ŠTĚPÁNKOVÁ O., LAŽANSKÝ J., ... : *Umělá inteligence (4)*. Academia, Praha, 2003.
- [2.] ZELINKA I., OPLATKOVÁ Z., ŠEDA M., OŠMERA P., VČELAŘ F.: *Evoluční výpočetní techniky, Principy a aplikace*. BEN - technická literatura, Praha, 2009.

Current address

Petr Skorkovský, Ing.

Areva NP o.s., JE Dukovany 269,
Dukovany 675 50,
tel. number: +420777084778,
e-mail: petrsk@centrum.cz

THE IMPLEMENTATION OF HYBRID ARIMA-NEURAL NETWORK PREDICTION MODEL FOR AGREGATE WATER CONSUMPTION PREDICTION

ŠTERBA Ján, (SK) , HILOVSKÁ Katarína (SK)

Abstract. Typical time series prediction methods used in many real-world applications – ARIMA models and Neural networks, achieve good prediction performance, however, both of them fits better the different type of time series. The ARIME models are generally better in prediction of linear time series, while Neural Networks are superior in predicting nonlinear time series. To create one, superior prediction method suited for prediction of general real-world time series, containing both linear and nonlinear parts, individual models can be hybridized to create single, ARIMA-Neural Network hybrid model. Using this approach, models can complement each other in capturing patterns and internal dependencies of time series. To examine the performance, proposed hybrid prediction method is used for prediction of water consumption based on time series collected from 1984 to 2007. As can be observed from results achieved from 12-step ahead prediction, the hybrid neural network outperforms the individual forecasting model.

Keywords. ARIMA-Neural Network, hybrid system, time series prediction

Mathematics Subject Classification: Primary 60G25; Secondary 62M20.

1 Introduction

The prediction of some real-world time series is a very important component for urban and industrial planning on both national and municipal level. Hence, it is critical to realize high-precision forecasting models to obtain reliable data. The state of the art of the time series forecasting methods in recent years can be divided into two areas. Ones are forecasting models based on traditional mathematical models, such as ARIMA model, Parametric Regressive model, Kalman filter model, Exponential Smoothing model, etc. Others are forecasting methods and models which does not pay attention to rigorous mathematical derivations and clear physical meaning, but emphasize on whether the model can fit the underlying relations of the investigated problems closely, including artificial neural network models, nonparametric regressive models, KARIMA algorithms, spectral basis analyses and others.

The accuracy of water consumption forecast has significant impact on planning and decision making of water facilities. Accurate water prediction is therefore important, especially with present rapid changes in water supplies and sources availability. Absolute necessity of exact water levels and water significance for industrial and municipal areas, and high costs for obtaining additional supplies makes accurate forecasts necessary. Therefore, it is important to search for models improving the performance of prediction and reducing the forecast error.

Box and Jenkins developed the autoregressive moving average to predict time series [1]. The ARIMA model is used for prediction non-stationary time series when linearity between variables is supposed. However, in many practical situations supposing linearity is not valid. For this reason, ARIMA models do not produce effective results when used for explaining and capturing nonlinear relations of many real world problems, what results in increased forecast error.

Artificial neural networks (ANN) models are part of an important class that has attracted a considerably attention in many applications. The use of ANN in many applied works is generally motivated by empirical results showing that under certain conditions, even simple ANN are able to approximate any measurable function to any degree [2-4]. As artificial neural networks are used as universal function approximations [5], they are very often used as to predict nonlinear time series [6,7]. Existing ANN models for forecasting generally use Multilayer Perceptron networks, which parameters - number of hidden layers, number of neurons in the layers and transfer function are often chosen through trial and error method with aim of finding the most feasible model for specific application [8].

However, the real-world time series problems are not absolutely linear or nonlinear – they often contain both linear and nonlinear parts. Furthermore, real time problems are often affected by irregularities and infrequent events, which make time series forecasting complicated and difficult [9]. Thus, using a single model for forecasting is not the best approach. Although both ARIMA and ANN models have achieved success in their own linear and nonlinear domains, neither ANN or ARIMA can adequately model and predict time series since the linear model cannot deal with nonlinear relationships, while ANN models alone are not able to handle both linear and nonlinear patterns equally well.

As a result, several researchers have proposed hybridizing ARIMA and ANN models, since different forecasting models can complement each other in capturing patterns of data set and time series. Both theoretical and empirical studies have revealed that a hybridization forecast outperforms individual forecasting models [7,10]. The merging of this structure can help the researchers in modelling complex structures in real-world time series more effectively. Moreover, by using ARIMA and ANN in single model can significantly assist in generating lower generalization variance of error.

To test the hybrid ARIMA-ANN prediction error on water consumption time series, data set consisting of 278 observation points was used, using aggregate water consumption ranging from January 1984 to June 2007.

The remaining of this paper is organized as follows. In section 2, description of ARIMA models is presented. Section 3 describes the hybrid ARIMA-ANN model and neural network used. Detailed discussion on the results is given in section 4, and finally in section 5 is a conclusion and projection of a future work.

2 ARIMA Models for Modelling of Time Series

ARIMA models, also known as Box-Jenkins models, are classical time series analyses method, which is being generally used for time series prediction. An Autoregressive Moving Average ARMA(p,q) is defined as

$$y_t(t) = \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=0}^q \theta_j \varepsilon_{t-j}, \quad (1)$$

where y_t is the time series value lagged by time moments $t=1,2,\dots,l$, the ϕ_i and the θ_j are the auto-regressive and moving averages model parameters, and the ε_t is purely a random process with zero mean and variance σ^2 .

One of the necessary conditions for applying ARMA model is the stationarity of the time series, which in practice, is very rarely met. For this reason, extension of ARMA model exist, which allows to apply model even on non-stationary time series, called autoregressive integrated moving average process (ARIMA). This extension transforms the time series by differencing them by the order of d , thus insuring stationarity of the time series. In other words, if the series y_t is non-stationary, but the d -th difference, $\Delta^d y_t = (1-B)^d y_t$, is stationary.

The ARIMA(p,d,q) model is then defined as

$$\phi(B)[(1-B)^d y_t - \mu] = \theta(B)\varepsilon_t \quad (2)$$

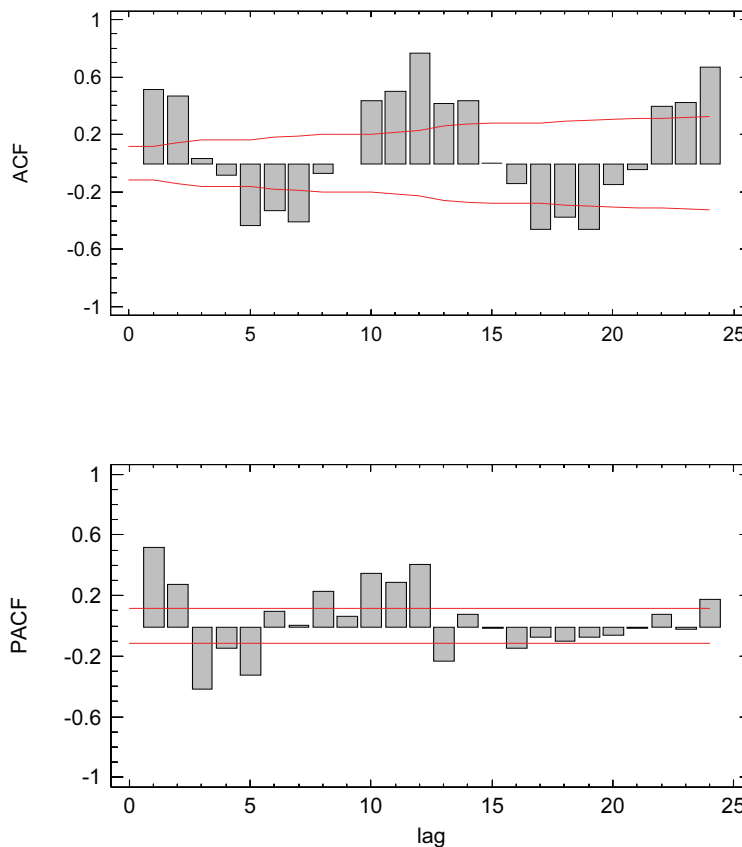


Figure 1. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of analyzed time series

Thus, ARIMA model of order (p,d,q) , often simply denoted as $ARIMA(p,d,q)$ can be used to predict the values of non-stationary time series.

Because real-world time series often fluctuates with seasonal patterns, the above mentioned ARIMA cannot model time series successfully, especially when seasonality presents a dynamic pattern. Therefore, $ARIMA(p,d,q)$ model can be extended, forming the seasonal ARIMA $(p,d,q)(P,D,Q)$ model of time series. For more information regarding seasonal ARIMA models of time series and their practical applications, see [1].

Process of evaluation of model parameters p, q, d , and seasonal parameters P, Q and D if they are present, is often referred to as Model Identification. The identification of ARIMA model is usually based on analyses of auto-correlation function (ACF) and Partial auto-correlation function (PACF), or alternatively it can be based on AIC criterion (Akaike Information Criterion) or FPE (Final Prediction Error) criterion. The PACF and ACF of the time series evaluated in this article can be seen in Fig. 1.

For more detail information regarding model identification, see [11].

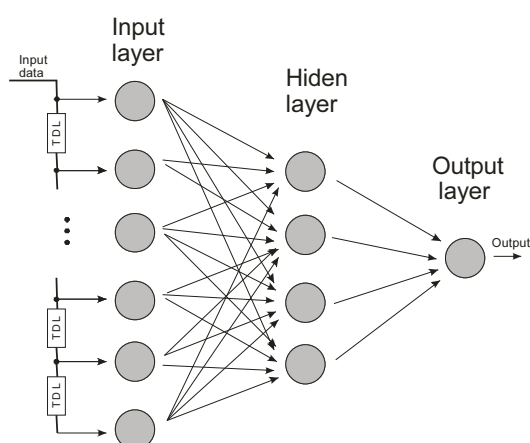


Figure 2. Artificial neural network with one hidden layer and tapped delay lines (TDL) at the input layer

TABLE I.
PERFORMANCE COMPARISON OF PREDICTION METHODS

Prediction method	RMSE	MAD	MAPE
ARIMA	8.1658×10^{-4}	5.4648	3.1638 %
ARIMA-NN	6.3163×10^{-4}	4.2099	2.4479 %

3 ARIMA-Neural Network Hybrid System for Time Series Prediction

Hybrid system is the combination of two or more than two systems in a one functioning system. Our hybrid system was obtained by combining neural networks with ARIMA time series model. Figure 2 demonstrates the framework of employed hybrid system.

The first step of investigated hybrid system involves usage of seasonal ARIMA model to model the linear part of the time series, and to create the ARIMA forecast. As the ARIMA model is based on linear relationship of system parameters, ARIMA can forecast linear relationships with high performance, but the performance often fail when the time series have non-linear relationship. On the other hand, the artificial neural networks have proven good performance when modelling non-linear relationships. For this reason, neural network is engaged in the following step to model the non-linearity, and all the remaining relationships which have not been absorbed by ARIMA model. Therefore in the second step, the ARIMA forecasts and times series data are used as inputs for artificial neural network, and trained using the known input and output training data to model the system responsible for creation of the time series. In the third and last stage, the neural network

is used to predict the future values of investigated time series 12-step ahead. As the network is trying to predict values in an interval out of known input values, the output of neural network from one time spot is used as an input of neural network in the following time spot. Thus, output from artificial neural network is used as estimate of time series value for the next forecast.

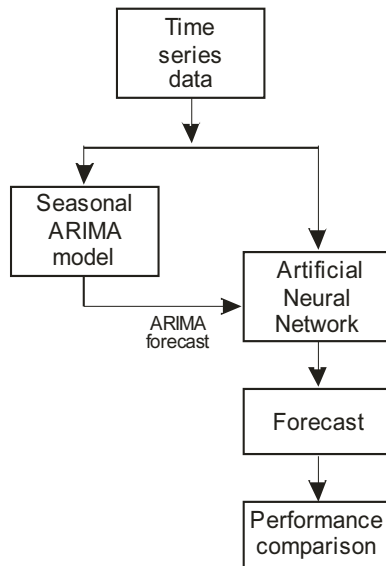


TABLE II.
PERFORMANCE OF BEST NEURAL NETWORKS

Number of neurons in input layer	Number of neurons in hidden layer	RMSE (validation set)	MAPE (validation set)
14	4	7.01 e^{+04}	2.447 %
5	2	7.03 e^{+04}	2.827 %
5	1	7.06 e^{+04}	3.316 %
4	1	7.08 e^{+04}	3.751 %
10	2	7.11 e^{+04}	4.127 %

Figure 3. Block type pilot arrangement

Three statistical tests are used to evaluate the performance of ARIMA and ARIMA-ANN hybrid models. These tests were realized using the well-known error functions as Root Mean Square Error (RMSE), Mean Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE). These tests are defined as

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (v_t - p_t)^2} \quad (3)$$

$$MAD = \frac{1}{n} \sum_{t=1}^n \frac{|v_t - p_t|}{n} \quad (4)$$

$$MAPE = \sum_{t=1}^n \left| \frac{v_t - p_t}{v_t} \right| \cdot \frac{100}{n} \quad (5)$$

where n is number of forecasting period, v_t is actual time series value at time t and p_t is the predicted value of time series.

Prior to prediction, the known time series data (Fig. 4) are first divided into training set, which is being used to train the neural network and ARIMA model, and into validation set, which is used only for evaluation of forecast performance, and thus data from validation set are not used as inputs in third, and all prior stages of system usage.

In investigated hybrid system, time-delayed feed-forward neural network was employed with one hidden layer, which general topology is illustrated in Fig. 3. Input data is send through set of tapped delay lines (TDL) into neurons in the first layer. The number of input neurons depends on the number of inputs and the values of input delays. Outputs from the neurons of input layer are

then connected with every neuron in hidden layer. Similarly, the outputs from neurons in hidden layer are connected to every neuron in output layer. As for a time series prediction, we are interested only in one forecasted value, therefore the number of neurons in the output layer is 1. The optimal number of neurons in hidden layer is obtained through trial and error method, e.g. using the computer simulations and choosing the neural network topology with the best performance.

Prior to prediction, near optimum network architecture should be found. In our experiments, three layered neural network has been used with different lengths of input data and number of neurons in hidden layer. Using different network architectures, we first trained all the networks using the back-propagation algorithm and then chose the network with best performance. The table 1 shows the results for few of the best neural networks.

4 Forecast Results

To show the efficiency of hybrid prediction model, the forecast performance is evaluated and compared with traditional ARIMA model. The time series data used for prediction shows significant seasonality tendency. For this reason, we modeled time series with seasonal ARIMA model, and determined the best model based on Akaike's information criterion as ARIMA (1,0,1) (1,1,1) 12 model. The output from ARIMA model was then used as input into artificial neural network. The optimal number of input delays of neural network and number of neurons in hidden layer was identified using the computer simulations and based on comparison of performance on validation data. The best neural network topology was identified as 14 4 1 topology, that is neural network with 14 neurons in the input layer, 4 neurons in the hidden layer and 1 neuron in the output layer. The neural network was trained using the back-propagation algorithm using the Levenberg-Marquardt optimization for updating weights and biases values, with tan-sigmoid transfer function in hidden layers and linear transfer function in the output layer. The lags {1,2, ... 14} were used for the input data. The performance comparison of prediction models is based on Root Mean Square Error (RMSE), Mean Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE). The results shown in Table 2 shows that hybrid model outperforms the ARIMA model. Figure 5 shows the forecast results for ARIMA and ARIMA-NN model, with real values of time series, and residuals of the models are shown in Fig. 6.

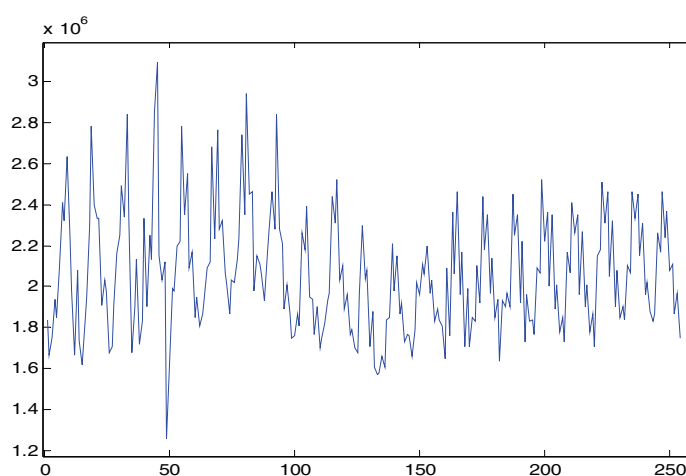


Figure 4. Time series data used for prediction

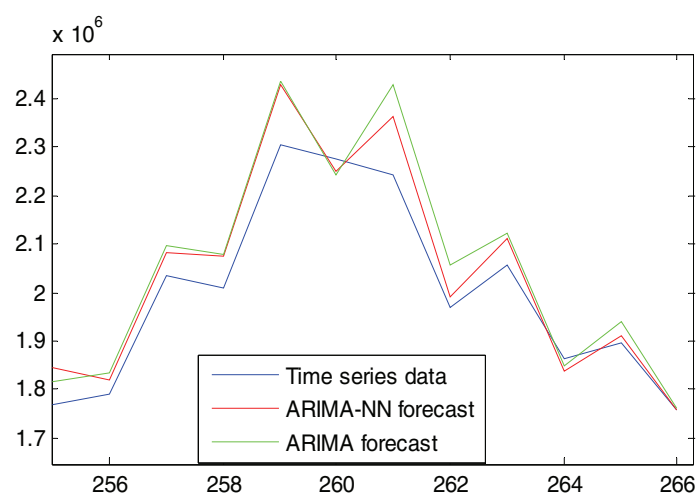


Figure 5. 12-step ahead forecast using ARIMA and ARIMA-NN hybrid system

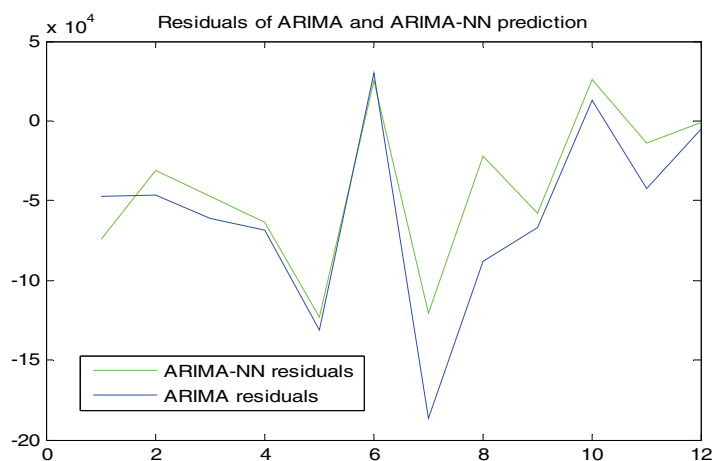


Figure 6. Residuals of ARIMA and ARIMA-NN hybrid system prediction

5

Conclusion

In this article we provided a forecast comparison of traditional seasonal ARIMA model and proposed ARIMA-NN hybrid model, based on time series data of aggregate water consumption of Spain. This case proved that ARIMA-NN hybrid prediction methods provide superior results than traditional ARIMA time series prediction model. The hybrid method takes advantage of the unique strength of ARIMA and NN in linear and nonlinear modelling of time series. The linear ARIMA model and nonlinear NN model are used jointly, aiming to improve the forecast prediction. For complex problems that have both linear and nonlinear relationships, the hybrid model can provide more accurate results. The results for water consumption prediction shows that this approach presents a superior and reliable alternative to traditional methods when choosing the appropriate number of delays and lagged input variables.

An extension to this work would be to investigate different alternative topologies of hybrid models and provide their performance comparison. In addition, the forecast models should be tested on different data sets in the future.

References

- [1] BOX, G.E.P., JENKINS, G.M.: *Time Series Analyses, Forecasting and Control*, San Francisco: Holden day, 1976.
- [2] CYBENKO, G.: *Approximation by superposition of sigmoidal functions*, Mathematics of Control, Signal and Systems, 2, 1989.
- [3] WHITE, H.: *Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings*, Neural Networks, 3, pp. 535-550, 1990.
- [4] GALANT, A.R., WHITE, H.: *On learning the derivatives of an unknown mapping with multilayer feedforward neural network*, Neural networks, 5, pp. 129-138, 1992.
- [5] HORNIK, K., STINCHCOMBE M., WHITE, H.: *Multilayer feedforward network are universal approximators*, Neural Networks, 2, pp. 359-366, 1989.
- [6] TANG, Z., FISHWICK P.A.: *Feedforward neural nets as models for time series forecasting*, Journal of Computing, pp. 374-385, 1993.
- [7] ZHANG, G.: *Time series forecasting using a hybrid ARIMA and neural network model*, Journal of Neurocomputing, 50, pp. 159-175, 2003.
- [8] GOMES, G.S.S., MAIA, A.L.S., LUDERMIR, T.B., CARVALHO, F., ARAUJO, A.F.R.: *Hybrid model with dynamic architecture for forecasting time series*, International Joint Conference on Neural Networks, pp.3742-3747, 2006.
- [9] SALLEHUDDIN, R., SHAMSUDDIN, S. M., ZAITON S., HASHIM, M.: *Hybridization Model of Linear and Nonlinear Time Series Data for Forecasting*, Second Asia International Conference on Modelling and Simulation, pp.597-602, 2008.
- [10] JAIN, A. & KUMAR, A.M.: *Hybrid neural network models for hydrologic time series forecasting*, Applied Soft Computing, Vol. 7, pp. 585-592, 2007.
- [11] AKAIKE H.: *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, 19, pp. 716-723, 1974.
- [12] SKOKAN, M., BUNDZEL, M., SINCAK, P.: *Pseudo-distance based artificial neural network training*, 6th International Symposium on Applied Machine Intelligence and Informatics, pp.59-62 2008.
- [13] DEHUI, Z., JIANMIN, X., JIANWEI, G., LIYAN, L., GANG, X.: *Short Term Traffic Flow Prediction Using Hybrid ARIMA and ANN Models*, Workshop on Power Electronics and Intelligent Transportation System, pp.621-625, 2008.

Current address

Ing. Ján Šterba

University of Economics in Bratislava, Department of Statistics, Dolnozemska cesta 1/b, 852 35 Bratislava, Slovak Republic, tel. number: +421 949 524 714,
e-mail: sterba.jan@gmail.com

Ing. Katarína Hil'ovská

Technical university of Košice, Faculty of Economics, Department of Banking and Investment, B. Nemcovej 32, 040 01 Košice, Slovak Republic, tel. number: +421 55 602 3263,
e-mail: katarina.hilovska@tuke.sk

STATISTICAL ANALYSIS II OF RESULTS OF PROJECT ESF STUDY SUPPORTS WITH PREVAILING DISTANCE FACTORS FOR SUBJECTS OF THE THEORETICAL BASE FOR STUDY

BOHÁČ Zdeněk, (CZ), DOLEŽALOVÁ Jarmila, (CZ), KREML Pavel (CZ)

Abstract. The project Study supports with prevailing distance factors for subjects of the theoretical base for study was dealt with at the VŠB – Technical University of Ostrava from January 5, 2006 to January 4, 2008. The submitted text evaluates the impact of study supports on the subject Algorithms and Data Structures if applied routinely.

Key words. E-learning, Mathematics, Evaluation

Mathematics Subject Classification: Primary 97U20, 97U40, 97U80; Secondary 28E10

1 Introduction

The project Study supports with prevailing distance factors for subjects of the theoretical base for study was dealt with at the VŠB – Technical University of Ostrava. More details are stated in [1]. In the framework of the project, twenty study supports were elaborated. They are available on website [7].

2 Study supports for the subject Algorithms and Data Structures

One of the first subjects for which electronic study supports were prepared and tested in pilot courses was the subject Algorithms and Data Structures.

The pilot course took place during the summer term of academic year 2006/2007 and was designed for first-year students in full-time form of study at the Faculty of Safety Engineering, further for students in combined form of study in Ostrava and at detached workplaces of the same Faculty in Prague and Most. Outputs of the project were available for the students of the target group on the web address www.studopory.vsb.cz, and for those whose access to the Internet was difficult, on CDs. Completed questionnaires were received from 459 respondents.

Results of exams after the completion of the pilot course showed a moderate increase in students' successfulness (57.6% in comparison with 53.5% in the previous academic year in full-time students and 50.7% in comparison with 47.1% in students in combined form of study). Although this was not a case of statistical evidential survey (reliable data from previous years are not available), it was possible to state (mainly with regard to the questionnaire survey) that study materials on Algorithms and Data Structures prepared in the framework of the project were a welcome teaching aid.

In the framework of project evaluation, a survey based on a questionnaire including ten questions as given below was performed.

1. In the course of study, I used these parts of electronic study text and/or CD
(more options may be circled)

- a) whole text
- b) explanation
- c) solved examples
- d) unsolved examples
- e) check questions
- f) check tests

2. How do you grade the CD (electronic study text)
1 (excellent) – 4 (unsatisfactory)

- a) whole text
- b) explanation
- c) solved examples
- d) unsolved examples.....
- e) check questions.....
- f) check tests

3. I consider the structuring of the study text to be (circle one option)

- a) very lucid
- b) lucid
- c) mostly lucid
- d) poorly lucid
- e) confused

4. Did you calculate unsolved examples? (circle one option)

- a) all
- b) about 75%
- c) about 50%
- d) about 25%
- e) less than 25%
- f) did not

5. You consider check tests to be (circle one option)

- a) very difficult
- b) difficult
- c) solvable
- d) easy

6. You consider the number of solved examples to be (circle one option)

- a) sufficient
- b) could be larger

- c) insufficient
7. State the number of hours you have devoted to study using the materials on CD and/or web pages (circle on option)
- 0-10
 - 10-25
 - 25-50
 - more
8. State the number of hours you have devoted to study using other materials (circle one option)
- 0-10
 - 10-25
 - 25-50
 - more
9. As for other study materials, you used (more options may be circled)
- printed textbooks
 - own notes of lectures
 - own notes of practicals
 - other sources; state them
10. I consider this subject to be (circle one option)
- very difficult
 - difficult
 - manageable with relatively small problems
 - easy

After pilot course conclusion, the study supports were put into routine use. In the framework of project sustainability we perform a questionnaire survey by means of the questionnaire, which was used for pilot course evaluation, in the case of subject Algorithms and Data Structures in the academic year 2008/2009. We received completed questionnaires from 229 respondents.

All the materials are evaluated by the students very positively, above all directly Delphi outputs. We shall get a picture of satisfaction of users with the study supports especially through answers to questions Nos. 1, 2, 3 and 6.

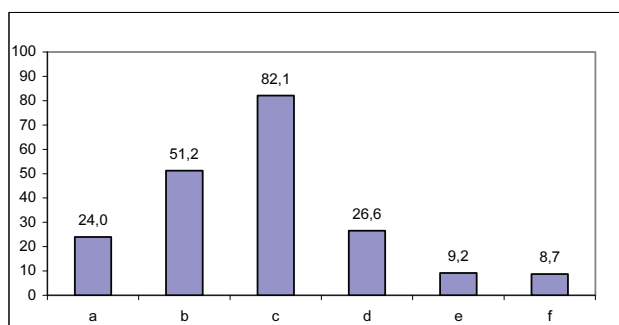


Fig. 1a Answers to question No. 1 in the academic year 2006/2007

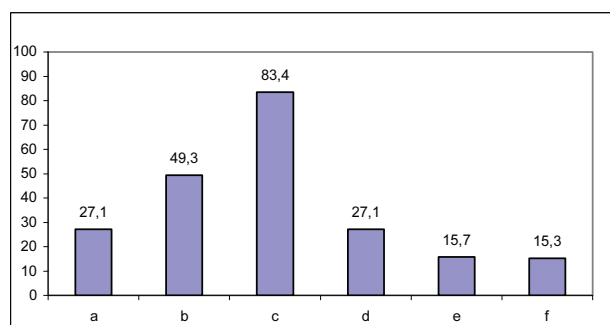


Fig. 1b Answers to question No. 1 in the academic year 2008/2009

From Figures 1a and 1b it follows that it is solved examples that are a “hit” of study supports. More than 80% of users prefer just this part of them. In the evaluation carried out by pilot course participants, a requirement for issuing a special collection of solved examples appeared

relatively frequently. On the basis of comments, a collection [5] was prepared and since the academic year 2008/2009 has been available for students in [6] as well.

In Figures 2a and 2b answers to question No. 2 are given. Individual parts of study supports are largely graded 1 and 2. A column designated 0 in the description of horizontal axis states the percentage of respondents who did not answer the question concerned.

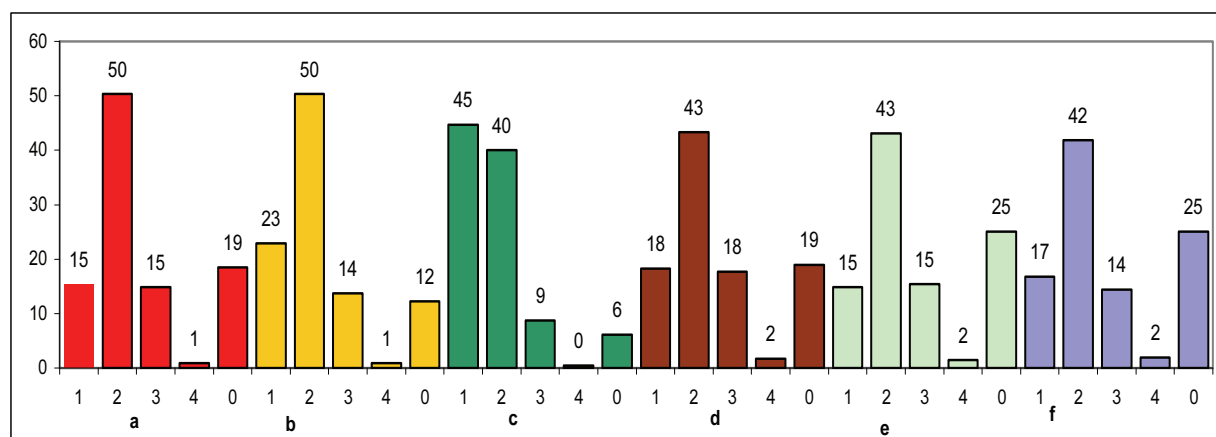


Fig. 2a Answers to question No. 2 in the academic year 2006/2007

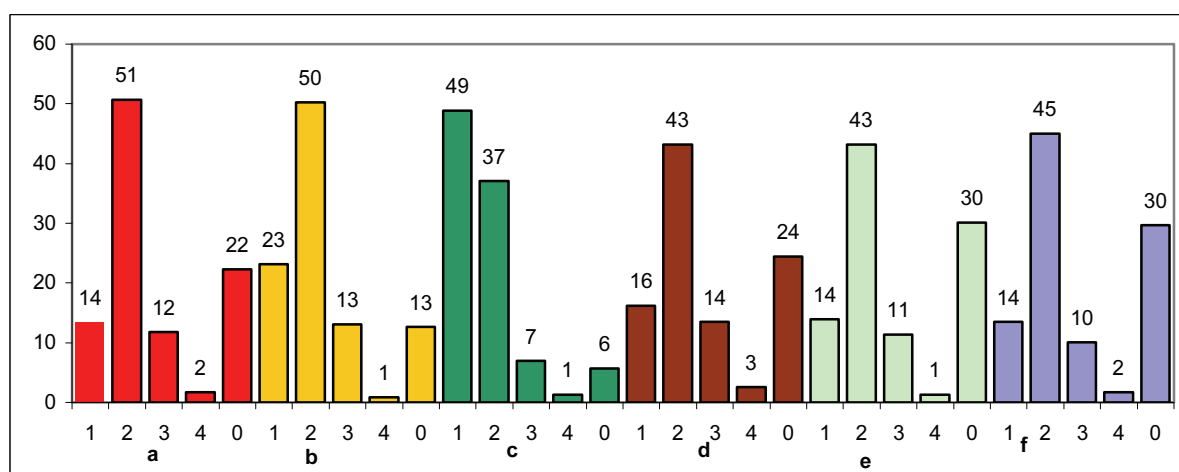


Fig. 2b Answers to question No. 2 in the academic year 2008/2009

In Figures 3a and 3b there are answers to question No. 3. In both the cases it is evident that more than 60% of respondents evaluate the study supports to be lucid or very lucid; if we include “mostly lucid” into the positive evaluation, more than 90% of respondents are satisfied with the structuring of the material.

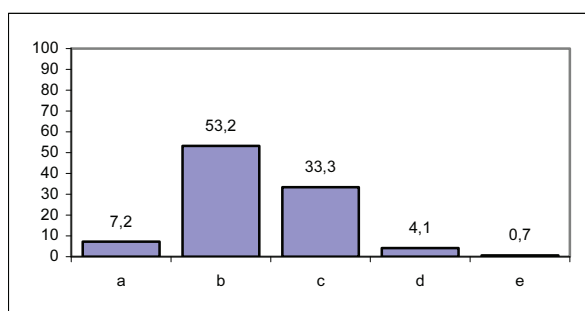


Fig. 3a Answers to question No. 3 in the academic year 2006/2007

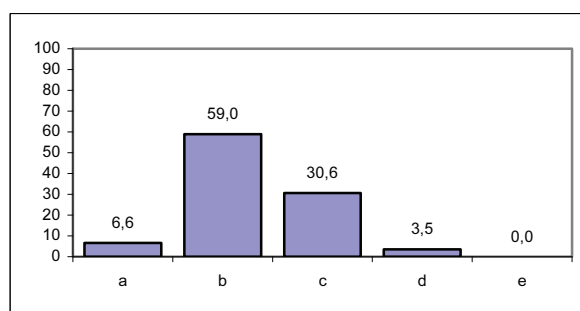


Fig. 3b Answers to question No. 3 in the academic year 2008/2009

3 Conclusion

The statistical comparison of results by means the two-sample Kolmogorov-Smirnov test shows that the evaluation of study supports on the part of users has not changed in time. A single exception is question No. 6 concerning the number of solved examples, where the test showed higher rating. This can be maybe explained by the fact that the study supports were after pilot course conclusion supplemented by a collection of solved examples.

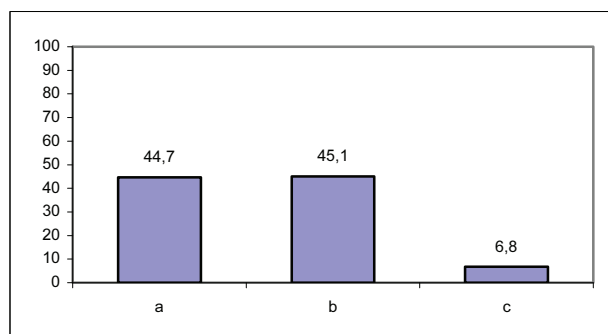


Fig. 4a Answers to question No. 6 in the academic year 2006/2007

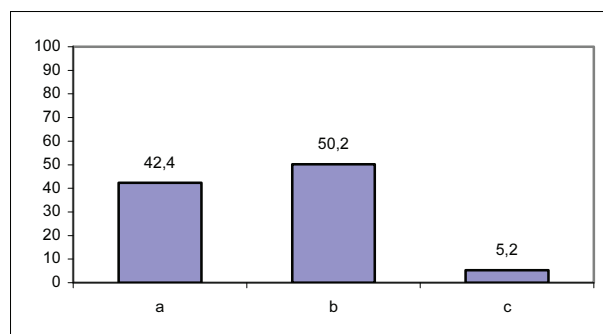


Fig. 4b Answers to question No. 6 in the academic year 2008/2009

Acknowledgement

The paper was prepared thanks to support provided by the project CZ.04.1.03/3.2.15.1/0016.

References

- [1.] BOHÁČ, Z., DOLEŽALOVÁ, J.: *Study support with prevailing distance factors for subjects of theoretical base for study*. In: *Zeszyty naukowe Politechniki Śląskiej, Górnictwo z. 279*, pp. 218-224, Polytechnika Śląska, Gliwice, 2007. PL ISSN 03729508.

- [2.] BOHÁČ, Z. – DOLEŽALOVÁ, J. – KREML, P.: *Projekt ESF Studijní opory s převažujícími distančními prvky pro předměty teoretického základu studia*. 7th International Conference APLIMAT, str. 911-916, Bratislava 2008, ISBN 978-80-89313-03-7.
- [3.] BOHÁČ, Z. – DOLEŽALOVÁ, J. – KREML, P.: *Využití elektronických studijních materiálů z matematiky v prezenčním a kombinovaném studiu*. 8th International Conference APLIMAT, str. 653-662, Bratislava 2009, ISBN 978-80-89313-31-0.
- [4.] BOHÁČ, Z., DOLEŽALOVÁ, J., KREML, P.: *Statistická analýza I výsledků projektu ESF Studijní opory s převažujícími distančními prvky pro předměty teoretického základu studia*. In: Sborník z 17. semináře Moderní matematické metody v inženýrství (3μ). VŠB-TU Ostrava, Ostrava, 2008. ISBN 978-80-248-1871-9.
- [5.] KRČEK, B., KOLOMAZNÍK, I.: *Algoritmy a datové struktury*. VŠB – TU Ostrava, Ostrava, 2007. ISBN 978-80-248-1306-6.
- [6.] BOHÁČ, Z., KOLOMAZNÍK, I.: *Algoritmy a datové struktury. Sbirka řešených příkladů*. VŠB – TU Ostrava, Ostrava, 2007. ISBN 978-80-248-1461-2.
- [7.] www.studopory.vsb.cz

Current address

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

DIFFERENCES IN REMEMBERING CALCULUS CONCEPTS IN UNIVERSITY SCIENCE STUDY PROGRAMMES

JUKIC Ljerka (HR)

Abstract: This study examines a retained university level of conceptual and procedural knowledge focusing mainly on derivatives and differentiation. A questionnaire, given two months after students were taught certain concepts from differential calculus, investigated whether the students are able to solve elementary tasks. Results are compared for five non-mathematics study programmes at one university in Croatia.

Keywords: university students, calculus course, conceptual and procedural knowledge, non-mathematics study programmes, derivatives, quantitative study

Introduction

Mathematical sciences are seen as fundamental to the economic and social well-being of nations ([6]). Calculus is one of the fundamental courses in mathematics. It serves not only to mathematicians as a basis for mathematical modelling and problem solving, but also has a great application in other fields such as in physics, chemistry, engineering ([10]). For a significant part of students who take mathematics courses at tertiary level, a transition from a secondary to a tertiary education presents a difficulty ([3, 9, 12,13]). Many problems that appear in this transition are linked with a new presentation of mathematics course, new ways of thinking on a higher level and are also connected with the lack of appropriate tools for learning mathematics ([9]).

Mathematics courses contain in great part topics regarding calculus, and many students who take calculus courses are not mathematics majors ([24]). Dahl [4] compared a level of competences that students should possess at a secondary and a tertiary education level. She found out that the level of mathematics competencies decreases from compulsory through to tertiary level. Hence, the transition from the secondary to the tertiary level mathematics reveals various challenges in terms of various knowledge forms. It shows that the nature of mathematics taught at the undergraduate level is different from the previous levels, either primary or secondary.

Background

Several studies had examined mathematical concepts of engineering students ([2, 9, 15, 16]). Studies compared mainly electrical and mechanical engineering students with mathematics students and showed that the non-mathematics students develop differently than the mathematics students, shaping their ways of thinking and understanding in a practical manner. Maull & Berry [15] concluded in their survey that lecturers were using words and ideas which mechanical engineering students interpreted in a different way from that the lecturers expected them to. Along with the students of electrical and mechanical studies, the students of other non-mathematics study programmes experience similar problems concerning undergraduate mathematics.

This study examines a level of students' knowledge related to the calculus course in various non-mathematics study programmes at one university in Croatia. The goal of this study is to investigate students' understanding and a retention of the selected number of concepts in the differential calculus and a comparison between non-mathematics study programmes.

One of the widely accepted categorization of mathematical knowledge is into *conceptual* and *procedural* knowledge. Conceptual knowledge provides an understanding of the principles and relations between pieces of knowledge in a certain domain, and procedural knowledge enables us to solve the problems quickly and efficiently ([11]). Many studies have investigated conceptual and procedural knowledge mainly on the primary school level (e.g. [19]). The structure of the university students has changed in last 20 years and the university classes do not contain only the best high school leavers, but also the students who diverse in many ways such as age, experience, socio-economic status and cultural backgrounds ([1]), thus studies have been emerging with an intention to investigate the tertiary education in a deeper way. Several studies examined the university level of conceptual and procedural knowledge with an emphasis on further exploration of deficiencies in students' understanding of calculus concepts ([5, 7, 8, 14, 17]). The results, that the new studies would provide, will be beneficial for further teaching, especially for calculus lecturers who should consider those implications to attain the goals of their courses as they prepare students for careers in science, mathematics, technology, and engineering ([8, 14]). Also, broadening the conception of procedural knowledge could lead to the better understanding of both surface and deep conceptual and procedural knowledge, and could broaden the ways of studying and assessing students' understanding ([21, 22]).

This study, therefore, examines a retained university level of conceptual and procedural knowledge in various non-mathematics study programmes, focusing mainly on topics concerning differentiation and derivatives.

Methodology

This survey was conducted at one university in Croatia in the academic year 2008/2009. The study was carried out through a questionnaire that was given to the first year students of five non-mathematics study programmes. 227 students participated in the survey. Students were enrolled in electrical engineering (EE), civil engineering (CE), food technology (FT), physics (P) and chemistry (C). The survey was conducted two months after the students were taught derivatives and were tested in concepts from differential calculus.

The questionnaire was carried out before exercise lessons in another mathematics course at each surveyed study programmes and was not announced, therefore the participants were those students

who came to the exercise lessons. Participating in the questionnaire was voluntary, so some students did not want to respond to some questions. Overall response rate was quite high, since 93% of surveyed students answered all questions. The questionnaire was given first to the chemistry students as a pilot. The chemistry students found no ambiguities in presented questions, thus questionnaire was given to the students of other study programmes. The questionnaire had four questions related to the topic of derivatives. The first question was a theoretical one, examining a definition of a geometric interpretation of a derivative of a certain function, and the other three questions were practical ones, asking students to differentiate the given function. The practical questions were sorted from easier to harder.

Analysis and Results

Question 1 was concerned with the geometric interpretation of the derivative of the function $f : R \rightarrow R$ at the point x_0 . The definition of the geometric interpretation of the derivative is one of the basic definitions and carries the motivation for the definition of the derivative. This question examined students' knowledge of the basic concepts two months after the students were taught derivatives. From the total number of surveyed students, 213 students answered this question. The following answers were offered:

1. the maximum/minimum of the function f at the given point (MAXMIN)
2. the slope of the tangent line to the curve $y = f(x)$ at the given point (TANGENT)
3. the continuity of the function f in the given point (CONT)
4. none of the above (N)

The correct answer to this question is coded as TANGENT. All answers were chosen from the author's personal experience teaching the course and marking the written exams for four years, and related to similar options that students wrote in their exams when asked to explain the geometric interpretation of the derivative of the function $f : R \rightarrow R$ at the point x_0 .

If we look at the distribution of the answers across the study programmes, all surveyed chemistry students gave the correct answer to the question. 50% of the electrical engineering students gave the correct answer and 29% choose the answer MINMAX. 45% of the civil engineering students gave the correct answer and 30 % have chosen the answer CONT. Only 29% of the food technology students selected the correct answer and 40% have chosen the answer CONT. 43% of the physics students gave the correct answer, and 40 % selected the answer CONT. The distribution of the selected answers for each surveyed study programme can be seen in the following table:

Study Programme	MAXMIN (%)	TANGENT (%)	CONT (%)	N (%)	Row Totals
CE	13.5	44.9	29.2	12.4	89
EE	27.9	50.8	11.5	9.8	61
P	0	43.5	39.1	17.4	23
FT	21.4	28.6	39.3	10.7	28
C	0	100	0	0	12

Table 1. A distribution of answers across study programmes for Question 1

If we present the data from the table above in a graphical form, we can see very clearly a domination of the chemistry students in Question 1.

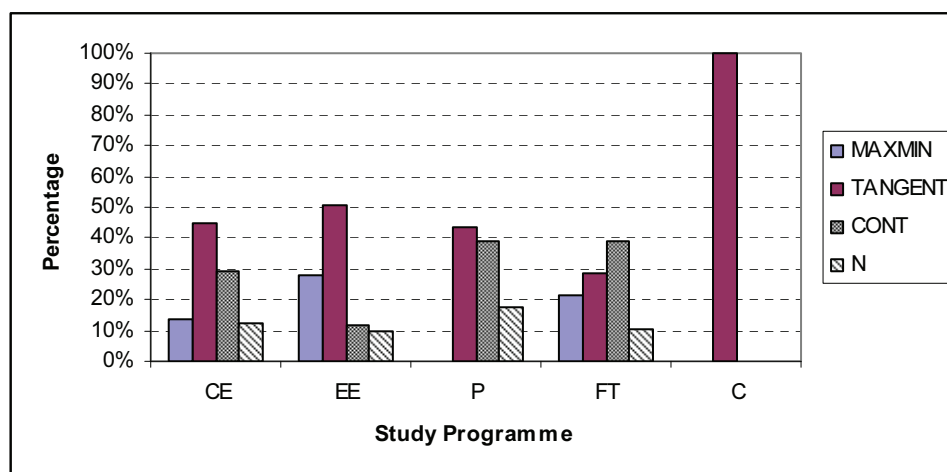


Figure 1. Answers across study programmes for Question 1

Question 2 asked the students to differentiate the function $f(x) = \frac{x^2 + 2}{x^2}$ and then to choose the correct answer. The answers that were offered beside the correct one were carefully chosen with anticipating mistakes. One answer contained an error in the numerator of derivative, and the other contained an error in the denominator of derivative. 220 students answered this question. The most successful were the chemistry students where 83% of surveyed students gave the correct answer. 78% of the civil engineering students answered correctly to this question. The electrical engineering, food technology and physics students had similar results in giving correct answers. Between 71% and 74% of surveyed students gave the correct answer. This can be seen from the following table:

Study Programme	DERt (%)	DERd (%)	DERn (%)	Row Totals
CE	77.8	12.2	10.0	90
EE	73.8	16.4	9.8	61
P	70.8	12.5	16.7	24
FT	72.7	6.1	21.2	33
C	83.3	8.3	8.3	12

DERt -code for the correct answer, DERd - code for a response in which the denominator of the differentiated function has an error, DERn - code for a response in which numerator of the differentiated function has error.

Table 2. A distribution of answers across study programmes for Question 2

If we present the data from the table above in the graphical form, we can better visualize a difference between the correct answer and the other answers selected by the students at the surveyed study programmes:

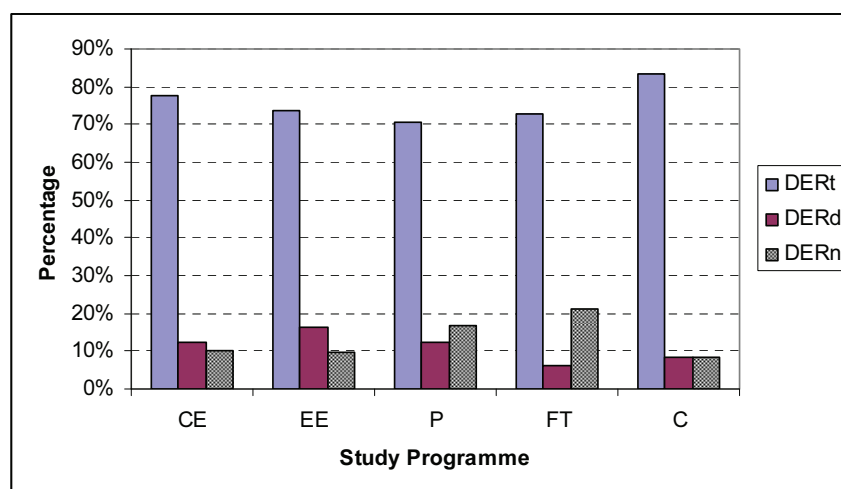


Figure 2. Answers across study programmes for Question 2

Question 3 asked the students to differentiate the function $f(x) = \sin^2 6x$. This question examined if the students know how to apply a chain rule for differentiation two months after the learning process was finished. As in the previous question, the answers that were offered beside the correct one were chosen with the anticipating mistakes. Certain steps in the chain rule of differentiation were omitted such as the differentiation of the argument of the sine function and the differentiation of the sine function itself. This question answered 218 students. The chemistry students were the most successful in answering this question where 90% of them gave the correct answer. The civil engineering students showed very bad results. Only 55% of them gave the correct answers. The correct answers for the electrical engineering, physics and food technology students are ranging between 70% and 80%. The distribution of the other answers for each study programme can be seen in the following table:

Study Programme	CHAINt (%)	CHAINsine (%)	CHAINarg (%)	Row Totals
CE	55.4	15.2	29.4	92
EE	71	16.1	12.9	62
P	78.2	8.7	13.1	23
FT	71	6.4	22.6	31
C	90	10	0	10

CHAINt = correct answer, CHAINsine = omitted differentiation of sine function,
CHAINarg = omitted differentiation of the argument of sine function

Table 3. A distribution of answers across study programmes for Question 3

Presenting the data above in the graphical form gives better visualisation of the difference between the correct answer and the other answers selected by the students at the surveyed study programmes. Also, it shows the evident domination of chemistry students' knowledge.

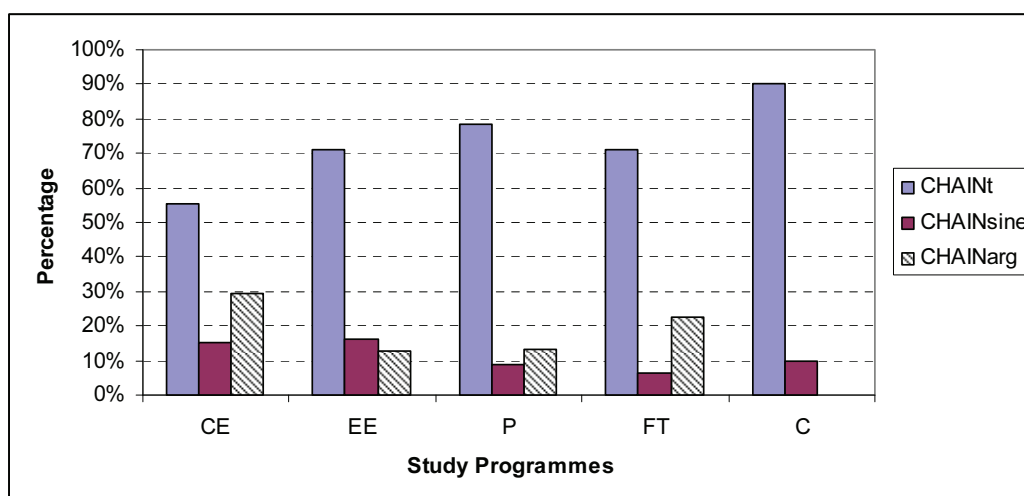


Figure 3. Answers across study programmes for Question 3

Question 4 was connected with the application of the derivative as the slope of the tangent line. For the given function $f(x) = (3x)^2$, the students had to calculate the slope of the tangent line to the curve $y = (3x)^2$ at the point $x = 1$. First, the students had to detect that the function has a simpler form, $f(x) = 9x^2$. Then they had to differentiate the function and finally insert the value for x into the derivative $f'(x)$. Among offered answers to this question was “9”. Students could get this number if they squared the given expression $(3x)^2$ as $9x$, and then continued with the differentiation. The other wrong answer was “6”, which students could get by differentiating the expression $(3x)^2$ as $6x$ and inserting the value for x . This question answered 217 students. The analysis of the students’ answers shows very poor knowledge and understanding of this question. All surveyed chemistry students gave the wrong answer to this question. Also 92% of the civil engineering students gave the wrong answer. The physics and food technology students were a bit more successful. Around 15% of them gave the correct answer. The electrical engineering students showed better results where 30% of them answered correctly. The detailed results can be seen from the table below:

Study Programme	ERRD (%)	NERR (%)	ERRS (%)	Row Totals
CE	76.4	7.9	15.7	89
EE	39.7	30.2	30.2	63
P	69.6	13.0	17.4	23
FT	37.5	15.6	46.9	32
C	30.0	0	70.0	10

ERRS- error in squaring, NERR-correct answer, EERD- error in differentiation

Table 4. A distribution of answers across study programmes for Question 3

The graphical representation of the data from the table above shows better ratio of the correct answer and the wrong answers. We can see more clearly that the chemistry students have chosen only the wrong answers, while the electrical engineering students selected all offered answers almost equally.

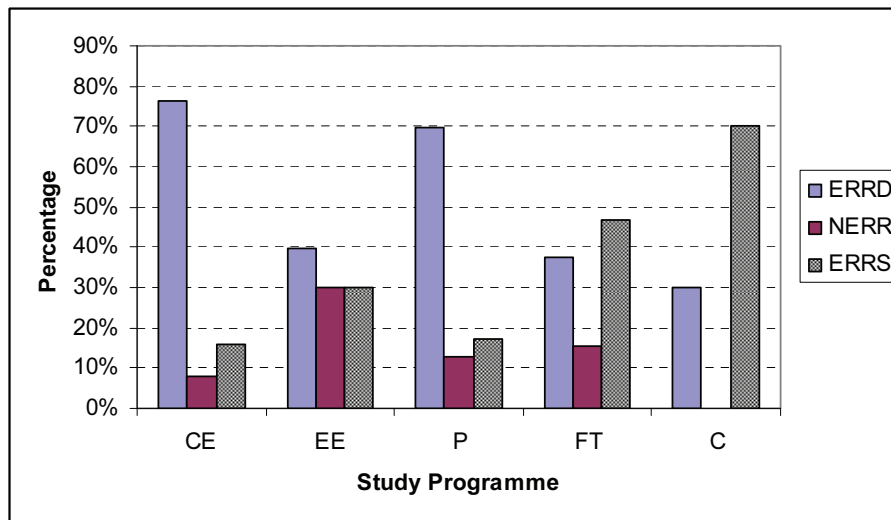


Figure 4. Answers across study programmes for Question 4

Discussion

Although this survey was given to the students of various non-mathematics study programmes, the tasks that were given examined basic knowledge from differential calculus. Students showed better procedural than conceptual knowledge two months after they were taught derivatives. This can be seen from results in the Question 1, 2, 3 & 4. Students were more successful in solving the Question 2 & the Question 3, than the Question 1 & the Question 4. In order to solve the Question 2 and the Question 3 students only had to apply the rules for differentiation. These questions tackled students' procedural knowledge. In order to solve the Question 4 it was not sufficient just to use some procedure, but also to connect several concepts in order to get the correct answer. The Question 1 and the Question 4 are strongly connected, and those questions tackled students' conceptual knowledge.

If we compare the results between the study programmes, we can see that the chemistry students demonstrated better procedural knowledge than the students of the other non-mathematics study programmes. All surveyed chemistry students gave the correct answer to Question 1, 90% of them gave the correct answer to the Question 2 and 83% to the Question 3. On the other hand, they showed bad results in Question 4. No one answered this question correctly. The electrical engineering students showed best results in the Question 4, but these results are considerably low, as well as their results in the Question 1, so we can not claim that they demonstrated better conceptual knowledge than the students of the other surveyed study programmes. Our results are showing that students are forgetting a subject matter really fast. Since the questionnaire was conducted only two months after the learning process was finished, we can not be satisfied with the obtained results. One could argue whether the given results are showing that students are forgetting

really fast or they just have never learnt the subject matter tested in the questionnaire. We can rule out the latter, because all the questions, theoretical and practical, have appeared in their tests.

The first year courses have “overloaded” syllabus quite often, as a compensation for reductions at the secondary level mathematics ([3]). Also, an overloaded syllabus is one factor among several situational factors that causes learning only the rules for solving computational problems ([20]). Guzman et al. [9] emphasized the need to decrease a quantity of the covered content and to engage students in deeper knowledge. Therefore, we can speculate that the overloaded syllabus leads to procedural knowledge. The surveyed study programmes have similar calculus course syllabus, but not completely the same, thus, it is worth to investigate their content, since the chemistry students demonstrated better knowledge than the students from the other surveyed study programmes.

It is important to understand the mathematical theory and as well, to possess good computational skills, since industries and businesses worldwide make demands for employees who can investigate and make predictions in a wide range of problems ([6]). In order to get time for these activities, this will imply the reduction in overloaded syllabus in most cases ([18]).

Acknowledgements

Thanks to my colleagues, teaching assistants, for help with the data collection, and The National Foundation for Science, Higher Education and Technological Development of the Republic of Croatia for funding. Thanks to Bettina Dahl Soendergaard for good advices.

Reference

- [1] BIGGS, J. (2003). *Teaching for Quality Learning at University*, Buckingham: Open University Press, Second edition
- [2] BINGOLBALI, E., MONAGHAN, J., & ROPER, T. (2007). Engineering students' conceptions of the derivative and some implications for their mathematical education, *International Journal of Mathematical Education in Science and Technology*, 38(6), 763-777
- [3] BRANDELL, G., HEMMI, K., & THUNBERG, H. (2008). The Widening Gap--A Swedish Perspective, *Mathematics Education Research Journal*, 20(2), 38-56
- [4] DAHL, B. (2009). Transition problems in mathematics that face students moving from compulsory through to tertiary level education in Denmark: Mismatch of competencies and progression, in *Proceedings, the 33th Conference of the International Group for the Psychology of Mathematics Education (PME33)*, 2, 369-376
- [5] CHAPPELL, K. & KILLPATRICK, K. (2003). Effects of concept-based instruction on students' conceptual understanding and procedural knowledge of calculus, *PRIMUS*, 13 (1), 17 – 37
- [6] CHINNAPPAN M., DINHAM S., HERRINGTON A., & SCOTT, D.(2007). Year 12 students and Higher Mathematics: Emerging issues, *AARE 2007 International education research conference*
- [7] ENGELBRECHT, J., HARDING, A., & POTGIETER, M. (2005). Undergraduate students' performance and confidence in procedural and conceptual mathematics, *International Journal for Mathematics Education in Science and Technology*, 36(7), 701- 712, 2005

- [8] GRUNDMEIER, T. A., HANSEN, J., & SOUSA, E. (2008). An exploration of definition and procedural fluency in integral calculus, *PRIMUS*, 16 (2), 178 -191
- [9] GUZMAN, M., HODGSON, B., ROBERT, A., & VILLANI, V. (1998). Difficulties in the passage from secondary to tertiary education , *Documenta Mathematica, Extra Volume ICM 1998*, 3, 747-762
- [10] HELFGOTT, M. (2004). Five guidelines in the teaching of first-year calculus, *Proceedings of the 10th International Congress of Mathematical Education*, Copenhagen, Denmark
- [11] HIEBERT, J., & LEFEVRE, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis, in J. Hiebert (Ed.), *Conceptual and Procedural Knowledge: The Case of Mathematics*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1–27
- [12] HING Sun Luk (2005). The gap between secondary and university mathematics, *International Journal of Mathematics Education in Science and Technology*, 32(2), 161-174
- [13] KAJANDER, A., & LOVRIC, M. (2005). Transition from secondary to tertiary mathematics: McMaster University Experience, *International Journal of Mathematical Education in Science and Technology*, 36(2-3), 149-160
- [14] MAHIR, N. (2006). Conceptual and procedural performance of undergraduate students in integration, *International Journal of Mathematical Education in Science and Technology*, 40(2), 201 - 211
- [15] MAULL, W., & BERRY, J. (2000). A questionnaire to elicit concept images of engineering students, *International Journal of Mathematical Education in Science and Technology*, 31(6), 899-917
- [16] MORGAN, A.T. (1990). A study of difficulties experienced with mathematics students in higher education, *International Journal of Mathematical Education in Science and Technology*, 21(6), 975-988
- [17] PETERSSON, K., & SCHEJA, M. (2008). Algorithmic contexts and learning potentiality: a case study of students' understanding of calculus, *International Journal of Mathematical Education in Science and Technology*, 39(6), 767 - 784
- [18] RUMP C., JAKOBSEN A., & CLEMMENSEN T. (1997). Improving Conceptual Understanding Using Qualitative Tests, 6th *Improving Student Learning Symposium*, in proceedings: C. Rust and G. Gibbs (ed.): *Improving Student Learning – Improving Student Learning Outcomes*, The Oxford Centre for Staff and Learning Development, Oxford
- [19] SCHNEIDER, M., & STERN, E. (2005). Conceptual and procedural knowledge of a mathematics problem: Their measurement and their causal interrelations, 27th *Annual Meeting of the Cognitive Science Society (CSS)*, Stresa, Italy
- [20] SKEMP, R. R. (1976). *The Psychology of Learning Mathematics*, Hillsdale, NJ: Lawrence Erlbaum Associates
- [21] STAR, J. R. (2000). On the relationship between knowing and doing in procedural learning. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Proceedings of fourth international conference of the Learning Sciences*, NJ: Lawrence Erlbaum, 80-86
- [22] STAR, J. R. (2005). Re-conceptualizing procedural knowledge, *Journal for Research in Mathematics Education*, 36(5), 127-155
- [23] SWELLER, J. (1999). *Instructional design in technical areas*, Melbourne: ACER Press
- [24] TALL, D., SMITH, D. & PIEZ, C. (2008). Technology and calculus, In M. Kathleen Heid and Glendon M Blume (Eds), *Research on Technology and the Teaching and Learning of Mathematics, Volume I: Research Syntheses*, 207-258

Current address

Ljerka Jukic, PhDstudent
Department of Mathematics, University of Osijek
Gajev trg 6, 31 000 Osijek, Croatia
tel. +38531224800
e-mail: ljukic@mathos.hr

THE USE OF INFORMATION TECHNOLOGY IN MATHEMATICS EDUCATION

ORSZÁGHOVÁ Dana, (SK)

Abstract. Modern education is associated with information and communication technologies that allow us to create electronic learning courses and then apply the e-learning method in the study. Electronic education has application in mathematical subjects, as well. Mathematical theory, a lot of tasks and applications can be found via the Internet, created in the electronic version from the simple text to the multimedia applets. In the paper we present some possibilities of applying means of information technologies in the study of mathematics subjects at the Faculty of Economics and Management of the Slovak University of Agriculture in Nitra.

Key words. Mathematical education, information technology, e-learning, LMS MOODLE

Mathematics Subject Classification: 97U50

1 Introduction

Requirements for data processing and information in electronic form have been motivation for the creation and development of the efficient tools of computing technology. Means of electronic communication and data transfer have become an important part of this development. Terms like Internet and information technology are rapidly becoming part of the working vocabulary in all areas. Moreover, the possibility of the Internet finds application in the everyday life of households. Therefore, even university education is no longer possible without information technology (IT). Education is under the direct influence of IT because of creating special educational tools and resources. Teaching without electronic educational materials is not attractive for students. Creating different kinds of electronic materials and using IT have become the part of the teacher's work.

It is possible to apply multimedia to the teaching of mathematics in different fields and options:

- Obtaining of mathematical knowledge by attractive means and ways,
- Writing of multimedia study materials,
- Usage of electronic interactive study books,
- Usage of World Wide Web,
- Testing by computers.

Education in general is in the process of transformation. We can see other possibilities and changes in mathematical education:

- The intensification of interdisciplinary relations,
- Extension of the teaching of mathematical subjects in a foreign language,
- Compatibility of mathematical education in European and international context,
- Teachers' study stays,
- Students' study stays,
- Interchange of study and professional literature [7, Országhová, Pokrivčáková, 2003].

Use of IT in education has brought new requirements – new educational competencies, that are important for the students and teachers.

The process of comprehensive integration of information and communication technology (ICT) into taught subjects placed high demands on the teacher, because it requires:

- appropriate and productive way to use new technologies to help achieve educational goals of syllabus,
- teacher should ensure the appropriate development of information literacy of their pupils and students to develop productive skills, independent and effective use of ICT.

In order for teachers to successfully integrate ICT into their subject:

- they should be familiar with effective methods for teaching their subject matter using ICT,
- teachers should know how to achieve their corporate goals using ICT,
- they should effectively use ICT for their preparation, teaching and administration,
- teachers should be able to assess the level of information literacy of their pupils and students, to know and develop it further [2, Božíková, 2004].

The Internet provides great opportunities in terms of information access, yet it also offers possibilities for directed study that effectively infer the learner's needs and desires. The fundamental prerequisite for this is creation of high-quality e-course incorporating LMS features in accordance with didactics of teaching and learning aids creation [1, Baraníková, 2009].

The important role in mathematical education and e-learning has software products. These include very well known Mathematica, Matlab, MathCAD, GeoBebra and others. Mathematical software GeoGebra combines geometry, algebra and mathematical analysis. Excluding construction and metric geometry options, we can also find in this program a wide range of applications in algebra or mathematical analysis [3, Drábeková, Rumanová, 2009].

2 E-learning and electronic educational courses

The advantage of electronic interactive study materials is accessibility for students without time limit. Both teachers and students consider e-learning as a method that has application in the study of mathematical objects.

In 2005 the LMS MOODLE was implemented into the education at the Faculty of Economics and Management (FEM) of the Slovak University of Agriculture (SUA) in Nitra. The system MOODLE contains variety of tools that allow the teacher creating modules for different subjects. Some of them will be briefly described and characterized in the following text:

- forum (it can be used for the purposes of communication, teaching or open),
- news (covers current information),
- assignment (any instruction given by the teacher to the students or the task outcome – a file sent to the teacher for review),

- quiz (the teacher creates, uses or modifies a database of test questions categorized into various categories from which the on-line tests are generated using random questions),
- journal (originated from the essays writings, suitable for teaching languages),
- resource (any teaching material in the form of a plain text, HTML, URL, attached file or other hyperlink to web site or reference material),
- survey (provides a number of verified survey instruments, which have been found useful in assessing and stimulating learning in online environments. Teachers can use these to gather data from their students that will help them learn about their class and reflect on their own teaching) [6, Országhová, Gregáňová, Majorová, 2007].

The Department of Mathematics created study materials for full-time students and for students of distance learning. Study materials are available for students at the web pages of the faculty:

<http://moodle.uniag.sk/fem/>

Exercises in mathematics – Winter Term (Cvičenia z matematiky ZS)

<http://moodle.uniag.sk/fem/course/view.php?id=44>

Exercises in mathematics – Summer Term (Cvičenia z matematiky LS)

<http://moodle.uniag.sk/fem/course/view.php?id=143>

Linear algebra (Lineárna algebra) (for illustration see Fig. 1)

<http://moodle.uniag.sk/fem/course/view.php?id=207>

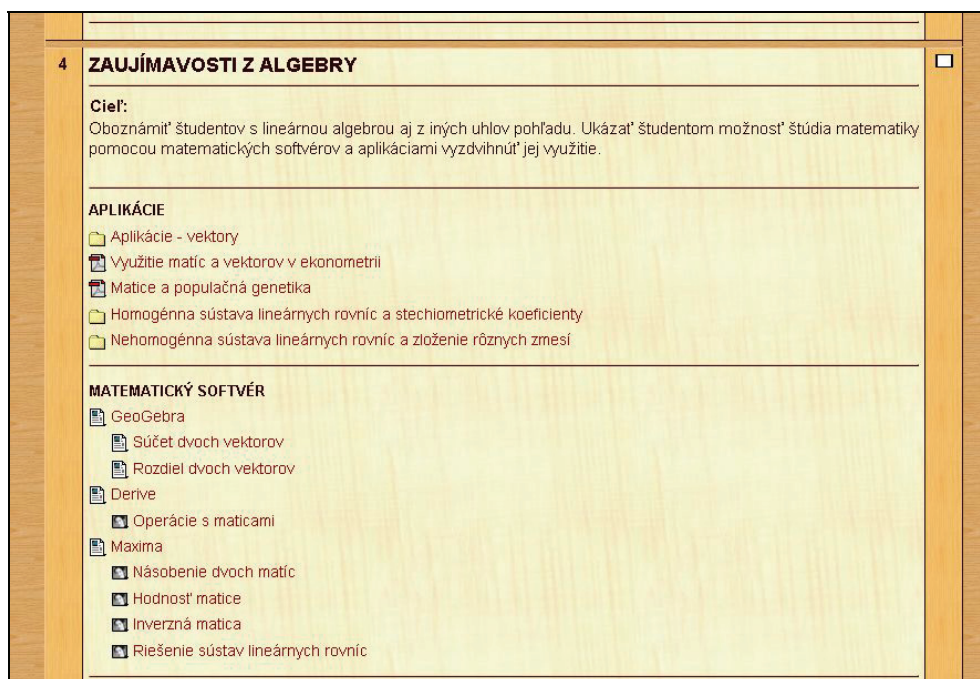


Figure 1: Web page of the course Linear algebra

Courses *Mathematics I* and *Mathematics II* were created for students of distance learning – the entrance is able only with the key. An illustration of the topic “Indefinite Integral” from the course *Mathematics II* is displayed in the Figure 2.

The advantages of e-learning courses available through the website include:

- creator (teacher) can update the site at any time,
- relatively inexpensive repair of typing or other mistakes in comparison with issuing of printed materials,

- students have unrestricted access to sites and their use to study,
- direct use in teaching sites in the auditorium with Network Connection,
- use for individual work of students (seminary work, tasks, etc.) [5, Országhová, 2008].

Students evaluated the access to electronic materials in mathematics via the Internet very positively. Moreover, this is one form of the communication students with mentor at distance learning. Those students who actively work with study materials achieve better academic results. They have enough information regarding the study of math and prepare for the test. In the course there are different kinds of materials that can students print and use them in paper form. By creating these electronic educational courses, we wanted to support and develop the students attitude to self-education.

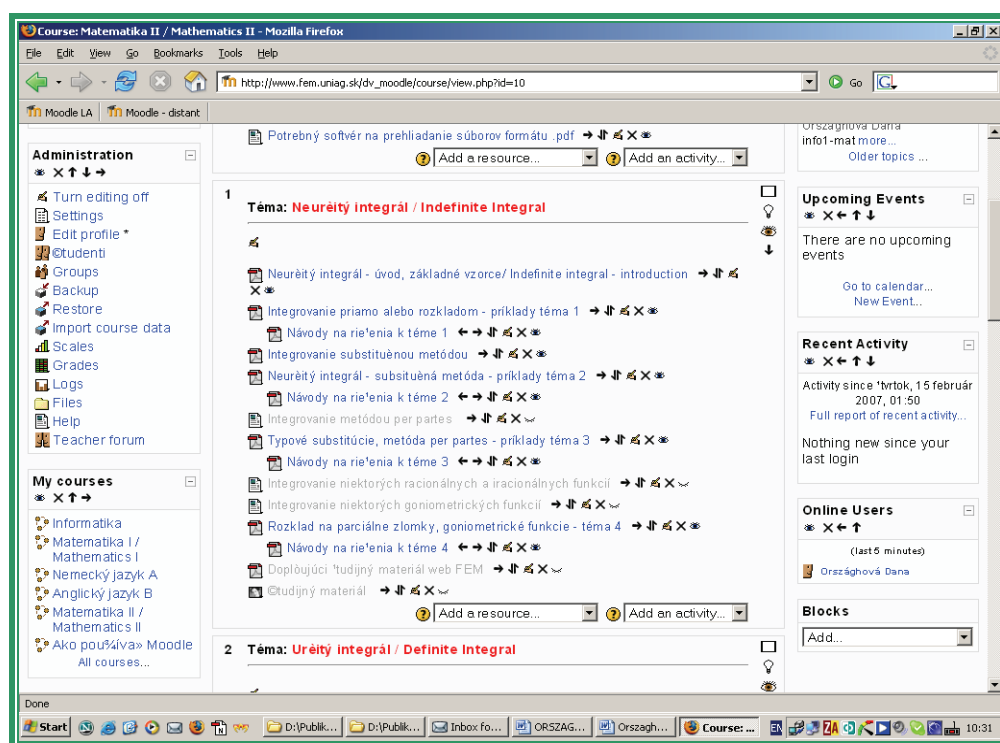


Figure 2: Illustration of the topic of the course Mathematics II

3 The structure and the content of the course “Exercises in Mathematics”

The priority of the current information society is the innovation of the contents, methods and forms of education based on information and communication technology (ICT), building virtual learning environments for education in modern schools with ICT [4, Gregáňová, 2006].

For the application of e-learning courses it is necessary teachers and students know to use tools of IT. Teachers of the Department of Mathematics of the SUA in Nitra create courses in the LMS MOODLE for several years. Students and teachers can enter created courses of the faculty departments directly from the website - <http://www.fem.uniag.sk/fem/>.

First created mathematics courses were “Exercises in Mathematics” (Winter Term and Summer Term) mentioned above. In the first course we can find these topics:

- Functions of one real variable, limits, derivatives, functions of two variables.

In the second one we can find the following topics:

- Indefinite integral, definite integral, vectors.

The structure of the topic is as follows [see Fig. 3]:

- Introduction to the topic with motivation example,
- Created dictionary with mathematical definitions and theorems,
- Set of examples with the procedure for solving,
- Set of exercises,
- Set of tests with the key.

Students can use tasks from exercises to prepare for the math seminar, test or exam. Distance learning students use these sets of tasks actively.

Another activity of the course is „Assignment“ (see Fig. 4, Fig. 5). We use on-line assignment and assignment in the form of attachment. Students elaborate their answers and send them via LMS to their teacher. Students can obtain points for the overall term assessment for correctly solved assignment. The assignments are the part of the feedback in the learning process; the teacher becomes a source of important information about students' knowledge.

9	Téma: Funkcia dvoch reálnych premenných	<input type="checkbox"/>
	Úvod témy 9	
	Slovník matematických pojmov 9	
	9.1 Pojem funkcie dvoch reálnych premenných, definičný obor a graf funkcie	
	cvičenie 1	
	9.2 Parciálne derivácie prvého rádu funkcie dvoch reálnych premenných	
	cvičenie 2	
	9.3 Parciálne derivácie vyšších rádov	
	cvičenie 3	
	9.4 Rovnica dotykovej roviny	
	cvičenie 4	
10	Téma: Lokálne extrémny funkcie dvoch reálnych premenných	<input type="checkbox"/>
	Úvod témy 10	
	Slovník matematických pojmov 10	
	10.1 Parciálne derivácie zloženej funkcie dvoch premenných	
	cvičenie 1	
	10.2 Parciálne derivácie vyšších rádov zloženej funkcie dvoch premenných	
	cvičenie 2	
	10.3 Lokálne extrémny funkcie dvoch premenných	
	cvičenie 3	
11	Téma: Viazané extrémny funkcie dvoch reálnych premenných	<input type="checkbox"/>
	Úvod témy 11	
	Slovník matematických pojmov 11	
	11.1 Viazané extrémny funkcie dvoch reálnych premenných	
	cvičenie 1	
12	Vyskúšajte sa!	<input type="checkbox"/>
	Test 7A	
	Výsledky testu 7A	
	Test 7B	

	Test 8 A	
	Výsledky testu 8A	
	Test 8B	

Figure 3: Topics of the course Exercises in Mathematics

In connection with the implementation of information technology into university education new requirements are given on the professional erudition of university teachers from the informatics branch. Because mathematics is included in the first year, it is necessary to explain students how to start work in the LMS MOODLE. The practical use of the LMS needs to know some of its tools. An administrative work includes registering students, create an account and password. Some students used the communication tools of LMS MOODLE, and sent teachers some technical questions and information about organizational problems.

Matematika z matematiky (LS)...

ZADANIE **ZADANIE** **ZADANIE** **ZADANIE** **ZADANIE** **ZADANIE**

1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1

Vyučujúca: doc. RNDr. Dana Országhová, CSc.

ZADANIA PRE ŠTUDIJNÉ SKUPINY: KME, UCT 1, UCT 2
Prihláste sa určite do kurzu, aby ste mohli pracovať na zadaniach za LS!

Úlohy 5-6-7-8 pre ŠS Účtovníctvo 1
Úlohy 5-6-7-8 pre ŠS Účtovníctvo 2
Úlohy 5-6-7-8 pre ŠS Kvantitatívne metódy v ekonómii

Vypracovali ste už Vaše zadanie do správneho hároku na odpovede?

Kľúč k zadaniu 2 (úlohy 5-6-7-8)

Vypracovali ste už Vaše zadanie?

ZADANIA

Vyučujúca: RNDr. Adriana Demová

ZADANIA PRE ŠTUDIJNÉ SKUPINY: EKP 2, EKP 3, EKP 1, EKP 5

A_Zadanie_LS_EKP_2
B_Zadanie_LS_EKP_3
C_Zadanie_LS_EKP_1.doc
D_Zadanie_LS_EKP_5.doc

Odpovedový hárok - stiahnite si tento vzor na odpovede (MS WORD)
Zadanie - pripojte váš súbor - odpovedový hárok

ZADANIA

Vyučujúca: Mgr. Radomíra Gregánová, PhD.

ZADANIA PRE ŠTUDIJNÉ SKUPINY: 1 MP, 2 MP, 4 MP

Seminárna práca_1MP
Seminárna práca_2MP
Seminárna práca_4MP

Figure 4: Reference to the assignments

The access to the mentioned courses is allowed for "guests", i.e. for unregistered participants of the course. However, some activities (e.g. assignments) may be invisible or inactive for „guests“.

Cvičenia z matematiky (ZS) FEM SPU

Moodle FEM - Cv(ZS)_KM - Zadania - Zadanie pre ŠS Účtovníctvo 1 (U6-U10)

Zadanie ZS: Účtovníctvo 1. študijná skupina (úloha 6 - 10)

6.úloha:
Zistite intervaly, na ktorých je funkcia $f : y = \frac{4+x^2}{4-x^2}$ rastúca, prípadne klesajúca (Do tabuľky pre odpovede napíšte k intervalom slovnú odpoveď: napr. rastúca, klesajúca).

A	B	C	D
$(-\infty, -2)$	$(-2, 0)$	$(0, 2)$	$(2, \infty)$

7.úloha:
Zistite, či v danom bode x má funkcia $f : y = \frac{4+x^2}{4-x^2}$ lokálne maximum alebo lokálne minimum. Odpoveď napíšte do tabuľky odpovedí.

A	B	C
$x = -2$	$x = 0$	$x = 2$

8.úloha:
Zistite intervaly konvexnosti, konkávnosti a inflexné body funkcie $f : y = 10 - 5x + 6x^2 - \frac{x^3}{3}$. Odpovede napíšte do tabuľky.

A	B	C
$(-\infty, 6)$	$(6, \infty)$	$x = 6$

Figure 5: Winter term assignment tasks

4 Conclusion

The current education gives a new requirement for students and teachers – the communication and computer literacy. Therefore, the main of this paper was to present possible ways of implementation of e-learning methods in mathematical subjects. The educational process via information and communication technologies has become the part of the university study. Means of information technologies affect the quality of mathematical education. Teachers can apply a variety of IT tools in the teaching to make the study of mathematics more attractive. And students can acquire new mathematical knowledge by modern and interesting methods of e-learning.

We can observe these notable features of the IT application in education:

- Changes in the technical support of education,
- Raising requirements for new educational competencies,
- Creation of electronic educational courses,
- Use of electronic educational courses (e-learning),
- Active study and self-study,
- Improving and streamlining the work of educator.

Method of e-learning also brings various advantages and disadvantages. It increases the attractiveness of education. Use of the Internet and web site supports the efficiency of education and reduces the financial burden of education. It provides administration tools for testing and assessing students' knowledge.

Pedagogical communication is the important part of education. From the aspect of this communication disadvantages of e-learning include the lack of interaction between participants of education, i.e. reduction of communication between students and teachers and students themselves. Technical support for electronic media, it also requires financial and time investment. Uncomfortable reading of electronic texts is another disadvantage. It is appropriate if an electronic educational material is created in an interactive format and also has a print version so that students can create (to print) themselves own textbook. To access new information and knowledge effectively – this is the future of using information technology in education. Tools to manage the learning process have application in distance education and lifelong learning. Different advantages of IT motivate teachers to search new ways of applying IT in teaching of mathematics.

Acknowledgement

The paper was supported by grant from Grant Agency KEGA with title “Theoretical and mathematical transformation of the educational training of agricultural engineers“, No. 3/7382/09.

References

- [1] BARANÍKOVÁ, H.: *Application of Informatics in Teaching of Mathematics at SAU*. In: Scientific papers (CD) of international seminary „New trends in university mathematical education“. Nitra: SPU, 2009, p. 169-174. ISBN 978-80-552-0197-9
- [2] BOŽIKOVÁ, M.: *Internet - a Means of Teaching Physics Content of Innovation*. Doctoral thesis, KF FPV UKF, Nitra 2004, 140 p.

- [3] DRÁBEKOVÁ, J., RUMANOVÁ, L.: *Use of GEOGEBRA Software in Examples Mathematical Analysis*. In: Scientific papers (CD) of international seminary „New trends in university mathematical education“. Nitra: SPU, 2009, p. 42-47. ISBN 978-80-552-0197-9
- [4] GREGÁŇOVÁ, R.: *Implementation of Web Pages of Mathematics in the Education Process*. In: CD Proceedings of the international scientific conference MVD 2006: Competitiveness in the EU - Challenge for the V4 countries. Nitra: SPU, 2006, p. 1233-1236. ISBN 80-8069-704-3
- [5] ORSZÁGHOVÁ, D.: *The Study of Mathematical Subjects with Using of LMS MOODLE*. In: Proceedings of the 5th international didactic conference DidZa 2008. Žilina: Žilinská univerzita, 2008, p. 6. ISBN 978-80-8070-688-3
- [6] ORSZÁGHOVÁ, D., GREGÁŇOVÁ, R., MAJEROVÁ, M.: *Professional Training of Economists and Managers in the Context of the Transformation of University Education*. In: Scientific papers (CD) “The Path of Internationalization and Integration in the Europe of Regions”. Curtea de Arges, Romania, 2007, p. 259-264. ISBN 978-80-8069-857-7
- [7] ORSZÁGHOVÁ, D., POKRIVČÁKOVÁ, S.: *To the Quality of Education in the Mathematical Subjects*. Proceedings of the International Seminar Quality Education in European Context and the Dakar Follow-up, Nitra, UKF, 2003, p. 122-127. ISBN 80-8050-636-1
- [8] URL <http://moodle.uniag.sk/fem/>
- [9] URL <http://moodle.uniag.sk/fem/course/view.php?id=44>
- [10] URL <http://moodle.uniag.sk/fem/course/view.php?id=143>
- [11] URL <http://moodle.uniag.sk/fem/course/view.php?id=207>

Current address

Dana Országhová, doc. RNDr. CSc.,

The Department of Mathematics, Faculty of Economics and Management, Slovak University of Agriculture in Nitra, 949 76 Nitra, Slovak Republic

Tel.: +421 0376414181

e-mail: Dana.Orszaghova@fem.uniag.sk

CIVIL IDENTIFICATION PROBLEMS WITH BAYESIAN NETWORKS USING OFFICIAL DNA DATABASES

ANDRADE Marina (P), FERREIRA Manuel Alberto M., (P)

Abstract. In forensic identification problems the study of DNA profiles is often used. DNA databases began to be used in England in 1995 and gave rise to new challenges when used in identification problems. In Portugal the legislation for the construction of a DNA database file was defined in 2008. So, it is important to determine how to use it in an appropriate way. An important forensic identification problem is body identification. That is, in general, the identification of a body found, or more than one, using the information of missing persons belonging to one or more known families for which there may be information of family members who claimed the disappearance. Here it is intend to discuss how to use the database: the hypotheses of interest and the database use to determine the likelihood ratios, i.e., how to evaluate the evidence in different situations.

Key words. Bayesian networks, DNA profiles, civil identification problems

Mathematics Subject Classification: 62C10, 68M10.

1 Introduction

The reason of this work is to propose a methodology, and give the adequate tools, to use correctly a DNA profiles database in the problem of civil identification case if there is a partial match between the genetic characteristic of an individual whose body was found, one volunteer who claimed a family member disappearance and one sample in the DNA database.

So in section 2 the civil identification case under study is presented and discussed. In section 3 a Bayesian network that allows the efficient computation of the probabilities determinant to evaluate the hypothesis in comparison is presented. Still in section 3 real life examples, which clarify the exposition, are presented. In the end a short list of references about these kind of problems is given.

2 Civil identification

The use of DNA profiles in forensic identification problems has become, in the last years, an almost regular procedure in many and different situations. Frequent examples of civil identification problems are the case of a body identification, together with the information of a missing person belonging to a known family, or the identification of more than one body resultant of a disaster or an attempt, and even immigration cases in which it is important to establish family relations.

This work focuses on civil identification problems. The establishment and use of DNA database files for a great number of European countries worked as a motivation to study in more detail the mentioned problems and the use of these database files for identification. In the context of the civil identification it may be very useful when unidentified corpses appear and may be identified by comparison of their DNA profiles with family volunteer's profiles.

The Portuguese law n°5/2008 establishes the principles for creating and maintaining a database of DNA profiles for identification purposes, and regulates the collection, processing and conservation of samples of human cells, their analysis and collection of DNA profiles, the methodology for comparison of DNA profiles taken from the samples, and the processing and storage of information in a computer file.

Here it is assumed that the database is composed of a file containing information of samples from convicted offenders with 3 years of imprisonment or more - α ; a file containing the information of samples of volunteers - β ; a file containing information on the "problem samples" or "reference samples" from corpses, or parts of corpses, or things in places where the authorities collect samples - γ . In this work the interest is to study problems of civil identification, particularly if there is a partial match between the genetic characteristic of an individual whose body was found and one volunteer who claimed a family member disappearance and one sample in the file γ of database.

2.1 A partial match with the volunteer and one γ - sample

In a problem of civil identification where there is an individual claiming for a disappeared person and gives his/her genetic information, C_{vol} , to be compared with the genetic characteristic of a body found, it is important to check first if there is a match between the genetic characteristic of the individual whose body was found, C_{BF} , and any sample of the DNA file, γ - sample, which is named "problem samples".

Considering it is checked and there is a partial match between the genetic profile of the individual whose body was found and one sample in the file γ , the evidence now is $E = (C_{BF}, \gamma - sample, C_{vol})$.

Regarding the problem it follows the establishment of the hypotheses of interest. The identification hypothesis (H_{ID}) versus the non identification hypothesis ($H_{not ID}$), as:

H_{ID} : It is possible to reach an identification of the individual whose body was found

vs

$H_{not ID}$: It is not possible to reach an identification of the individual whose body was found.

The first approach is to check the possibility of a partial match between the profile of the individual whose body was found, C_{BF} , the sample in the file γ , γ -sample, and the volunteer, C_{vol} . Thus, two different comparisons are made in order to obtain a measure either of the possible genetic relation between the individual whose body was found with the γ -sample ($bf_match_gs?$), or of the possible genetic relation between the individual whose body was found and the volunteer ($bf_match_vol?$). These comparisons may have as an answer: *yes* or *no*.

Combining the states of each comparison (*yes*, *no*); (*no*, *yes*); (*yes*, *yes*) and (*no*, *no*) are the resulting pairs.

State: (*yes*, *no*) – defines the possibility of genetic relationship between the individual whose body was found and the γ -sample but not the volunteer;

State: (*no*, *yes*) – defines the possibility of genetic relationship between the individual whose body was found and the volunteer but not the γ -sample;

State: (*yes*, *yes*) – defines the possibility of genetic relationship between the individual whose body was found and both the volunteer and the γ -sample;

State: (*no*, *no*) – defines the possibility of genetic relationship between the individual whose body was found neither with the volunteer nor with the γ -sample;

The first two states define the identification hypothesis, H_{ID} , and the last two define the non identification hypothesis, $H_{not ID}$. The state (*no*, *yes*) is a particular one that is the simple problem studied in Andrade and Ferreira (2009b). Each of the four possible states probabilities provide a measure for each event, and the four are pairwise incompatible.

After the probabilities computation it is important to have in mind the comparison between the state (*no*, *no*) versus the others; i.e., to evaluate the event the individual whose body was found is not genetically related either with the γ -sample or the volunteer. This first step comparison intends to evaluate the situation “the genetic information of the individual whose body was found is not compatible with the other genetic information available” and “the genetic information of the individual whose body was found is compatible with at least one of the remaining genetic information”, that is, compares the sets $\{(no, no)\}$ with $\{(no, yes), (yes, no), (yes, yes)\}$.

If $\{(no, no)\}$ is accepted the process ends and the body genetic information joins the file γ in the database. If $\{(no, no)\}$ is discarded next it is necessary to perform a comparison between $\{(no, yes), (yes, no)\}$ and $\{(yes, yes)\}$ events. If $\{(yes, yes)\}$ is accepted the process ends and police intelligence investigations must be done. If $\{(yes, yes)\}$ is discarded finally $\{(no, yes)\}$ and $\{(yes, no)\}$ must be compared. If $\{(no, yes)\}$ is accepted the conclusion is that the individual whose

3.1 Examples

In order to exemplify the described methodology in Table 1 are presented the allele frequencies (real ones) for some genetic markers¹ and, for each marker, possible evidence profiles for the body found (C_{BF}), the γ -sample and the volunteer (C_{vol}).

Marker	Allele Frequencies				$\{(C_{BF}), (\gamma\text{-sample}), (C_{vol})\}$
FGA	p_{21}	p_{22}	p_{23}	p_{24}	$\{(21,24), (21,21), (22,24)\}$
	0.1750	0.1950	0.1550	0.1750	
D21S11	p_{28}	p_{29}	p_{30}	$p_{31.2}$	$\{(29,30), (30,30), (29,31.2)\}$
	0.1674	0.2136	0.2437	0.1138	
F13A1	p_5	p_6	p_7	p_8	$\{(6,7), (7,8), (5,6)\}$
	0.1985	0.2890	0.3377	0.0112	
SE33	$p_{22.2}$	$p_{25.2}$	$p_{27.2}$	$p_{28.2}$	$\{(22.2,27.2), (22.2,28.2), (27.2,25.2)\}$
	0.1043	0.0764	0.1458	0.0695	
TH01	p_6	p_7	p_9	$p_{9.3}$	$\{(7,9), (9,9.3), (6,7)\}$
	0.2044	0.1696	0.1984	0.2748	
TPOX	p_8	p_9	p_{10}	p_{11}	$\{(8,11), (8,10), (9,11)\}$
	0.5053	0.0974	0.0647	0.2893	
VWA31	p_{15}	p_{16}	p_{17}	p_{18}	$\{(16,17), (17,17), (16,18)\}$
	0.1216	0.2300	0.2649	0.1859	

Table1: Allele frequencies and genetic profiles.

In Table 2 the state probabilities (the node counter states, see Figure 1) are presented.

States	FGA	D21S11	F13A1	SE33	TH01	TPOX	VWA31
(no, no)	0.0744	0.1108	0.2574	0.0751	0.1343	0.3112	0.1108
(no, yes)	0.1063	0.1296	0.2226	0.1287	0.1978	0.2688	0.2251
(yes, no)	0.2124	0.2274	0.1904	0.1797	0.1692	0.1539	0.2092
(yes, yes)	0.6069	0.5322	0.3296	0.6165	0.4987	0.2661	0.4548

Table 2: State probabilities.

¹ <http://www.uni-duesseldorf.de/WWW/MedFak/Serology/dna.html>

And in Table 3 the decisions, consequence of the procedures proposed in section 2.1, are presented for each example evidence profile.

Evidence Profiles	Decision
$\{(21,24), (21,21), (22,24)\}$	Police intelligence investigations must be done
$\{(29,30), (30,30), (29,31.2)\}$	Police intelligence investigations must be done
$\{(6,7), (7,8), (5,6)\}$	The individual whose body was found is a volunteer relative
$\{(22.2,27.2), (22.2,28.2), (27.2,25.2)\}$	Police intelligence investigations must be done
$\{(7,9), (9,9.3), (6,7)\}$	Police intelligence investigations must be done
$\{(8,11), (8,10), (9,11)\}$	The individual whose body was found is a volunteer relative
$\{(16,17), (17,17), (16,18)\}$	Police intelligence investigations must be done

Table 3: **Decisions for each evidence profile**

4 Conclusions

Performing the sequence of three hypothesis tests proposed with the probabilities computed through the Bayesian network, built specifically for a civil identification problem in which there is a partial match between an individual whose body was found, a volunteer who claimed a relative disappearance supplying his/her own genetic information and a DNA database file sample existent, it is possible to decide first if an identification is possible or not; second if an effective identification is possible or not; third to make the identification. The comparison between the hypotheses in each test is made through its probabilities values computing the respective likelihood ratio. The accepted hypothesis is the one that corresponds to the greatest probability event.

So with a procedure technically simple, since it was defined the Bayesian network and the chain of hypothesis tests it is possible to make an adequate and correct use of a DNA database.

And as the examples illustrate, the procedure leads almost surely to a decision: whether it is to close the case identifying the individual, or concluding that it is not possible any identification, or to go on with the police investigations.

Acknowledgement

The authors are members of the UNIDE/ISCTE research group StatMath/ISCTE which support they gratefully thank

References

- [1.] ANDRADE, M., FERREIRA, M. A. M.: “*Bayesian networks in forensic identification problems*”. Aplimat - Journal of Applied Mathematics. Volume 2, number 3, 13-30, 2009.
- [2.] ANDRADE, M. and FERREIRA, M. A. M.: *Criminal and Civil Identification with DNA Databases Using Bayesian Networks*. International Journal of Security. Volume 3, issue 4, 65-74, 2009.
- [3.] ANDRADE, M., FERREIRA, M. A. M., FILIPE, J. A.: “*Evidence evaluation in DNA mixture traces*”. Journal of Mathematics and Allied Fields (Scientific Journals International-Published online). Volume 2, issue 2, 2008.
- [4.] ANDRADE, M., FERREIRA, M. A. M., FILIPE, J. A., COELHO, M.: “*Paternity dispute: is it important to be conservative?*”. Aplimat – Journal of Applied Mathematics. Volume 1, number 2, 2008
- [5.] BALDING, David J.: “*The DNA database controversy*”. Biometrics, 58(1):241-244, 2002
- [6.] CORTE-REAL, F.: “*Forensic DNA databases*”. Forensic Science International, 146s:s143-s144, 2004.
- [7.] COWELL, R. G., DAWID, A. P., LAURITZEN, S. L., SPIEGELHALTER, D. J.: “*Probabilistic Expert Systems*”, Springer, New York, 1999.
- [8.] DAWID, A. P., MORTERA, J., PASCALI, V. L. Van BOXEL, D. W.: “*Probabilistic expert systems for forensic inference from genetic markers*”. Scandinavian Journal of Statistics, 29:577-595, 2002.
- [9.] EVETT, I., WEIR, B. S.: “*Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*”, Sinauer Associates, Inc., 1998.
- [10.] GUILLÉN, M., LAREU, M. V., PESTONI, C., SALAS, A., CARRECEDO, A.: “*Ethical-legal problems of DNA databases in criminal investigation*”. Journal of Medical Ethics, 26:266-271, 2000.
- [11.] MARTIN, P.: “*National DNA databases – practice and practability. A forum for discussion*”. In International Congress Series 1261, 1-8, 2004.
- [12.] NEAPOLITAN, R. E.: “*Learning Bayesian networks*”, Pearson Prentice Hall, 2004.

Current address

Marina Andrade, Professor Auxiliar

Iscte – Lisbon University Institute
Av. Das Forças Armadas
1649-026 Lisboa
Telefone: + 351 21 790 34 05, Fax: + 351 21 790 39 41
e-mail: marina.andrade@iscte.pt

Manuel Alberto M. Ferreira, Professor Catedrático

ISCTE – Lisbon University Institute
Av. Das Forças Armadas
1649-026 Lisboa

Telefone: + 351 21 790 37 03, Fax: + 351 21 790 39 41
e-mail: manuel.ferreira@iscte.pt

MIXTURE MODEL CLUSTERING FOR HOUSEHOLD INCOMES

BARTOŠOVÁ Jitka, (CZ), FORBELSKÁ Marie, (CZ)

Abstract. Finite mixture models are often used to study data from a population that is suspected to be composed of a number of homogeneous subpopulations. Mixture-model-based clustering has become a popular approach for its statistical properties and the implementation simplicity of the EM algorithm. Therefore, we focused on the partitions of household income into homogeneous subpopulations using the mclust library of R (see [3], [6]).

Keywords. Household income, lognormal distribution, finite mixture model, clustering

Mathematics Subject Classification: Primary 62H30, Secondary 30C40.

1 Economical basis and data sets

The survey of income and various demographic and regional characteristics of households, their level of housing, facilities, etc., with lots of interesting knowledge and information may further serve as a basis for analysis in the social and economic fields, respectively, for the formulation of measures that may affect these areas. When interpreting the obtained results, however, be borne in mind that this is a sample survey and therefore, the published results are representative estimates burdened with statistical error.

1.1 Mikrocensus and EU-SILC in the Czech Republic

Sample survey of household income in the Czech Republic is made by the Czech Statistical Office (CSO). From the fifties of the last century there was an irregular survey, which took place at intervals of 2 to 5 years under the name Microcensus. After the establishment of the Czech Republic, Microcensus was realized in 1989, 1992, 1996 and 2002. After the entrance to the European Union, Microcensus was replaced by annual survey of income and living conditions of households called EU - SILC. For the first time this investigation was carried out by the Czech Statistical Office in 2005 under the name Living Conditions 2005.

The purpose of this investigation is to obtain representative data on the receiving division of the different types of households, as well as information on method, quality and financial cost of housing, household durables as well as labor, material and health conditions of adults living in the household. The obligation to carry out an annual sample survey called "Living conditions" as the national-wide investigation module EU-SILC (European Union - Statistics on Income and Living Conditions), the CSO shows the amendment Regulation (EC) 1177/2003 and the subsequent implementation of the European Commission. The EU participates on financing of this investigation.

Investigation is carried out by the so-called rotating panel, where the same households were re-interviewed in the annual intervals for four years. After this time are replaced by other households living in the newly visited homes that are added to the investigation file continuously by the random selection. Longer monitoring of a household permits building image of their social situation, not only in the year, but also the changes and developments over time. Czech modification of EU-SILC survey maintains continuity with previous surveys Microcensus by its methodology for the selection of households. The same principle is also based on following data correction including estimates of underestimated revenues and avoids the influence of different interviewer's success in different regions and types of households.

The basic segmentation of the surveyed subjects is the so-called classification of a household. This definition means the voluntary declaration of persons living in dwellings that live together and operate, i.e., pay expenses for meals, accommodation, etc. The important person is called the head of household. In two-parent families as a head of household takes every man, regardless of its economic activity, regardless of whether the income is actually greater, or at least a substantial portion of family income, or even whether he was unemployed. In single-parent families where is only one parent with children and then non-family households headed by a person assessed on the basis of its economic activity respectively of height of income.

We can obtain information from data sets not only on total disposable household income (the income per household) and income calculated per household member (the income per capita), but also the equivalent disposable household income (the income per unit) which is comparable with other countries.

The Czech Statistical Office uses two methods for the transformation of income into an equivalent range:

1. OECD methodology, where
 - the head of household is taken with a coefficient of 1.0,
 - children aged 0 to 13 with a coefficient of 0.5,
 - and the other children and people with a coefficient of 0.7.
2. EU methodology, where
 - children aged 0 to 13 are taken with a coefficient of 0.3,
 - and the other children and people with a coefficient of 0.5.

Comparison of the financial potential of households within the EU is determined by the equivalent scale with the number of consumer units according to EU methodology. This scale is especially designed to assess the financial situation of households in Western Europe, where households spend

much higher amounts of common expenses (housing, water, electricity, fuel, etc.) than in post-communist countries of Eastern Europe. The financial situation in the countries of Eastern Europe is more representative with the equivalent scale from the OECD methodology.

1.2 Modelling of income distribution in the Czech Republic

Statistical models of income distribution provide basis for evaluation of the living standards of the population of the country at whole as well as comparison of the living standards of different social classes or regions. They also represent an indicator of the relative living standards in a chosen country in comparison with other countries.

Before 1990 (“Velvet Revolution”), planned economy experienced high homogeneity of income in the population in all social classes. Character of income distribution before Velvet revolution was determined by principals of remuneration in socialistic régime. After “Velvet Revolution”, former administrative rules for remuneration were immediately changed. Bracket of wages in the Czech Republic started to grow after 1990. Finally, it has reached the level of differentiation of wages usual in Western Europe. At the same time, the structure of incomes changed. The variety of income sources and present process of differentiation of wages has resulted in

- Rise of discrepancies between the empirical income distribution and the theoretical model
- Occurrence of considerably different incomes that may be concerned as outliers
- Commixtures of two (or more) simple models of distribution (see e.g. Bartošová, Bína, 2007, [1])

2 Mixture model

The distribution of income in most populations is highly skewed, with a long right-hand-side tail and high density at the lower percentiles. The logarithm is the natural transformation for such data. So we shall consider the lognormal distribution as an approximative mathematical model of the household income distribution. The lognormal distribution is positive valued and has a large tail that can describe the presence of extreme variability of observations.

We assume that the population can be broken down into k homogeneous subpopulations (strata) with proportions π_1, \dots, π_k and that each household income Y follows lognormal distribution. Here, homogeneity refers to similar income sources, geographical, social, demographic, and professional characteristic of its representatives.

Note that, as with number of modes used to detect heterogeneity, the number of components in the mixture is invariant under a continuous and monotonic transformation of income Y . So, if Y is a mixture of k lognormal densities, then $\log(Y)$ is a mixture of k normal densities.

We let $x_i = \log(y_i)$ denote the value of $X = \log(Y)$ corresponding to the i th entity ($i = 1, \dots, n$). With the mixture approach to clustering, x_1, \dots, x_n are assumed to be an observed random sample from mixture

of a finite number of groups in some unknown proportions π_1, \dots, π_k . The mixture density of x_i is expressed as

$$f(x_i; \Psi) = \sum_{j=1}^k \pi_j f_j(x_i; \theta_j) \quad (2.1)$$

where the mixing proportions π_1, \dots, π_k sum to one and the group-conditional density $f_j(x_i; \theta_j)$ is specified up to a vector θ_j of unknown parameters ($j = 1, \dots, k$). The vector of all the unknown parameters is given by $\Psi = (\pi_1, \dots, \pi_{k-1}, \theta_1, \dots, \theta_k)$. Using an estimate of Ψ , this approach gives a probabilistic clustering of the data into k clusters in terms of estimates of the posterior probabilities of component membership,

$$\omega_j(x_i) = \frac{\pi_j f_j(x_i; \theta_j)}{f(x_i; \Psi)}, \quad (2.2)$$

where $\omega_j(x_i)$ is the posterior probability that x_i (really the entity with observation x_i) belongs to the j th component of the mixture ($i = 1, \dots, n, j = 1, \dots, k$).

In the Bayesian framework, we use the rule which assigns observation $x_i = \log(y_i)$ to the class for which x_i has the highest posterior probability.

The parameter vector Ψ can be estimated by maximum likelihood (*MLE*) and can be obtained via the expectation-maximization (*EM*) algorithm of Dempster et al. (1977, see [2]).

In practice, the number of components k is unknown and can be chosen as that which minimizes some criterion, e.g. Bayesian Information Criterion *BIC* of Schwarz (1978, see [7]), see also McLachlan and Peel (2000, see [5]).

The use of mixture models for clustering is sometimes referred to as model-based probabilistic clustering, since a particular functional form for the component densities must be assumed.

3 Household income clustering

Estimation of parameters in a mixture model includes model order selection, i. e., determination of the number of mixture components.

3.1 Initial model order selection via kernel density estimates

In order to identify the number of components, we first construct the kernel estimates of the density functions of logarithm of household income $X = \log(Y)$. The kernel density estimate, $f_n(x)$, of a set of n points X_1, \dots, X_n from a density $f(x)$ is defined as:

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \quad (3.1)$$

where K is the kernel function (e.g. *Gaussian kernel function*), h is the smoothing parameter or window width (see e.g. Silverman, 1978, [8], Horová, Zelinka, 2000, [4]).

Figures 1a, 1b and 1c show the shape of estimated densities using a Gaussian kernel for the natural logarithm of incomes over years 2005, 2006 and 2007. We may observe the formation of three modes in the distribution of $\log(\text{income})$.

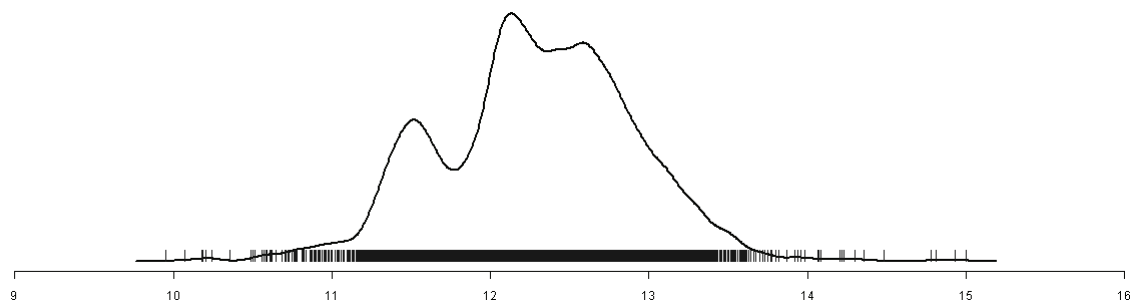


Figure 1a. Kernel density estimate of $\log(\text{income})$ in 2005. (Source: SILC 2005)

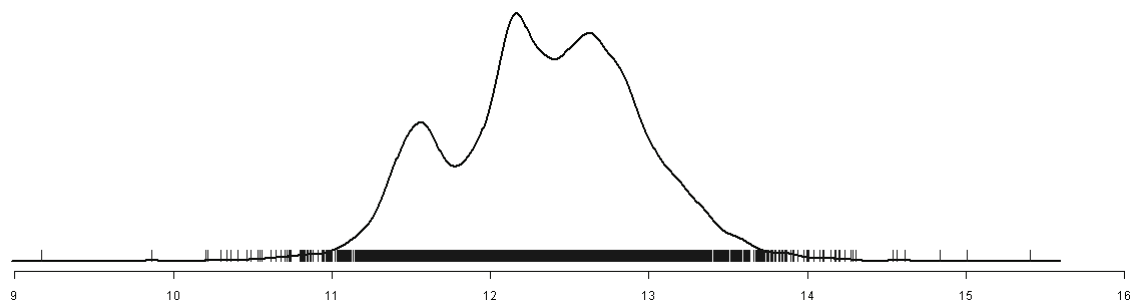


Figure 1b. Kernel density estimate of $\log(\text{income})$ in 2006. (Source: SILC 2006)

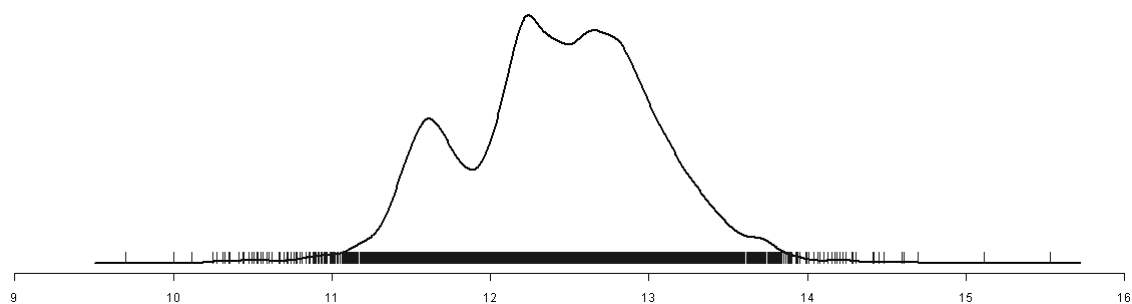


Figure 1c. Kernel density estimate of $\log(\text{income})$ in 2007. (Source: SILC 2007)

Based on the kernel density estimates (see Figures 1a, 1b, 1c), we first consider three-component mixtures. Results given by EM algorithm are presented via Figures 2a, 2b, 2c and the Table 1.

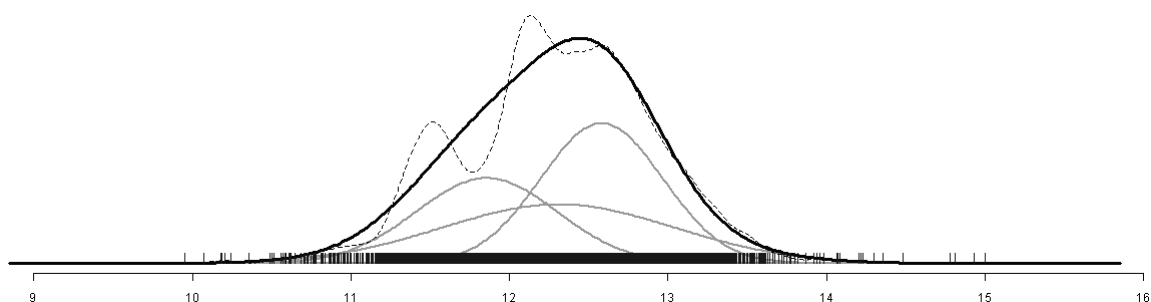


Figure 2a. Solution of the EM algorithm for three-component mixtures of $\log(\text{income})$ in the year 2005 (dashed line denotes kernel estimate) – data source SILC 2005

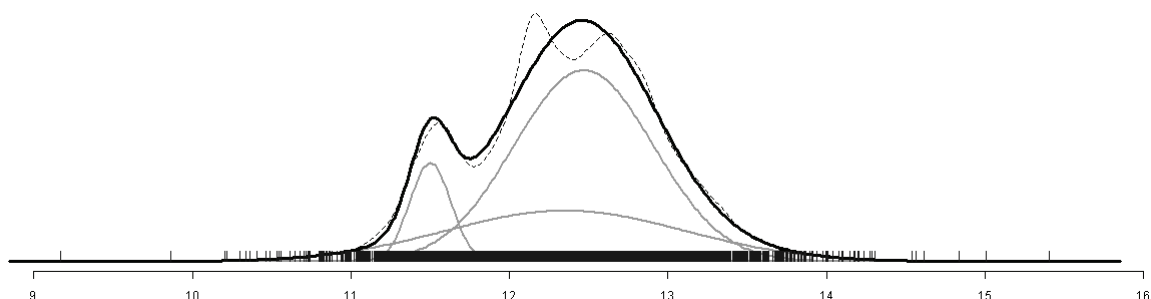


Figure 2b. Solution of the EM algorithm for three-component mixtures of $\log(\text{income})$ in the year 2006 (dashed line denotes kernel estimate) – data source SILC 2006

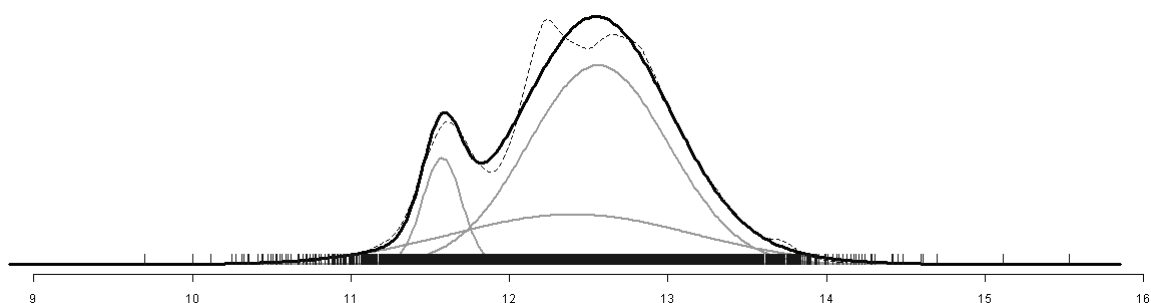


Figure 2c. Solution of the EM algorithm for three-component mixtures of $\log(\text{income})$ in the year 2007 (dashed line denotes kernel estimate) – data source SILC 2007

Table 1: Estimated parameters for the three-component mixtures of $\log(\text{income})$ of incomes over years 2005, 2006, 2007

Year	Component	Proportion π_j	Mean μ_j	Variance σ_j^2
2005	1	0.281	11.855	0.200
	2	0.400	12.581	0.150
	3	0.320	12.309	0.542
2006	1	0.094	11.501	0.017
	2	0.623	12.468	0.194
	3	0.282	12.351	0.566
2007	1	0.094	11.574	0.015
	2	0.632	12.560	0.196
	3	0.274	12.400	0.588

However, Figures 2a, 2b, 2c show clearly that modelling household income by means of the three components is inadequate.

3.1 Optimal model order selection

Choosing the right number of components is a crucial question in a clustering problem. With the Gaussian mixture model, we have a *MLE* likelihood criterion that we must optimize. We can evaluate various values of the number of components and select the best one. Unfortunately, the *MLE* estimate of k is not well defined because the likelihood may always be made better by choosing a large number of subclusters. Therefore we use the Bayesian Information Criterion *BIC*

$$BIC = 2\log(\text{maximized likelihood}) - m \log(n), \quad (3.2)$$

where m is the number of parameters and n is number of points. Thus the *BIC* criterion is a penalized likelihood criterion.

Figures 3a, 3b, 3c and Table 2 show the resulting six-component partitions given by EM algorithm with *BIC* criterion.

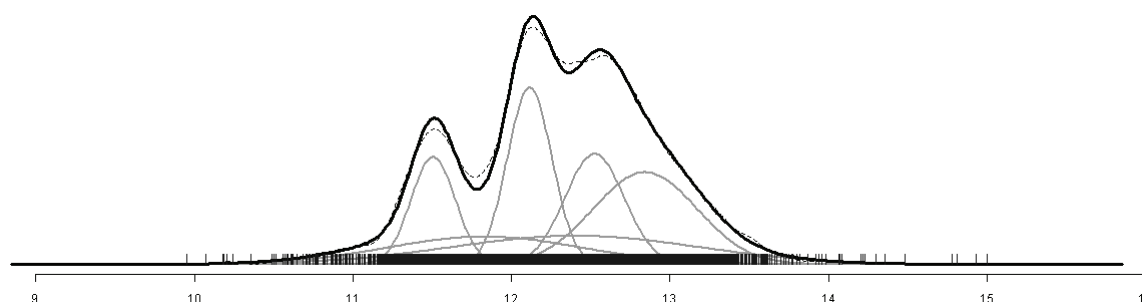


Figure 3a. *BIC*-optimal six-component mixtures of $\log(\text{income})$ in the year 2005 (dashed line denotes kernel estimate) – data source SILC 2005

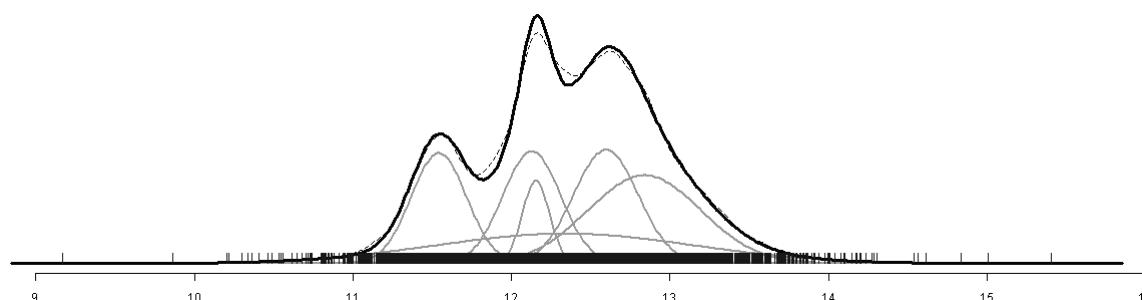


Figure 3b. *BIC*-optimal six-component mixtures of $\log(\text{income})$ in the year 2006 (dashed line denotes kernel estimate) – data source SILC 2006

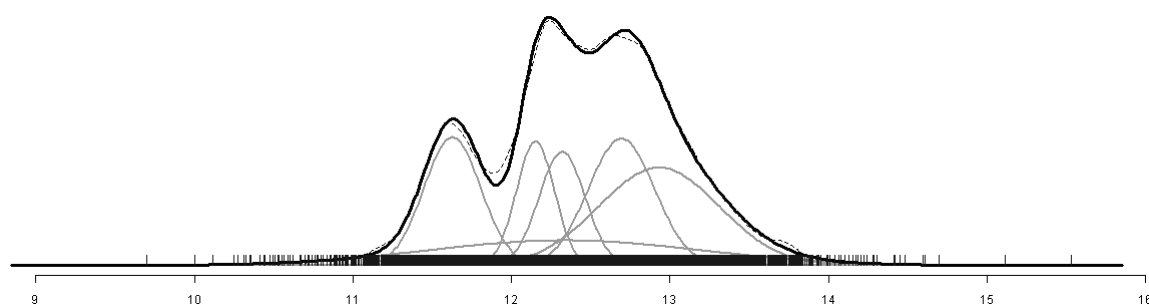


Figure 3c. BIC-optimal six-component mixtures of $\log(\text{income})$ in the year 2007 (dashed line denotes kernel estimate) – data source SILC 2007

Table 2: Estimated parameters for the optimal six-component mixtures of $\log(\text{income})$ over years 2005, 2006, 2007

Year	Component	Proportion π_j	Mean μ_j	Variance σ_j^2
2005	1	0.130	11.828	0.375
	2	0.116	11.506	0.020
	3	0.191	12.114	0.020
	4	0.159	12.524	0.035
	5	0.232	12.843	0.107
	6	0.171	12.400	0.610
2006	1	0.153	11.542	0.030
	2	0.056	12.154	0.007
	3	0.164	12.128	0.034
	4	0.185	12.596	0.042
	5	0.250	12.839	0.127
	6	0.192	12.336	0.666
2007	1	0.165	11.627	0.032
	2	0.107	12.151	0.014
	3	0.118	12.319	0.021
	4	0.190	12.694	0.043
	5	0.267	12.929	0.144
	6	0.154	12.334	0.736

The BIC plot is shown in the Figure 4.

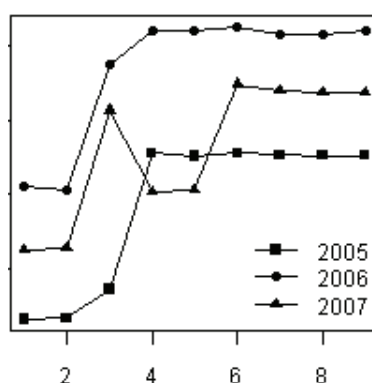


Figure 4. BIC plot for $\log(\text{income})$ over years 2005, 2006, 2007.

A mixture model with a large number of components provides a good fit but may have poor interpretive value. However, even with a suboptimal choice of order parameter k ($k=4$ for years 2004 and 2005), we can obtain interesting results (see next Figures 5a, 5b and Table 3).

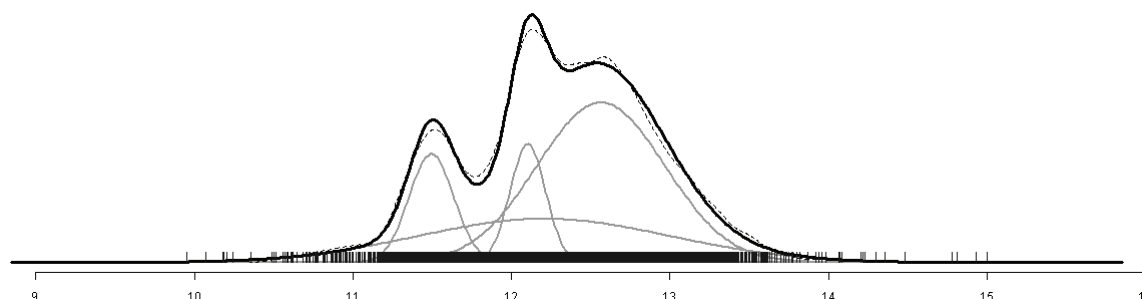


Figure 5a. Suboptimal four-component mixtures of $\log(\text{income})$ in the year 2005 (dashed line denotes kernel estimate) – data source SILC 2005

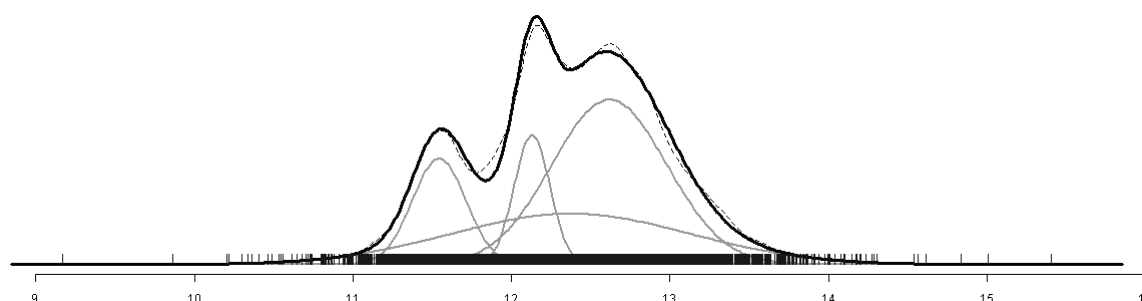


Figure 5b. Suboptimal four-component mixtures of $\log(\text{income})$ in the year 2006 (dashed line denotes kernel estimate) – data source SILC 2006

Table 3: Estimated parameters for the suboptimal four-component mixtures of $\log(\text{income})$ over years 2005 and 2006.

Year	Component	Proportion π_j	Mean μ_j	Variance σ_j^2
2005	1	0.120	11.495	0.020
	2	0.104	12.103	0.013
	3	0.505	12.564	0.163
	4	0.271	12.204	0.631
2006	1	0.138	11.543	0.029
	2	0.112	12.129	0.013
	3	0.458	12.618	0.132
	4	0.292	12.374	0.567

4 Conclusions

The mixture model arises as a simple and natural way to model heterogeneity typically observed in the household income data. Therefore the household income distribution could not be modeled by any known unimodal distribution. Typically, a mixture distribution will produce a better fit to a data set than a single component distribution, because there are more parameters in the mixture distribution than the single component case.

Decomposition of mixtures could be done in two different ways (logical or data oriented). Logical methods of decomposition should always be preferred. Logical decomposition allows us to find real cause of occurrence of mixture and gives us a chance to model single components using models with different parameters. This decomposition (decomposition by known defined rules) doesn't have to be optimal in the view of income distribution modeling. Its success depends on discovering suitable sorting criteria, and so it is limited by specific characteristics of data file and by an experience of researchers.

Acknowledgement

The research was supported by project of Grant Agency of the Czech Republic no. 402/09/0515 with title: "Analysis and modelling of financial power of Czech and Slovak Households".

References

- [1.] BARTOŠOVÁ, J., BÍNA, V.: *Mixture Models of Household Income Distribution in the Czech Republic*. In Kováčová, M. (ed.) 6th International Conference APLIMAT 2007, Part I. Slovak University of Technology, Bratislava, pp. 307-316, 2007.
- [2.] DEMPSTER, A. P., LAIRD, N. M. RUBIN, D. B.: *Likelihood from Incomplete Data via the EM Algorithm*. In Journal of the Royal Statistical Society. Series B (Methodological) **39** (1), pp. 1–38, 1977.
- [3.] FRALEY, C., RAFTERY, A. E.: *MCLUST: Normal Mixture Modeling and Model-Based Clustering*. R package version 3.0-0; 2006.
- [4.] HOROVÁ, I., ZELINKA, J. *Contribution to the bandwidth choice for kernel density estimates*. In Computational Statistics, Springer, 22, 1, pp. 31-47, 2007.
- [5.] McLACHLAN, G. J. , PEEL, D.: *Finite mixture models*. New York: Wiley & Sons, 2000.
- [6.] R Development Core Team: *A language and environment for statistical computing*. R. Foundation for Statistical Computing, Vienna, Austria. 2008. URL <http://www.R-project.org>
- [7.] SCHWARTZ, G.: *Estimating the Dimension of a Model*. In The Annals of Statistics, 6 (2), pp. 461-464, 1978.
- [8.] SILVERMAN, B. W.: *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

Current address

Jitka Bartošová, RNDr., PhD.

University of Economics Prague, Department of Management of Information of the Faculty of Management, Jarošovská 1117/II, Jindřichův Hradec, 377 01, Czech Republic,
tel.: +420 384 417 221,
email: bartosov@fm.vse.cz

Marie Forbelská, RNDr., PhD.

Masaryk University, Department of Mathematics and Statistics of the Faculty of Science, Kotlářská 2, Brno, 611 37, Czech Republic,
tel.: +420 549 493 811
email: forbels@math.muni.cz

SOME FACTORS AFFECTING EXPENDITURE ON HOUSING IN THE CZECH REPUBLIC

BARTOŠOVÁ Jitka, (CZ), BÍNA Vladislav, (CZ)

Abstract. Difference between financial power of the poorest and the wealthiest inhabitants in the Czech Republic continually grows. We also know that Czech households are more and more indebted. And continuation of the similar trend could be expected even in future. The most significant part of Czech household's debt are the mortgages. Expenditure on housing is very significant part of expenditures of the Czech households. Therefore, we focused on the analysis of the average total monthly expenditure on housing in the Czech Republic.

Key words. Analysis of dependence, expenditure on housing, SILC 2005

Mathematics Subject Classification: Primary 46N30; Secondary 62E15

1 Introduction

The former Czechoslovakia during the socialist era was ranked among the countries with the lowest wage differences not only in Europe but also in the whole world. This was caused partly by equalitarian tradition and partly by political and economical authority of the socialist system. After the year 1990 we can observe the first symptoms of non-uniform distribution of incomes. The reason was a transformation to free market economy.

After fifty years of the stable income distribution there emerge inequalities caused by free market principles. During the transformation process the question of economical inequality is shifting towards the forefront. Economical reform entails significant financial difficulties to many people and in some cases even poverty (see e.g. [3], [6]). Thus the question of inequality is a very sensitive theme. The lowest incomes are ensured by the implementation of living wages but the highest incomes are perpetually increasing. Thus the gap between the wealthiest and the poorest households continually grows (see e.g. [1], [5], [7]).

Household expenditure (or consumption of households) is also important economical indicator. If the expenditures are lower than incomes we speak about living in surplus and the difference is called savings. Nowadays, we can observe rather opposite trend – towards indebtedness. E.g. the

total debt of Czech households in the relation to gross disposable income grew in 2006 to 40%. And the continuation of similar trend could be expected even in future (see e.g. [2], [4]).

The progression of ratio of savings relative to disposable income has in case of the Czech household's undesirable trend. In ten years (from 1995 to 2005) the ration of savings decreased nearly to one tenth (from 10% to 1,5%). The growing indebtedness of Czech households constitutes a significant threat. Also the share of households on the national savings in the Czech Republic in last 10 years significantly decreased.

Expenditure on meals and non-alcoholic beverages belongs among the essential expenditures of the Czech households. Other very important charges are expenditures on housing and related costs for energies, water supply, maintenance, installment of mortgages etc. Another important factor is the charge for transport, especially for the people dependent upon daily commutation.

2 Applied methods

For the representation of the level of expenditures, variability and irregularity of their distribution different summary characteristics are employed. Momentum type characteristics contain information formed from all surveyed entities but usually are remarkably biased by outliers appearing in the sample. Thus our perception of reality can be distorted. On the other hand, quantile characteristics bear the information formed only on the bases of few significant values. This evident disadvantage is compensated by their robustness. For complex description of expenditure data we will take advantage of the combination of both types of characteristics, since data files can contain outliers, especially in the scope of very high financial amounts.

For the same reason, for the analysis of expenditures dependence on different social-economic and demographic factors is better to use non-parametrical methods since assumptions of application of parametrical methods will be violated in most cases. Graphs can also be a very useful tool providing rapid and transparent information about the problem (see e.g. [7]).

3 Expenditure on housing

The results of sample survey of incomes and expenditures (Household Budget Statistics - HBS) show that the financial burden of the Czech households is strongly influenced by expenditure on housing. Therefore, we focused on the analysis of the average total monthly housing expenditure in the Czech Republic. We are interested in the description of expenditure on housing depending upon various factors that can significantly affect these costs. It is especially the type and size of municipality and type of dwelling according to legal grounds of use. In all cases the 5% level of significance was used. For the analysis data set from 2005 was used; it contains 4351 households.

3.1 Basic characteristics

For describing of the distribution of housing expenditure some of the basic location and variability characteristics are used (specifically: mean, median, maximum, minimum, standard deviation and coefficient of variation). In the following tables, chosen characteristics of the total monthly expenditure on housing are shown. They are divided according to the

- type of municipality,

- size of municipality,
- type of dwelling.

Categories of these three variables are taken from the SILC data file.

3.1.1 Expenditure on housing according to the type of municipality

Type of municipality can be one of the factors significantly influencing the expenditures of household on housing. The notion of type of municipality means division of the households into four categories. First category is the capital Prague and its boroughs. In the second category regional centers are included. The third category contains all municipalities having characteristic attributes of town (at least 3000 inhabitants, the professional and social structure differs and should have square).

Table 1. Basic characteristics of the total monthly expenditures on housing according to the type of municipality.

Type of municipality	Mean	Min.	Max.	Median	Standard deviation	Coefficient of variation	Number of households
capital Prague	4264.412	1107	17000	3883.0	2114.638	0.496	469
regional centers	3644.072	0	10006	3383.0	1502.296	0.412	677
towns	3370.957	0	12605	3234.0	1293.786	0.384	1711
villages	2951.514	0	21903	2862.5	1313.751	0.445	1494

In the table 1 we can see that in three categories (regional centers, towns and villages) there exist households with zero expenditure on housing. Quite interesting is also the fact that maximal expenditure was registered in the village. This apparent paradox could be caused by contemporary trend not to live in the capital but in suburban zone. This is mainly the case of Prague surroundings.

3.1.2 Expenditure on housing according to the type of dwelling

The housing “costs” should certainly depend on the ownership relation of the household and the dwelling. According to statistical survey “Living conditions 2005” dwellings were divided into six categories according to the legal ground of usage. Generally we distinguish three basic categories of flats according to the person of the owner. It depends whether this person actually owns the dwelling or whether it is rented or in the cooperative property. Those categories show the distribution of expenditures according to the attitude of the household to the dwelling.

Table 2. Basic characteristics of the total monthly expenditure on housing according to the type of dwelling.

Type of dwelling	Mean	Min.	Max.	Median	Standard deviation	Coefficient of variation	Number of households
own house	3090.585	0	11125	2959.0	1236.512	0.400	1785
own flat	3105.150	676	8250	2942.0	1108.813	0.357	749
cooperative flat	3902.396	1208	9098	3708.5	1275.571	0.327	568

rented flat	3966.585	576	21903	3597.0	1885.852	0.475	1059
company flat	3104.429	773	4500	3825.0	1458.711	0.470	7
other use	1983.372	0	12605	1875.0	1321.343	0.666	183

The table shows that zero expenditures were registered in the case of the own house and other unspecified legal grounds. On the contrary the household housing in rented flat experienced the highest expenditures.

3.1.3 Expenditure on housing according to the size of municipality

Finally, we show the basic characteristics of the total monthly expenditure on housing, depending on the size of the village.

Table 3. Basic characteristics of the total monthly expenditure on housing according to the size of municipality.

Size of municipality	Mean	Min.	Max.	Median	Standard deviation	Coefficient of variation	Number of households
to 199 inhabitants	2659.660	442	5236	2827.0	1156.173	0.434	47
200 – 499	2792.378	430	9523	2689.5	1271.589	0.455	286
500 – 999	2966.107	0	7671	2893.5	1255.256	0.423	394
1 000 – 1 999	2951.867	30	8393	2871.0	1171.632	0.396	384
2 000 – 4 999	3101.298	0	8610	3032.0	1246.016	0.401	510
5 000 – 9 999	3234.239	160	21903	3045.0	1502.973	0.464	339
10 000 – 49 999	3396.377	0	12605	3242.0	1305.709	0.384	970
50 000 – 99 999	3539.353	0	10006	3312.5	1460.645	0.412	538
100 000 and more	4011.100	0	17000	3637.0	1860.754	0.464	883

As could be expected the size of municipality have a strong influence on the expenditure on housing. The zero costs were registered in five cases even in the case of the town with 100 000 or more inhabitants. Maximum was registered in the case of small town (between 5 and 10 thousand of inhabitants).

3.2 Dependence of housing expenditure on selected factors

First we will check whether the data fulfils conditions of the use of parametrical methods. Therefore, we will investigate whether there has not been violation of the assumption of normality and homoskedasticity of residuals. For this purpose, we can use both graphical and numerical methods (box – plot, Q – Q graph, Shapiro – Wilk test of normality, Levene test of homogeneity of variance, etc.). If we find a violation of the above mentioned assumptions, there is still possibility to employ for testing the nonparametric Kruskal – Wallis rank sum test (with weaker assumptions).

3.2.1 Detection of violation of normality and homoskedasticity of residuals

As the box - plots indicate, normality of the data is probably corrupted. Also Shapiro – Wilk tests of normality rejected the hypothesis of normality (see table 4). Even Levene tests of homogeneity of variance indicate rejects the assumption of homoskedasticity. And as we can see in the results of tests (see table 4), similar results are obtained in all cases – in the case of dependence on the type of municipality, on the type of dwelling and on the size of municipality. Thus we will employ a non-parametrical variant – Kruskal – Wallis rank sum test.

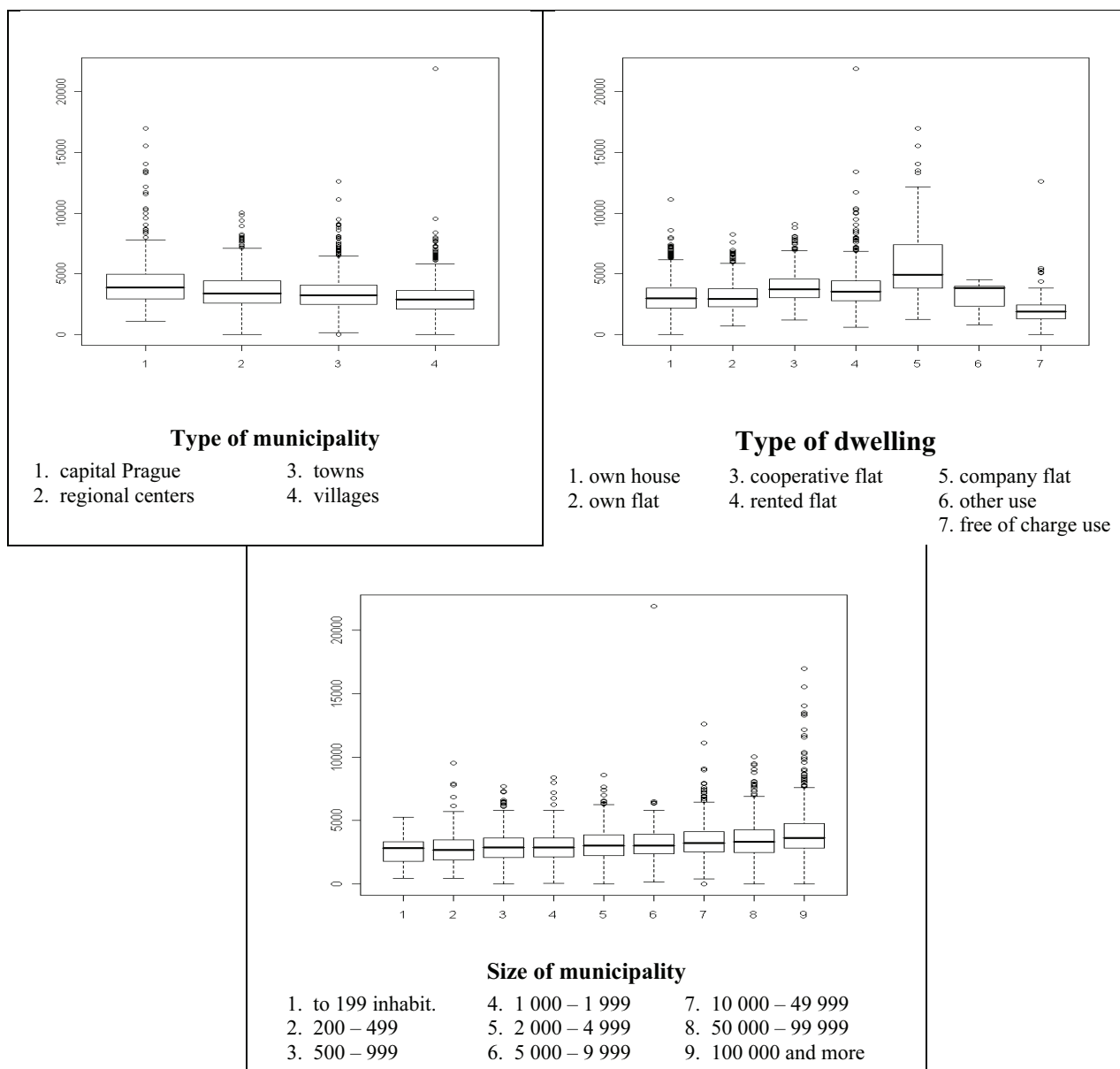


Figure 1. Boxplots of the total monthly expenditure on housing according to the type and size of municipality and type of dwelling. (Data source: SILC 2005)

3.2.2 Testing the differences of expenditures on housing according to the selected factors

Resulting values of statistics and p-values of the Kruskal – Wallis rank sum test for expenditures on housing according to the selected factors are located in the Table 4. The values given in the table shows that the first result is to confirm the hypothesis that expenditure on housing at 5% level of significance depends on all the selected factors – both the type and size of the municipality, and the type of housing.

Table 4. Verification of assumptions for parametric tests and resulting values of statistics and p-values of the Kruskal – Wallis rank sum test.

Tests for total monthly expenditure on housing according to the type of municipality

Shapiro – Wilk test of normality	$W = 0.9064$	$p - value < 2.2 \cdot 10^{-16}$
Levene test of homogeneity of variance	$Levene's F = 29.692$	$p - value < 2.2 \cdot 10^{-16}$
Kruskal – Wallis rank sum test	$Kruskal-Wallis \chi^2 = 266.3490$	$p - value < 2.2 \cdot 10^{-16}$

Tests for total monthly expenditure on housing according to the type of dwelling

Shapiro – Wilk test of normality	$W = 0.9149$	$p - value < 2.2 \cdot 10^{-16}$
Levene test of homogeneity of variance	$Levene's F = 39.2291$	$p - value < 2.2 \cdot 10^{-16}$
Kruskal – Wallis rank sum test	$Kruskal-Wallis \chi^2 = 578.5075$	$p - value < 2.2 \cdot 10^{-16}$

Tests for total monthly expenditure on housing according to the size of municipality

Shapiro – Wilk test of normality	$W = 0.9033$	$p - value < 2.2 \cdot 10^{-16}$
Levene test of homogeneity of variance	$Levene's F = 8.8466$	$p - value = 4.38 \cdot 10^{-12}$
Kruskal – Wallis rank sum test	$Kruskal-Wallis \chi^2 = 278.7588$	$p - value < 2.2 \cdot 10^{-16}$

3.2.3 Testing the differences of expenditures on housing in the particular types of municipalities and dwellings and in the municipalities of various size

Now we use the pair comparison and determine in which cases the differences are statistically significant. We will use two sample Wilcoxon test with Bonferroni correction. The significance level will be in the first case 0.00833, in the second and third 0.002380952. Resulting p-values of the Wilcoxon rank sum test with continuity correction are shown in the tables 5 – 7.

Table 5. Test of the differences of housing expenditure in the particular types of municipalities.

Categories	2	3	4
1	$3.103 \cdot 10^{-7}$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
2		0.001488	$< 2.2 \cdot 10^{-16}$
3			$< 2.2 \cdot 10^{-16}$

Table 5 shows that the expenditures on housing are on the 5% significance level different in all types of municipalities.

Table 6. Test of the differences of housing expenditure in the particular types of dwellings.

Categories	2	3	4	5	6	7
1	0.6764	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	0.6061	$< 2.2 \cdot 10^{-16}$
2		$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	0.5573	$< 2.2 \cdot 10^{-16}$
3			0.00067	$1.145 \cdot 10^{-11}$	0.3647	$< 2.2 \cdot 10^{-16}$
4				$2.864 \cdot 10^{-16}$	0.6321	$< 2.2 \cdot 10^{-16}$
5					0.01439	$< 2.2 \cdot 10^{-16}$
6						0.03533

Table 6 shows situation in various types of dwellings. The cases when the null hypothesis is rejected (the equality of median expenditures on housing) are highlighted. Those results means that households from the categories 1 and 2 (i.e. own house or flat) does not vary in their expenditures. It is obvious that median expenditures of households in category 6 do not differ from any category. But we should mention that only 7 households belong to this category.

Table 7. Test of the differences of housing expenditure in the particular municipalities of various size.

Categories	2	3	4	5	6	7	8	9
1	0.7412	0.1512	0.1286	0.02552	0.006788	0.0002044	$9.139 \cdot 10^{-5}$	$6.333 \cdot 10^{-8}$
2		0.02998	0.01982	$5.95 \cdot 10^{-5}$	$1.705 \cdot 10^{-6}$	$1.046 \cdot 10^{-14}$	$1.821 \cdot 10^{-14}$	$< 2.2 \cdot 10^{-16}$
3			0.8863	0.465	0.003861	$2.797 \cdot 10^{-9}$	$1.314 \cdot 10^{-9}$	$< 2.2 \cdot 10^{-16}$
4				0.05502	0.003513	$2.391 \cdot 10^{-9}$	$1.274 \cdot 10^{-9}$	$< 2.2 \cdot 10^{-16}$
5					0.2493	$4.092 \cdot 10^{-5}$	$9.864 \cdot 10^{-6}$	$< 2.2 \cdot 10^{-16}$
6						0.01567	0.003955	$3.263 \cdot 10^{-13}$
7							0.3088	$3.729 \cdot 10^{-12}$
8								$1.647 \cdot 10^{-6}$

Table 7 shows the difference of expenditures on housing in the case of various sizes of municipalities.

Generally, we can say that smaller villages and towns (less then circa 5000 inhabitants) do not differ but greater rather do. But it does not hold without exception. Even in the case of bigger towns the differences are sometimes insignificant.

4 Conclusions

Analysis of the data from a sample survey “Living conditions in 2005” elicits that the monthly expenditures on housing in 2005 significantly depend on the type and size of the municipality and the type of dwelling according to its ownership.

Testing the differences between different groups, we found out that expenditure on housing varies in all types of municipalities. But households living in their own home or their own flat do not differ in their expenditure on housing. Households in the category 6 (other legal ground of use) do not differ from any category. But into this category belong only 7 households. Thus this result

should not be taken very seriously. Small villages and towns (less than circa 5000 inhabitants) do not differ but larger towns rather do. But this fact does not hold without an exception. Even in the case of bigger towns the differences are insignificant in a few cases.

Acknowledgement

The research was supported by project of Grant Agency of the Czech Republic no. 420/09/0515 with title: “Analysis and modelling of financial power of Czech and Slovak Households”.

References

- [1] BARTOŠOVÁ, J.: *Analysis and Modelling of Financial Power of Czech Households*. In APLIMAT – Journal of Applied Mathematics, Vol. 2, Nr. 3, Slovak Technical University, Bratislava, pp. 31-36, 2009.
- [2] BARTOŠOVÁ, J., NOVÁK, M.: *Analýza ekonomického chování sektoru domácností v České republice z hlediska zadluženosti*. In MSED – Mezinárodní statisticko-ekonomické dny na VŠE, Vysoká škola ekonomická, Praha, [CD-ROM], pp. 1-6, 2009.
- [3] CHAUDHURI, S. – RAVALLION, M. (1994): How well do static indicators identify the chronically poor. *J. Public Economic* 53, pp. 367-394.
- [4] HRONOVÁ, S., HINDLS, R.: *Ekonomické chování sektoru domácností ČR – spotřeba a zadluženost*. *Statistika* 3/2008, Český statistický úřad, Praha, pp. 189-204, 2008.
- [5] LONGFORD, N.T., PITTAU, M.G.: *Stability of household income in European countries in the 1990s*. *Computational Statistics & Data Analysis* 51, pp. 1364-1383, 2006.
- [6] PAAP, R. – van DIJK, H.K. (1998): Distribution and mobility of wealth of nation. *European Economic Review* 42, pp.1269-1293.
- [7] STANKOVIČOVÁ, I., BARTOŠOVÁ, J.: *Príspevok k analýze subjektívnej chudoby v SR a ČR*. In FORUM STATISTICUM SLOVACUM 3/2009, Slovenská statistická a demografická spoločnosť, Bratislava, [CD-ROM], pp. 1-12, 2009.
- [8] STANKOVIČOVÁ, I., VOJTKOVÁ, M.: *Viacrozmerné štatistické metódy s aplikáciami*. Iura Edition, Bratislava, 261 p., 2007.

Current address

Jitka Bartošová, RNDr., PhD.

University of Economics Prague,
Faculty of Management, Jarošovská 1117/II, Jindřichův Hradec, 377 01, Czech Republic,
tel.: +420 384 417 221,
email: bartosov@fm.vse.cz

Vladislav Bína, Ing.

University of Economics Prague,
Faculty of Management, Jarošovská 1117/II, Jindřichův Hradec, 377 01, Czech Republic,
tel.: +420 384 417 221,
email: bina@fm.vse.cz

INCOME AND EXPENDITURE OF CZECH HOUSEHOLDS

BARTOŠOVÁ Jitka, (CZ), NOVÁK Michal, (CZ)

Abstract. The paper focuses on the analysis of income and expenditure of Czech households with regard to their indebtedness. The main goal is capturing the global differences between indebted and money saving households. Attention is also given to the nature of the distribution of income and expenditures and changes in the structure of household expenditure that occurred in the period 2002-2007. The paper has been created with the support of the grant project of the Grant Agency of the Czech Republic 402/09/0515 "Analysis and Modelling of Financial Power of Czech and Slovak Households".

Keywords. Households, Income, Statistical analysis, Expenditures, Indebtedness

Mathematics Subject Classification: Primary 46N30; Secondary 62E15

1 Introduction

Income and expenditure of households are one of the main indicators of macroeconomic stability of a state. Especially at the beginning of the 21st century, there were significant changes in the economic behavior of households in the Czech Republic (see [1], [2]). There is a talk about the ever-increasing indebtedness of households in the last few years which many journalists show in the light of impending disaster.

Compared with the rest of the Europe is a measure of household indebtedness of Czech households relatively low (see [3], [6]). Due to the ongoing financial crisis and a higher risk of job loss, too high financial obligations of the individual may have fatal impact not only on themselves but also on the entire household. This paper deals with the analysis of total household expenditure data file of the Czech Statistical Office - Household Budget Survey (HBS).

Drahomíra Dubská deals in her article "Czech Households: Escalating consumption and strong changes in the structure of savings during period 1995-2004" with income, expenditure and total indebtedness of households [5]. Another relevant article is a contribution from Jana Čermáková entitled "Vliv diferenciace příjmů na strukturu výdajů domácností" [4], which uses an analysis of data directly from the HBS. Luboš Smrčka in his article "Zadlužení rodin – klíčové téma současnosti" (2008) analyzes indebtedness of Czech families and places them in context with other data of the Czech economy [7].

2 Description of the data base and application software

Data that was used for analysis come from a sample survey of the Czech Statistical Office, called the Family Budget Survey (HBS). This is an investigation pursued by the management of private households and provides information on their spending and consumption patterns, together with information on consumption patterns of households surveyed. Identified information can be used for strategic, tactical and operational decisions in the social policy of the state or for further research.

Used data set contains fourteen types of expenditures. Specifically, the expenditure on:

- food and soft drinks,
- alcoholic beverages and tobacco,
- clothing and footwear,
- housing, water, energy, fuel,
- furnishings and household equipment,
- health,
- transportation,
- post and telecommunications,
- recreation and culture,
- education,
- catering and accommodation,
- other goods and services,
- insurance, financial, administrative and other services,
- taxes.

For basic data analysis was used spreadsheet program Microsoft Excel 2007. Advanced statistical processing was done in the “R” statistical program.

3 The structure and distribution of income and expenditure of Czech households

3.1 Structure of expenditure

Structure of household expenditure is largely stable, however changes in prices also changes the ratio of expenditure incurred for the acquisition of different types of commodities. Household consumption behavior is influenced not only by prices but also by the interest rate and by various socio-economic and demographic factors. The current consumption is influenced by their current values, by past values and by (expected) future values.

Table 1 is made of horizontal and vertical analysis of total household expenditures in 2002, 2005, 2006 and 2007. Table shows the percentage representation of different groups in the total expenditure budget of households. The table shows that in 5 years, from 2002 to 2007, the structure of household expenditures is substantially unchanged. Most noticeable change is seen in spending on food and soft drinks, which took place between 2006 and 2007 to a relative decline in spending (from 18.51 % in 2002 to 16.94 % in 2006 and 16.96 % in 2007). This trend can be attributed to reduction in spending due to the relative cheaper compared to other essential food items in the family budget. On the contrary, the relative increase is seen primarily in travel expenses (from 8.39 % in 2002 to 9.26 % in 2005, 9.25 % in 2006 and 9,17 % in 2007). This can be explained by the increasing financial demands on mobility and the shift in the thinking of people who are willing

to commute to work. Also, household expenditure on other goods and services relative increase - from 2002 to 2007 they increased spending by 1.22 % (from 7.90 % to 9.12 %). Slightly increasing trend shows the proportion of expenditure on health increased by 0.45 % (from 1.47 % to 1.92 %). Interesting is also the difference in tax burden, which in 5 years declined by 0.7 % (from 7.28 % to 6.58 %).

Table 1. Structure of total monthly household expenditures in 2002, 2005, 2006 and 2007.

expenditure structure	food soft drinks	alcoholic beverages tobacco	clothing footwear	housing water, energy, fuel	furnishings household equipment	health	transportation	post telecommunication	recreation culture	education	catering accommodation	other goods services	insurance	taxes
2007	16.96 %	2.41 %	4.55 %	16.77 %	5.96 %	1.92 %	9.17 %	3.93 %	8.90 %	0.50 %	4.37 %	9.12 %	8.88 %	6.58 %
2007-2006	0.02 %	0.02 %	-0.04 %	-0.62 %	0.16 %	0.20 %	-0.09 %	-0.07 %	0.27 %	0.04 %	0.10 %	0.34 %	0.03 %	-0.36 %
2006	16.94 %	2.39 %	4.59 %	17.39 %	5.80 %	1.72 %	9.25 %	4.00 %	8.63 %	0.46 %	4.26 %	8.78 %	8.85 %	6.94 %
2006-2005	-0.21 %	0.03 %	-0.07 %	0.68 %	0.27 %	0.09 %	-0.01 %	0.21 %	-0.20 %	0.01 %	0.03 %	0.35 %	-0.28 %	-0.89 %
2005	17.14 %	2.36 %	4.66 %	16.71 %	5.53 %	1.64 %	9.26 %	3.78 %	8.83 %	0.45 %	4.23 %	8.43 %	9.13 %	7.83 %
2005-2002	-1.37 %	-0.23 %	-0.78 %	0.15 %	-0.24 %	0.17 %	0.87 %	0.43 %	0.01 %	-0.02 %	-0.04 %	0.53 %	-0.04 %	0.55 %
2002	18.51 %	2.59 %	5.43 %	16.56 %	5.77 %	1.47 %	8.39 %	3.36 %	8.82 %	0.48 %	4.28 %	7.90 %	9.17 %	7.28 %
2007-2002	-1.55 %	-0.18 %	-0.89 %	0.21 %	0.19 %	0.45 %	0.78 %	0.57 %	0.08 %	0.02 %	0.09 %	1.21 %	-0.29 %	-0.70 %

3.2 Basic characteristics of income and expenditure

This paper focuses on the analysis of total income and expenditure of households, which are the sum of all incremental income and expenditure. Puts to the fore the question of whether the consumer behavior of Czech households prevailed the trend of debt or money saving in 2007. From a global perspective the behavior of households can be described in a given year by capturing the global characteristics.

Tables 2 and 3 contain information of the level and variability of income and expenditures and their difference. Table 2 shows that both the average and the median income of households have always exceeded the value of the corresponding expenditure characteristics. Similarly, they were also at the maximum and minimum values. For example, the average monthly expenditure of Czech households in 2007 is amounted to approximately 98 % of average income. Even more interesting are the minimum and maximum income and expenditure. Minimum expenses constituted 78 % of the minimum income and maximum expenditure reached even this year, only 52 % of the maximum income. From Table 3, which contain the information about absolute and relative variability of income and expenditure, can be seen that the costs has slightly higher relative variability (coefficient of variation is a practically identical, differing only by 0.2 %, but the relative quartile deviation is about 5 % higher).

Table 2. The level of monthly income and expenditure of Czech households in 2007 and their comparison.

Level characteristics	Level of income and expenditure		Comparison of levels of income and expenditure	
Mean (CZK)	income	24199	difference (expenses – income)	-516
	expenditure	23683	proportion (expenses / income)	97.9 %
Median (CZK)	income	21743	difference (expenses – income)	-677
	expenditure	21066	proportion (expenses / income)	96.9 %
Minimum (CZK)	income	3210	difference (expenses – income)	-702
	expenditure	2508	proportion (expenses / příjmy)	78.1 %
Maximum (CZK)	income	208187	difference (expenses – income)	-100181
	expenditure	108006	proportion (expenses / income)	51.9 %

Table 3. Variability of monthly income and expenditure of Czech households in 2007 and their comparison.

Variability characteristics	Variability of income and expenditure		Comparison of variability of income and expenditure	
Standard deviation (CZK)	income	13803	difference (expenses – income)	- 236
	expenditure	13567	proportion (expenses / income)	0.98
Quartile deviation (CZK)	income	7996	difference (expenses – income)	793
	expenditure	8789	proportion (expenses / income)	1.10
Coefficient of variation (%)	income	57.04 %	difference (expenses – income)	0.2 %
	expenditure	57.29 %	proportion (expenses / income)	1.004
Relative quartile deviation (%)	income	34.30 %	difference (expenses – income)	5.1 %
	expenditure	39.42 %	proportion (expenses / income)	1.15

3.3 Distribution of income and expenditure

For further analysis of income and expenditures will be appropriate to use a logarithmic transformation. The reason for using this transformation is both skew distribution and a significant proportion of both right outliers, i.e. very high income, respectively expenditure. Distribution of income and expenditure is approximately log-normal, so by the logarithmic transformation we get approximately normal distribution and we gain elimination of a large portion of outliers. Changes in the distribution of income and expenditure in 2007 after logarithmic transformation, including the QQ-graphs documenting the normality of transformed data, are presented in Charts 2 and 3.

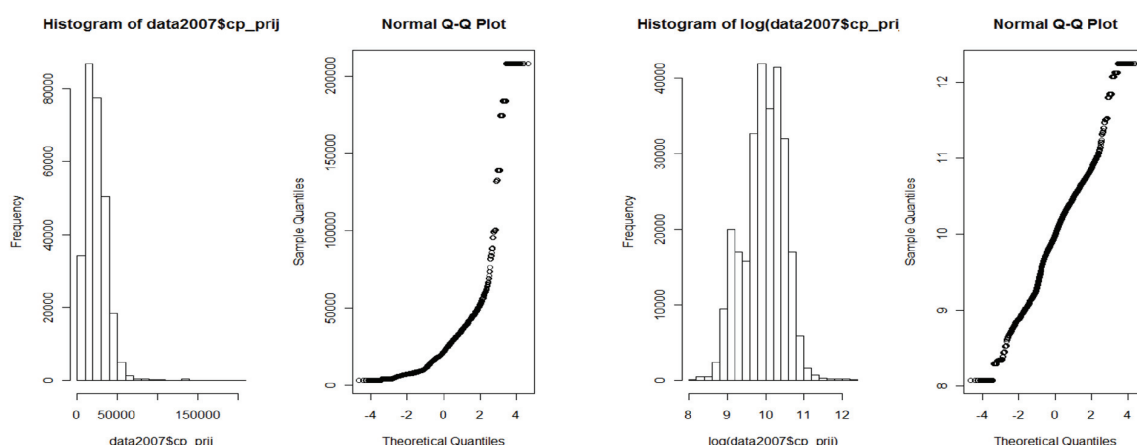


Figure 2. Diagnostic plots of the distribution of total household income in 2007; left – original; right – after transformation. (Data source: HBS)

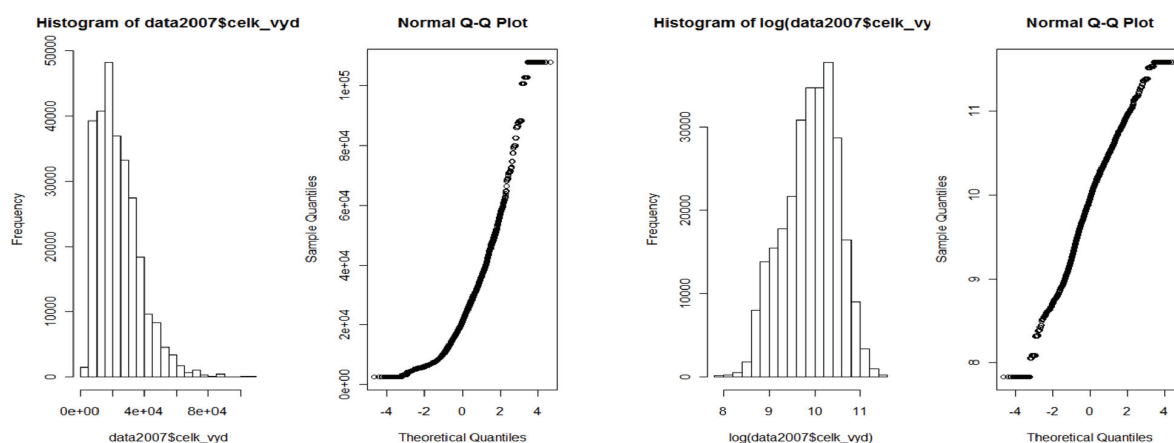


Figure 3. Diagnostic plots of the distribution of total household expenditures in 2007; left – original; right – after transformation (Data source: HBS)

4 Saving and indebting households

4.1 Income and expenditure compliance survey

In terms of the level of total income and expenditure of households can be divided into two groups - a saving household, i.e. those that have total incomes exceeding total expenditures, and indebting households are those who have total expenditures higher than total income. Monitoring characteristics of households in the two groups separately and determining their differences enables us to identify factors that affect the classification of households into saving or indebting group.

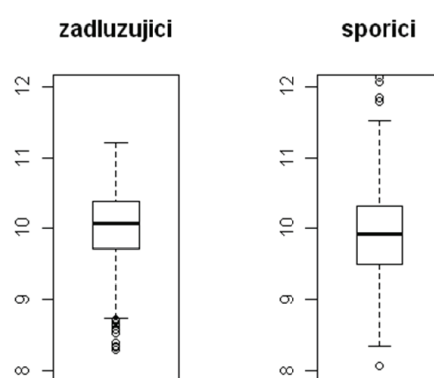


Figure 1. Boxplots of total income distribution of indebting and saving households; left – indebting; right – saving. (Data source: HBS)

An interesting and unexpected result is the finding that a graphic comparison of income of saving households and indebting households leads to the conclusion that indebting households had a median income higher than the total household savings in 2007 (see Figure 1). One possible explanation is that households with higher income have a better opportunity to obtain and repay the loan, and therefore better placed to manage the risks associated with debt. Figure 1 also shows apparent differences of income variability in both groups.

Both hypotheses are still needed to confirm or disprove quantitatively. Therefore, the compliance tests performed mean values and variances of intergroup consensus income for both types of households. Results of Levene's test of homogeneity of variance showed that the variance of total income in 2007 for saving a indebting households are different (F statistic = 7.6677, p-value = 0.005657). Among other things, this means that there is a breach of the condition of homoscedasticity and compliance testing of high income is necessary to use Welch's test, which compares the median values for unequal variance. It is therefore necessary to calculate an approximation of the number of degrees of freedom of the referenced t-distribution. Welch's test proved that in 2007, median of total income of saving and indebting households is different (p-value < $2.2 \cdot 10^{-16}$).

Similarly, we proceed in determining whether the saving and indebted households are also distinguished by their household expenditure. Before the quantitative determination of high compliance costs again perform for both types of households, first compliance testing intergroup variances expenditure. On the basis of homogeneity of variance test Levene's managed to show that the variance of the total expenditure for saving and indebted households in 2007 was different (F statistic = 7.6677, p -value = 0.005657). Welch's test also showed that the median of total expenditure of indebted and saving households was different in 2007 (p -value < $2.2 \cdot 10^{-16}$).

4.2 Distribution of income and expenditure

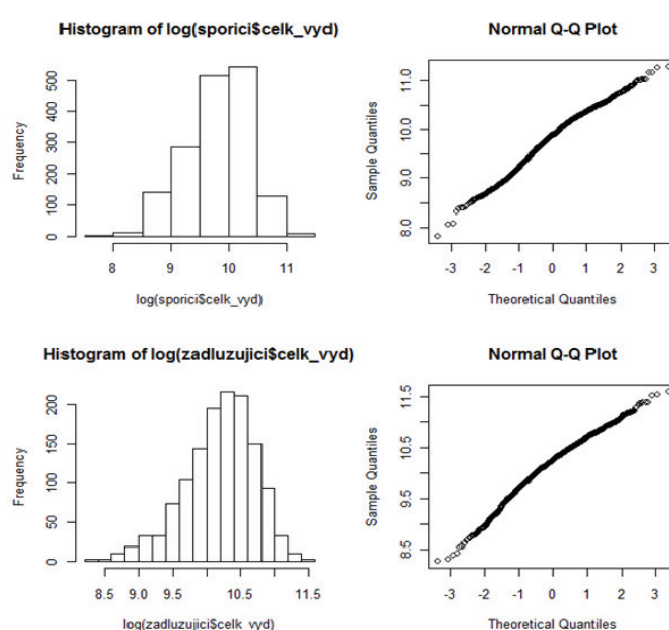


Figure 4. Diagnostic plots of the distribution of total expenditure in 2007 after logarithmic transformation; above – saving households; under – indebted households. (Data source: HBS)

Like the distribution of income and expenditure of all Czech households, also in case of division of indebted and saving households it is slanted and contains the right outliers. We have to transform the data before further analyses. There is the distribution of expenditures in graph 4 after logarithmic transformation in the group of saving and group of indebted households.

5 Conclusions

Presented article followed several goals. It deals with identifying changes in the structure of Czech households' expenditure in the period 2002–2007. Furthermore, the article dealt with characterization of the distribution of income and expenditure and last but not least there was analyzed the diversity of income and expenditure of indebted and saving households.

The above analysis used samples from the HBS for the years 2002, 2005, 2006 and 2007, which shall contain information of the income and expenditure as well as a number of other socio-economic and demographic characteristics of households. Graphically displayed data confirmed the assumption that income and expenditure of households have approximately log-normal distribution with the right amount of outliers. To maintain the assumption of normality distribution, which requires some analysis, it is appropriate to work on the analysis of the logarithms of income and expenditure. It also occurs to a significant reduction or even elimination of values that are considered outlying.

On the basis of horizontal and vertical analysis of total household expenditures in 2002, 2005, 2006 and 2007 can be argued that the structure of expenditure in those years did not substantially change. Interesting can be considered that there is a decrease in expenditure on food and soft drinks, which can be explained by the relative cheaper food compared with other necessary items in the family budget. Percentage of other types of expenses is generally stagnant or slightly increasing. The largest increase can be seen in travel expenses and other (unspecified) goods and services. The growth of travel expenses related to the increasing financial demands on mobility and the shift in the thinking of people who are willing to commute to work. It is also interesting that the proportion of expenditure on health has a slightly upward trend, unlike the tax burden, which in turn decreased in this period.

When analyzing the total income and total expenditure of households from HBS in 2007 we learned that the total income of Czech households have been higher this year than the total expenditure. This confirms the fact that in 2007 there was a higher growth of total savings than the growth of household debt.

After the division of households into those whose incomes are less than the expenditure (in text as indebted households) and those whose income is higher than the total expenditure (in text as saving households), was detected statistically significant difference between the two groups, namely both the level and variability of income and expenditure. Be considered interesting finding that indebted households in 2007 had a median of total income higher than the saving households. This result is likely to be subject to their better ability to earn and repay loans, which of course increase their tendencies to run the risk of debt.

Acknowledgement

The research was supported by project of Grant Agency of the Czech Republic no. 420/09/0515 with title: "Analysis and modelling of financial power of Czech and Slovak Households".

References

- [1.] BARTOŠOVÁ, J.: *Income Distribution in the Czech Republic after Velvet Revolution*. In Kováčová, M. (ed.) 5th International Conference APLIMAT 2006, Bratislava, February 7-10, 2006, Part I, Slovak University of Technology, Bratislava, pp. 417-423, 2006.
- [2.] BARTOŠOVÁ, J.: *Analysis and Modelling of Financial Power of Czech Households*. In APLIMAT – Journal of Applied Mathematics, Vol. 2, Nr. 3, Slovak Technical University, Bratislava, pp. 31-36, 2009.

- [3.] BARTOŠOVÁ, J., NOVÁK, M.: *Analýza ekonomického chování sektoru domácností v České republice z hlediska zadluženosti*. In MSED – Mezinárodní statisticko-ekonomické dny na VŠE, Vysoká škola ekonomická, Praha, [CD-ROM], 2009, pp. 1-6,.
- [4.] ČERMÁKOVÁ, J.: *The influence of income differentiation on the structure of household expenditures*. Finance a úvěr, vol. 51, issue 1, pp. 33-45, 2001.
- [5.] DUBSKÁ, D.: *Czech households: Escalating consumption and strong changes in the structure of savings during period 1995 – 2004*. ČSÚ Praha, 2006. URL: <http://panda.hyperlink.cz/cestapdf/pdf06c2/dubska.pdf>
- [6.] HRONOVÁ, S., HINDLS, R.: *Ekonomické chování sektoru domácností ČR – spotřeba a zadluženost*. Statistika 3/2008, Český statistický úřad, Praha, pp. 189-204, 2008.
- [7.] SMRČKA, L.: *Zadlužení rodin – klíčové téma současnosti*. Ekonomika a management [online], Vol. 1, Nr. 11. 2008.

Current address

Jitka Bartošová, RNDr., PhD.

University of Economics Prague,
Faculty of Management, Jarošovská 1117/II, Jindřichův Hradec, 377 01, Czech Republic,
tel.: +420 384 417 221,
e-mail: bartosov@fm.vse.cz

Michal Novák, Bc.

University of Economics Prague,
Faculty of Management, Jarošovská 1117/II, Jindřichův Hradec, 377 01, Czech Republic,
tel.: +420 384 417 221,
e-mail: xnovm132@fm.vse.cz

ECONOMIC FORECASTS WITH BAYESIAN AUTOREGRESSIVE DISTRIBUTED LAG MODEL: CHOOSING OPTIMAL PRIOR IN ECONOMIC DOWNTURN

BUŠS Ginters, (LV)

Abstract. Bayesian inference requires an analyst to set priors. Setting the right prior is crucial for precise forecasts. By using an autoregressive distributed lag model, this paper analyzes how optimal Litterman prior changes when an economy is hit by a recession. The results show that a sharp economic slowdown changes the optimal prior in two directions. First, it changes the structure of the optimal weight prior by setting smaller weight on the lagged dependent variable compared to variables containing more recent information. Second, greater uncertainty brought by a rapid economic downturn requires more space for coefficient variation which is set by the overall tightness parameter. It is shown that the optimal overall tightness parameter may increase to such an extent that Bayesian ADL becomes equivalent to frequentist ADL.

Key words and phrases. Forecasting, Bayesian inference, Bayesian autoregressive distributed lag model, optimal prior, Litterman prior, business cycle, mixed estimation.

Mathematics Subject Classification. Primary 91B84; Secondary 62P20.

1 Introduction

Bayesian inference requires an analyst to set a prior. Setting the right prior is crucial for precise forecasts. This paper analyzes how optimal Litterman prior changes when an economy is hit by a recession. By an 'optimal Litterman prior' in this paper we define Litterman hyperparameters that minimize the root mean squared error from one-period ahead forecasts.

Although the question about what hyperparameters to use has been addressed in a series of papers by, among others, Litterman and coauthors (Litterman (1979), Doan, Litterman and Sims (1984), Litterman (1986)) and LeSage and coauthors (LeSage and Magura (1991), LeSage and Pan (1995), LeSage and Krivelyova (1999)), the role of a business cycle on the optimal prior,

to the best of our knowledge, has not been discussed. Thus, this paper analyzes how (if any) prior hyperparameters should be altered for the best one-period ahead forecasting performance when there is a switch in a phase of a business cycle. For this task, an autoregressive distributed lag model (ADL) is chosen. The prior is set up like in Litterman (1979). The model is solved by ‘mixed estimation’ set forth in Theil and Goldberger (1961). Latvia’s gross domestic product (GDP) was found to be well suited for the analysis. The results show that a sharp economic slowdown changes the optimal prior in two directions.

First, a lagged dependent variable loses its dominance as the key explanatory variable and, instead, more current information contained in leading indicator-type variables is of greater importance to improve forecasts. This changes the structure of the optimal weight prior, setting smaller weight on the lagged dependent variable compared to variables containing more recent information.

Second, greater uncertainty brought by a swift economic downturn requires more space for coefficient variation, which is set by the overall tightness parameter. Particularly, the results show that, in economic downturn, the optimal overall tightness parameter may increase to such an extent that Bayesian ADL becomes equivalent to frequentist ADL, which may imply that a greater uncertainty in an economy requires more skills from an analyst to set the right prior such that, during great economic uncertainty, one may become more comfortable using frequentist rather than Bayesian inference.

The paper is organized as follows. Section 2 describes the model and its estimation procedure. Section 3 presents the results from a case study. Finally, Section 4 concludes.

2 Methodology

2.1 The Model

Consider an autoregressive distributed lag model (ADL) of order (p, q) :

$$y_t = \sum_{m=1}^p \beta_m y_{t-m} + \sum_{n=0}^q \gamma'_n x_{t-n} + \xi' z_t + \epsilon_t \quad (1)$$

where y_t is the dependent variable, x_t is a $d \times 1$ vector of key explanatory variables $x = [x_1 \ x_2 \ \dots \ x_d]$, z_t is (a vector of) other explanatory variable(s) potentially containing a constant, a dummy variable for an outlying effect, etc., and $\epsilon_t \sim N(0, \sigma^2)$. The Bayesian prior is set to

$$\begin{aligned} \beta_m &\sim N(\chi_{\{1\}}(m), \sigma_m^2) \\ \gamma_{in} &\sim N(0, \sigma_{in}^2) \end{aligned} \quad (2)$$

where $\chi_{\{1\}}()$ is an indicator function, $m = 1, 2, \dots, p$, $i = 1, 2, \dots, d$, and $n = 0, 1, \dots, q$. The specification of the standard deviation of the prior is *à la* Doan, Litterman and Sims (1984):

$$\begin{aligned} \sigma_m &= \theta k m^{-\phi} \\ \sigma_{in} &= \theta l (1 + n)^{-\phi} \left(\frac{\hat{\sigma}_{u,i}}{\hat{\sigma}_{u,y}} \right) \end{aligned} \quad (3)$$

where $\hat{\sigma}_{u,y}$ and $\hat{\sigma}_{u,i}$ are the standard errors from a univariate autoregression involving y and x_i , respectively, so that $\hat{\sigma}_{u,i}/\hat{\sigma}_{u,y}$ is a scaling factor that adjusts for varying magnitudes of the involved variables. The parameter θ is referred as the overall tightness. The terms $m^{-\phi}$ and $(1+n)^{-\phi}$ are referred as lag decay functions for y and x_i , respectively, with $\phi \geq 0$ reflecting a shrinkage of the standard deviation with increasing lag length. The parameters k and l specify the relative tightness of the prior for variables y and x_i , respectively. Note that, for simplicity, we set l the same for all x_i .

2.2 Estimation

The model (1) to (3) can be estimated using the ‘mixed estimation’ method set forth in Theil and Goldberger (1961). For ease of exposition, drop z_t from (1) and rewrite it as

$$y = X\beta + \epsilon \quad (4)$$

where y is the $T \times 1$ vector of observations on the dependent variable, X the $T \times (p + (q + 1)d)$ matrix of observations on the explanatory variables with rank $p + (q + 1)d$, β the $(p + (q + 1)d) \times 1$ vector of coefficients, and ϵ the $T \times 1$ vector of disturbances such that

$$E\epsilon = 0, \quad \Sigma := E(\epsilon\epsilon') = \sigma^2 I_{T \times T}. \quad (5)$$

The Bayesian prior is included in

$$r = R\beta + \nu, \quad (6)$$

where r is a $(p + (q + 1)d) \times 1$ vector $[1 \ 0 \ 0 \ \dots \ 0]'$, R is a $(p + (q + 1)d) \times (p + (q + 1)d)$ identity matrix, and ν is a $(p + (q + 1)d) \times 1$ vector of disturbances such that

$$E\nu = 0 \quad (7)$$

and $E(\nu\nu')$ is a $(p + (q + 1)d) \times (p + (q + 1)d)$ diagonal matrix with diagonal elements being the variances specified in (3),

$$\Omega := E(\nu\nu') = \begin{bmatrix} \sigma_1^2 & 0 & & \dots & & 0 \\ 0 & \sigma_2^2 & & & & \\ 0 & 0 & \ddots & & & \\ & & & \sigma_p^2 & & \\ \vdots & & & \sigma_{10}^2 & & \vdots \\ & & & & \sigma_{11}^2 & \\ & & & & & \ddots \\ & & & & & & \sigma_{d,q-1}^2 & 0 \\ 0 & & & \dots & & & 0 & \sigma_{dq}^2 \end{bmatrix} \quad (8)$$

The sample and the independent extraneous information may be combined by writing

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} u \\ \nu \end{bmatrix}; \quad E \begin{bmatrix} u \\ \nu \end{bmatrix} = 0; \quad E \left(\begin{bmatrix} u \\ \nu \end{bmatrix} \begin{bmatrix} u' & \nu' \end{bmatrix} \right) = \begin{bmatrix} \Sigma & 0 \\ 0 & \Omega \end{bmatrix}. \quad (9)$$

An application of generalized least squares (GLS) procedure leads to estimating β as

$$\hat{\beta} = \left([X' \ R'] \begin{bmatrix} \Sigma & 0 \\ 0 & \Omega \end{bmatrix}^{-1} \begin{bmatrix} X \\ R \end{bmatrix} \right)^{-1} [X' \ R'] \begin{bmatrix} \Sigma & 0 \\ 0 & \Omega \end{bmatrix}^{-1} \begin{bmatrix} y \\ r \end{bmatrix} \quad (10)$$

or

$$\hat{\beta} = [X' \Sigma^{-1} X + R' \Omega^{-1} R]^{-1} [X' \Sigma^{-1} y + R' \Omega^{-1} r]. \quad (11)$$

Normalizing R :

$$\tilde{R} := \begin{bmatrix} \frac{\sigma}{\sigma_1} & 0 & & \dots & & 0 \\ 0 & \frac{\sigma}{\sigma_2} & & & & \\ & & \ddots & & & \\ & & & \frac{\sigma}{\sigma_p} & & \\ \vdots & & & & \frac{\sigma}{\sigma_{10}} & \vdots \\ & & & & \frac{\sigma}{\sigma_{11}} & \\ & & & & & \ddots & \\ & & & & & & \frac{\sigma}{\sigma_{d,q-1}} & 0 \\ 0 & & & \dots & & & 0 & \frac{\sigma}{\sigma_{dq}} \end{bmatrix}$$

and r :

$$\tilde{r} := \begin{bmatrix} \frac{\sigma}{\sigma_1} & 0 & 0 & \dots & 0 \end{bmatrix}$$

gives $E(\nu\nu') = \sigma^2 I$, and the GLS estimator in (11) reduces to an ordinary least squares estimator:

$$\hat{\beta} = [X' X + \tilde{R}' \tilde{R}]^{-1} [X' y + \tilde{R}' \tilde{r}]. \quad (12)$$

3 Results

The dependent variable of the model (1) is Latvia's quarterly GDP series from 1995Q1 till 2009Q1. The key explanatory variables x are two quarterly series, the output in manufacturing industry (according to Nace revision 1.1 subsequently called D) and output in electricity, gas and water supply industry (E). All three series are chained priced as of year 2000 and twice regularly and once seasonally differenced. The second regular difference is performed for better forecasting performance during the latter part of the GDP series due to a sharp economic downturn (see Buss, 2009 for a discussion). Series D and E are published before the GDP flash estimate is released, thus we can potentially use these series to forecast GDP before its other components are known. The model may contain a constant and other explanatory variables, all contained in z in (1). All calculations are performed in Scilab with the aid of its econometrics toolbox Grocer.

Model	RMSE	RMSE1sthalf	RMSE2ndhalf
SARMA(01)(01)	0.0328737	0.0160291	0.0436398
AR(1)	0.0275043	0.0194567	0.0336810
AR(2)	0.0263058	0.0203990	0.0311106
FADL(1,0)(D)	0.0277540	0.2011203	0.0330832
FADL(2,0)(D)	0.0289995	0.0272706	0.0306310
FADL(2,1)(D)	0.0253833	0.0196827	0.0300202
FADL(2,1)(E)	0.0257016	0.0216257	0.0292142
FADL(2,1)(D+E)	0.0247125	0.0220415	0.0271218
FADL(3,2)(D)	0.0260984	0.0216730	0.0298754
FADL(3,2)(E)	0.0257382	0.0217008	0.0292230
FADL(3,2)(D+E)	0.0253316	0.0251711	0.0254912
BADL(2,1)(D+E)(.95,.1,.8,0)	0.0239113	0.0196482	0.0275217
BADL(2,1)(D+E)(.05,1,2,0)	0.0264237	0.0258526	0.0269828
BADL(3,2)(D+E)(1,.35,.2,0)	0.0223288	0.0171109	0.0265400
BADL(3,2)(D+E)(.8,.25,.2,0)	0.0225414	0.0166686	0.0271732

Table 1: A brief comparison of SARMA, AR, FADL and BADL. The two latter models are specified by their orders, (p, q) , key exogenous variables, e.g. (D+E), and the Bayesian ADL with a single key exogenous variable is specified by its prior, (k, l, θ, ϕ) , where prior weight $w := [k \ l]$. The least RMSE in each sample space is framed.

3.1 Warm-up

To start, Table 1 shows root mean squared forecast errors (RMSE) for the whole sample, the first half of the sample (RMSE1sthalf) and the second half of the sample (RMSE2ndhalf) from one-period ahead pseudo real-time forecasts beginning at sample size 17 from simple benchmark seasonal autoregressive moving average model (SARMA), autoregressive models (AR), and frequentist and Bayesian autoregressive distributed lag models (FADL and BADL, respectively) of order (p, q) with explanatory variable in parenthesis. Notation (D+E) means the variables are summed to result in a single explanatory variable. The Bayesian counterpart of ADL requires to specify the hyperparameters for (3), called Litterman prior consisting of four parameters, k , l , θ , and ϕ , with $w := [k \ l]$ for one-dimensional x . The forecasts are called *pseudo* real-time because they are made on the revised values of left-hand-side and right-hand-side variables in (1); although the revisions for the specific variables used in this analysis tend to be relatively small, they might underestimate RMSE. Nonetheless, this does not harm for our purpose.

The sample is split in halves because the first half contains a smooth growth whereas the second half contains rapid economic downturn, so we can analyze how the forecasting performance of the models changes with the business cycle and, especially, how Bayesian prior has to be altered for the best forecasting performance.

The least RMSE in each column is framed. It can be seen that Bayesian ADL models compare well with other models. It can also be seen that the BADL(3,2) models give the most

precise one-period ahead forecasts for the whole sample as well as for the first half of the sample among all the ADL models considered, but they are outperformed by FADL for the second half of the model. This observation suggests that the optimal Bayesian prior might be different for the first half of the model (smooth positive growth) compared to the second half of the sample when there is a rapid economic downturn. We check this hypothesis further by employing grid search for the optimal prior.

3.2 Search for optimal priors

First, the grid search is performed for BADL(2,1)(D+E). The weight vector $[k \ l]$ is 2-dimensional, one element, k , for the dependent variable and one, l , for a single explanatory variable x , both ranging from .05 to 1 with step size .05. The overall tightness, θ , is set to range from .6 to 2.5 with step .1, and the lag decay, ϕ , from 0 to 1 with step .2. So, the grid size is $20 \times 20 \times 20 \times 6$ containing overall 48000 prior combinations for each one-period ahead forecast with sample size ranging from 17 to 51. The minimum RMSE for the whole sample is attained at the coordinate $[19 \ 2 \ 3 \ 1]$ with the corresponding values $[k \ l \ \theta \ \phi] = [.95 \ .1 \ .8 \ 0]$ with a boundary value at $\phi = 0$. The boundary for ϕ can not be decreased further since negative values would presume lags of a higher order be more informative which is counterintuitive. Figures 1(a) and 1(b) show the inverse of the RMSE as a function of the prior for the whole sample.

Figure 1 about here

Figure 1(a) shows the inverse of the RMSE as a function of the weight vector (the x and y axes represent k and l , respectively) given the rest of parameters, θ and ϕ , at their RMSE-minimizing values. It can be seen that the values of k have the major impact on the RMSE with acceptable range about (.4,1), otherwise the RMSE increases substantially. On the contrary, values of l have less influence on the RMSE given k , nonetheless, a peak is evident at $l = .1$ for all acceptable values of k .

Similarly, Figure 1(b) shows the inverse of the RMSE as a function of θ and ϕ (representing x and y axes, respectively) given the RMSE-minimizing weight vector. It can be seen that the values of both θ and ϕ have a nontrivial impact on RMSE at its optimum with the maximizing values .8 and 0, respectively. The maximizing value of $\phi = 0$ might be due to the small number of lags, which is one for each RHS variable in this model.

Now, calculating the minimum RMSE for the second half of the sample, the optimum value is attained at the coordinate $[1 \ 20 \ 15 \ 1]$ with the corresponding values $[k \ l \ \theta \ \phi] = [.05 \ 1 \ 2 \ 0]$ with three boundary values for k , l and ϕ . It can already be seen that the optimal prior weight is different compared to the full sample. Figures 1(c) and 1(d) show the inverse of the RMSE as a function of the prior for the second half of the sample. Figure 1(c) looks almost like the inverse of Figure 1(a). Now, the RMSE is increasing with k , with an optimum at the lowest k considered; other values of k would significantly increase the RMSE at all levels of l , the latter being also critical for optimal RMSE with acceptable range about (.3,1), otherwise the forecast error increases substantially. This observation is in line with our hypothesis that, during sharp decline in the economy, explanatory variables containing most recent information are more important than the lagged dependent variable.

Figure 1(d) shows that, for the second half of the sample, the optimal tightness parameter is higher compared to the full sample, with acceptable values in about (1,2.5), otherwise the forecast error increases substantially. This observation is as expected since the model coefficients should be given more flexibility during a rapid change in an economy. For acceptable θ , the values of lag decay parameter, ϕ , is of less importance. The forecasting performance of BADL(2,1)(D+E) for the first half of the sample is not impressive and thus not presented here.

Having explored BADL(2,1)(D+E), we now check the results for BADL(3,2) (D+E) whose forecasting performance for all sample spaces considered, as it can be seen in Table 1, is promising. The grid space is formed by k and l being from .05 to 1 with step .05, θ from .1 to 1 with step .1, and ϕ from 0 to 1 with step .1. The coordinate for the least RMSE for full sample is [20 7 2 1] with the prior values $[k \ l \ \theta \ \phi] = [1 \ .35 \ .2 \ 0]$, showing some resemblance with the results for BADL(2,1)(D+E). The inverse RMSE for full sample around the optimal prior values is shown in Figures 2(a) and 2(b). The behavior of the inverse RMSE around its optimal value is similar to that of BADL(2,1)(D+E).

Figure 2 about here

We can see from Table 1 about the model's BADL(3,2)(D+E) comparatively competitive forecasting performance for the first half of the sample. Figures 2(c) and 2(d) show the inverse RMSE around its optimum as a function of prior parameters for the first half of the sample. We see that the results are similar to the results from a full sample with optimal $k = .8$, $l = .25$, $\theta = .2$ and $\phi = 0$. It can also be seen that l has more influence on the RMSE compared to the full sample, with lowest RMSE concentrating on the lowest part of l space.

Regarding the results for the second half of the sample, the coordinate of the optimal value is [20 20 10 1], with all values being at a boundary and suggesting a greater θ (i.e., more flexibility for coefficient values). An extensive search for the optimal θ resulted to its value around 10^5 with RMSE being the same as for FADL(3,2)(D+E) at least up to and including the 7th digit after a comma, shown in Table 1. The latter result might suggest that during a sharp decline in an economy one might wish to set the overall tightness parameter, θ , so loose that one is more comfortable to use frequentist version of ADL.

4 Conclusions

Bayesian inference requires an analyst to set priors. Setting the right prior is crucial for precise forecasts. This paper analyzes how optimal prior changes with business cycle, specifically, when an economy is hit by a recession. Latvia's GDP is well suited for this analysis. The results show that when economy is growing, the optimal overall tightness parameter is less than one, and the optimal weight vector sets a higher weight on a lagged dependent variable compared to other explanatory variables. However, a swift economic downturn changes the optimal prior considerably in two directions.

First, a lagged dependent variable loses its dominance as the key explanatory variable and, instead, more current information contained in leading indicator-type variables is of greater importance to improve forecasts. This changes the structure of the weight prior, setting smaller weight on the lagged dependent variable compared to variables containing more recent information.

Second, greater uncertainty brought by a rapid economic downturn requires more space for coefficient variation, which is set by the overall tightness parameter. Particularly, the results show that, in economic downturn, the optimal overall tightness parameter may increase to such an extent that Bayesian ADL becomes equivalent to frequentist ADL, which may imply that a greater uncertainty in an economy requires more skills from an analyst to set the right prior such that, during great economic uncertainty, one may become more comfortable using frequentist rather than Bayesian inference.

Acknowledgement

The author is thankful to his supervisor, Viktors Ajevskis, for his guidance and support, as well as to the seminar participants at Riga Technical University for their valuable comments.

References

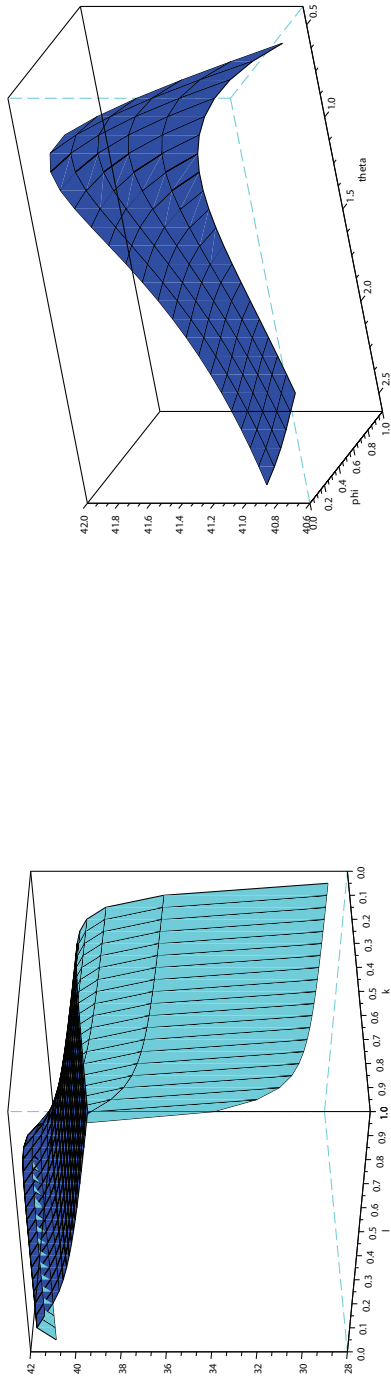
- [1] BUSS, G.: *Comparing Forecasts of Latvia's GDP Using Simple Seasonal ARIMA Models and Direct Versus Indirect Approach*. MPRA Paper 16832, University Library of Munich, Germany, 2009.
- [2] DOAN, T., LITTERMAN, R. B., SIMS C. A.: *Forecasting and Conditional Projection Using Realistic Prior Distributions*. In *Econometric Reviews*, Vol. 3, pp. 1-100, 1984.
- [3] LESAGE, J. P.: *Applied Econometrics using MATLAB*, 1999.
- [4] LESAGE, J. P., KRIVELYOVA, A.: *A Spatial Prior for Bayesian Vector Autoregressive Models*. In *Journal of Regional Science*, Vol. 39, pp. 297-317, 1999.
- [5] LESAGE, J. P., MAGURA, M.: *Using Interindustry Input-Output Relations as a Bayesian prior in Employment Forecasting Models*. In *International Journal of Forecasting*, Vol. 7, pp. 231-238, 1991.
- [6] LESAGE, J. P., PAN, Z.: *Using Spatial Contiguity as Bayesian Prior Information in Regional Forecasting Models*, In *International Regional Science Review*, 18, pp. 33-53, 1995.
- [7] LITTERMAN, R. B.: *Techniques of Forecasting Using Vector Autoregressions*. Working Paper 115, Federal Reserve Bank of Minneapolis, 1979.
- [8] LITTERMAN, R. B.: *Forecasting with Bayesian Vector Autoregressions - Five Years of Experience*. In *Journal of Business & Economic Statistics*, Vol. 4, pp. 25-38, 1986.
- [9] THEIL, H., GOLDBERGER, A. S.: *On Pure and Mixed Statistical Estimation in Economics*. In *International Economic Review*, Vol. 2, pp. 65-78, 1961.

Current address

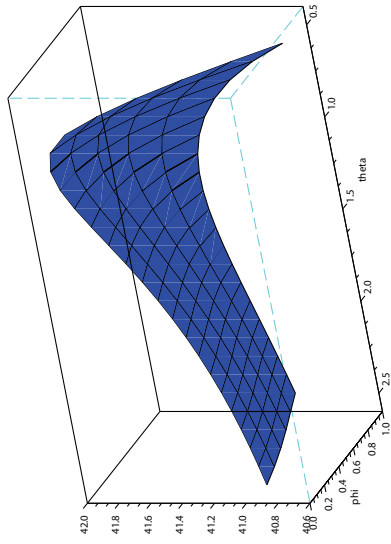
Ginters Bušs, MA

Department of Probability Theory and Mathematical Statistics, Faculty of Computer Science and Information Technology, Riga Technical University, Meza iela 1k4, Riga, LV-1048, Latvia, e-mail: ginters.buss@rtu.lv

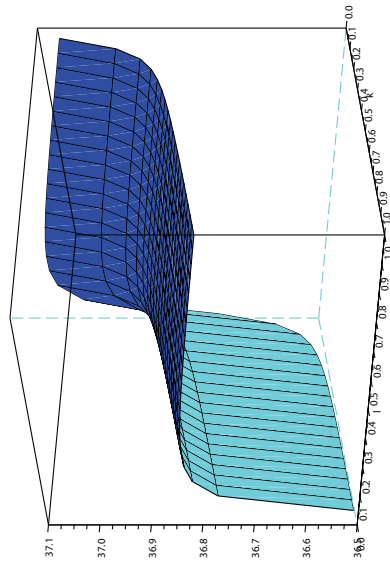
Mathematical Support Division, Central Statistical Bureau of Latvia, Lacplesa iela 1, Riga, LV-1301, e-mail: ginters.buss@csb.gov.lv



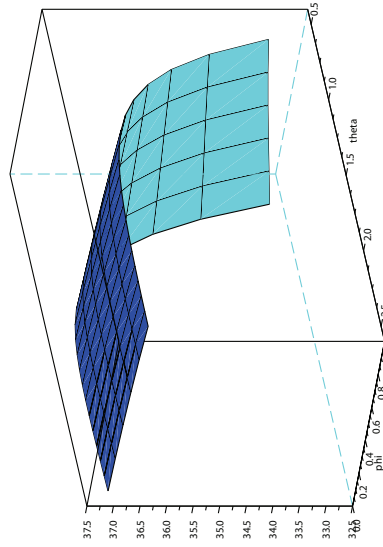
(a) Full sample. Optimal $k = .95$ and optimal $l = .1$.



(b) Full sample. Optimal $\theta = .9$ and optimal $\phi = 0$.



(c) Second half of the sample. Optimal $k = .05$ and optimal $l = 1$.



(d) Second half of the sample. Optimal $\theta = 2$ and optimal $\phi = 0$.

Figure 1: Results from grid search for optimal prior for BADL(2,1)(D+E). Figures 1(a) and 1(b) represent a full sample, whereas Figures 1(c) and 1(d) represent the second half of the sample. The figures on the left (1(a) and 1(c)) show $RMSE^{-1}$ (z axis) as a function of a weight vector (k, l) (x and y axis, respectively) at the RMSE-minimizing θ and ϕ . The figures on the right (1(b) and 1(d)) show $RMSE^{-1}$ (z axis) as a function of θ and ϕ (x and y axis, respectively) at the RMSE-minimizing weight vector.

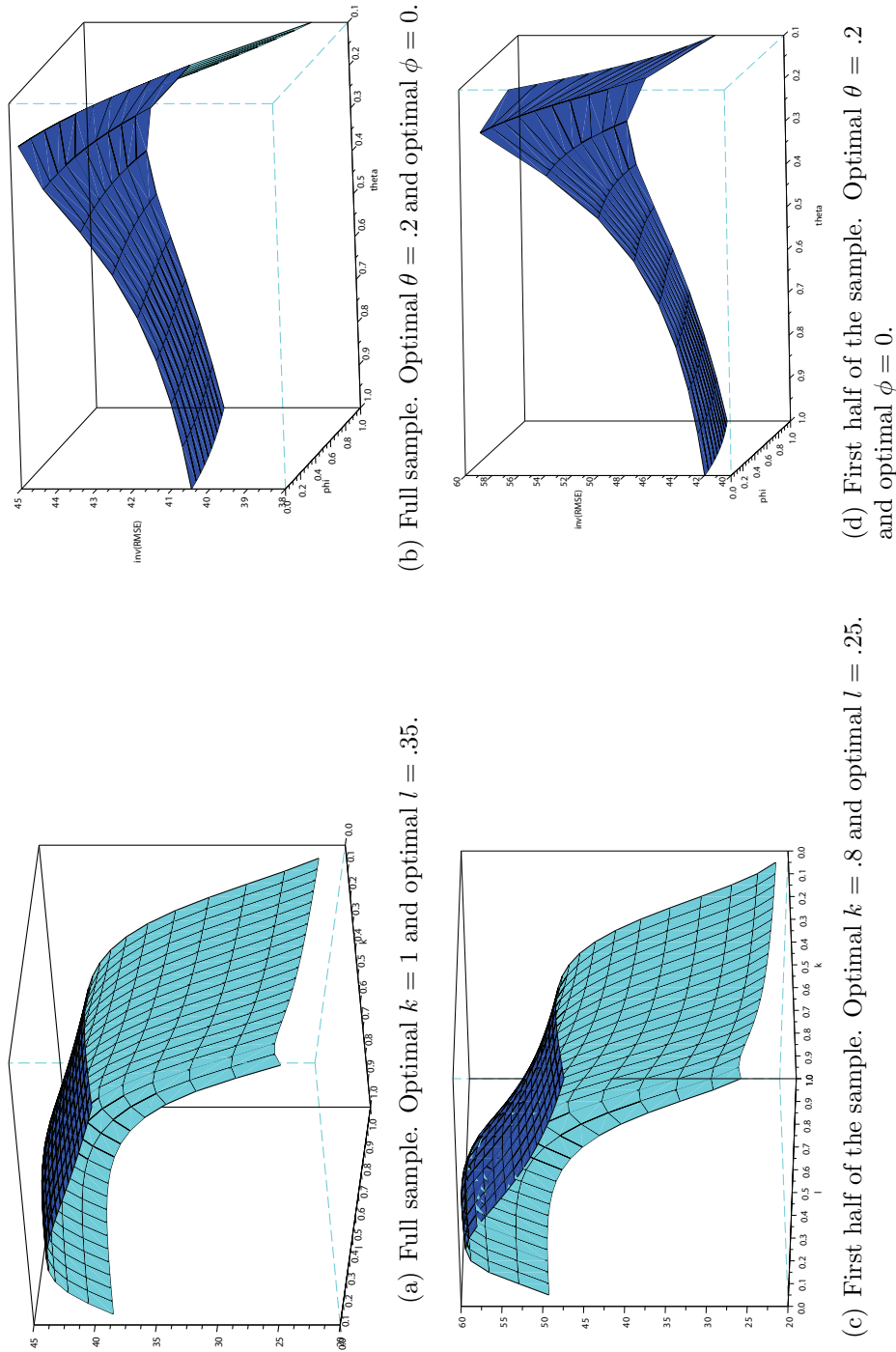


Figure 2: Results from grid search for optimal prior for BADL(3,2)(D+E). Figures 2(a) and 2(b) represent a full sample, whereas Figures 2(c) and 2(d) represent the first half of the sample. The figures on the left (2(a) and 2(b)) show $RMSE^{-1}$ (z axis) as a function of a weight vector (k, l) (x and y axis, respectively) at the RMSE-minimizing θ and ϕ . The figures on the right (2(c) and 2(d)) show $RMSE^{-1}$ (z axis) as a function of θ and ϕ (x and y axis, respectively) at the RMSE-minimizing weight vector.

ANALYSIS OF TREE BARK COMPOSITION IN NON-BALLAST REGIONS

DOLEŽALOVÁ Jarmila, (CZ), VALOVÁ Marie, (CZ)

Abstract. One of the dominant factor, that affect the chemical composition of tree bark, is the air pollution. The analysis of measured and laboratory processed data contributes to the determination of past and present air deposition ballast of the landscape, to the determination of regrettable elements and compounds in the air.

Key words. Chemical composition of tree bark, the influence of air pollutants, statistic data analysis

Mathematics Subject Classification: Primary 62J10, 64-07; Secondary 47N30.

1 Introduction

Air pollution is one of the problems that are intensively discussed these days. The bio-indicators belong between the instruments that help to monitor air pollution. The bryophytes can be used for determination of past and present air deposition ballast of the landscape, to the determination of regrettable elements and compounds in the air [1]. Especially the epiphytic bryophytes belong to the effectual bio-indicators due to their sensibility towards the air quality. The importance of bryophytes as bio-indicators is in their very narrow linkage to the abiotic conditions of the environment. This linkage is caused by their eco-physiological characteristics, when the surface of thallus is giant compared to the surface of bryophyte body, as well as there are no barriers that would inhibit the solutions from penetrating to the physiologically active centres [3]. Their appearance in given area is affected by the chemical composition of air, the representation of chemical elements and organic carbon in substrate (tree bark) to which the epiphytic bryophytes attach. Therefore the information concerning the content of biogenous elements in the water leach from each of the tree barks as well as the total content of risk elements in tree barks are very important. The natural content of given elements in the tree barks can vary depending on the age of tree and they can vary between different species (biogenous elements). Secondarily the chemical composition of bark can vary due to the influence of air pollutants, which can get to the tree bark by dry or wet deposition [4].

2 Collection of data

Three regions were used for chemical composition of tree bark including Osoblažsko, Odersko and Vsetínsko. These regions can be considered to be not affected by the industry or influenced only a little bit in light of air pollution. The samples were collected from 16 species of trees (*Salix fragilis*, *Larix decidua*, *Tilia cordata*, *Tilia platyphyllos*, *Carpinus betulus*, *Betula pendula*, *Picea abies*, *Pinus sylvestris*, *Acer pseudoplatanus*, *Acer platanoides*, *Malus* sp, *Quercus robur*, *Fraxinus excelsior*, *Juglans regia*, *Fagus sylvatica*, *Aesculus hippocastanum*). Always four trees of different size and different age were chosen randomly from in advance selected tree species in each locality. The collection of data was very complicated in this stage because not always was the selected tree found in given locality or it often grew in distant and inapproachable places. The samples of tree bark were collected from all points of compass in height of 1-1,5 meter above the ground surface using a sharp knife. The size of collected samples from all points of compass was approximately 25 cm² in thickness to phloem. The tree was consequently treated with tree healing balsam. The tree bark samples were preserved in paper bags and transported to the laboratory for further analysis. The chemical composition of tree bark was determined by analysis of water leach, further the content of volatile organic substances (burnt at temperature 470°C) and by the determination of heavy metals in dry matter of tree bark. The values of following 12 quantities were stated for each sample: perimeter of trunk (m), pH, DOC (mg/kg), total phosphorus (mg/kg), total nitrogen (mg/kg), potassium (mg/kg), calcium (mg/kg), sodium (mg/kg), sulphate (mg/kg) and conductivity (μS/m).

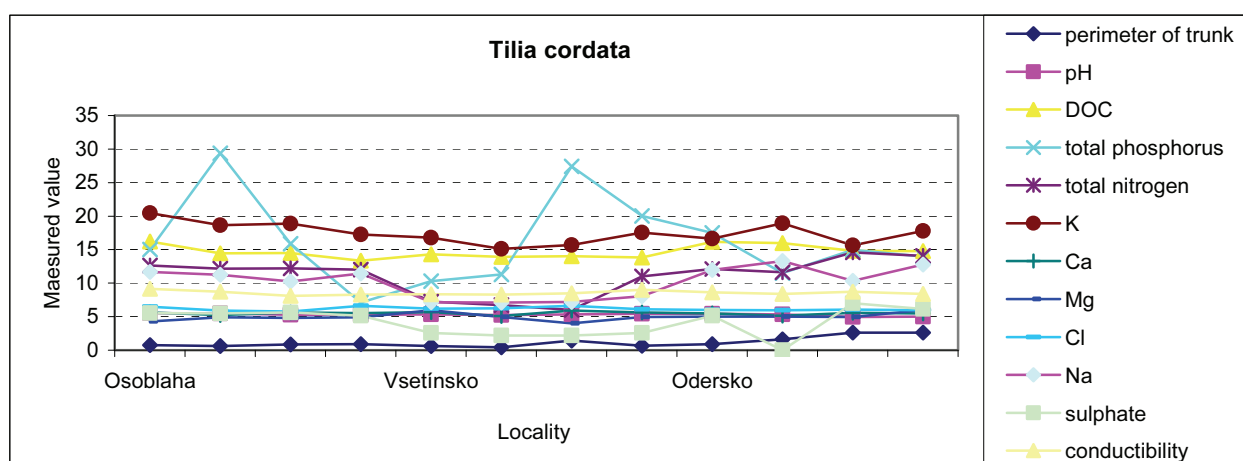
3 Data analysis

The submitter required only basic statistics information about the data file for the primary orientation including arithmetic mean, standard deviation and range of variation.

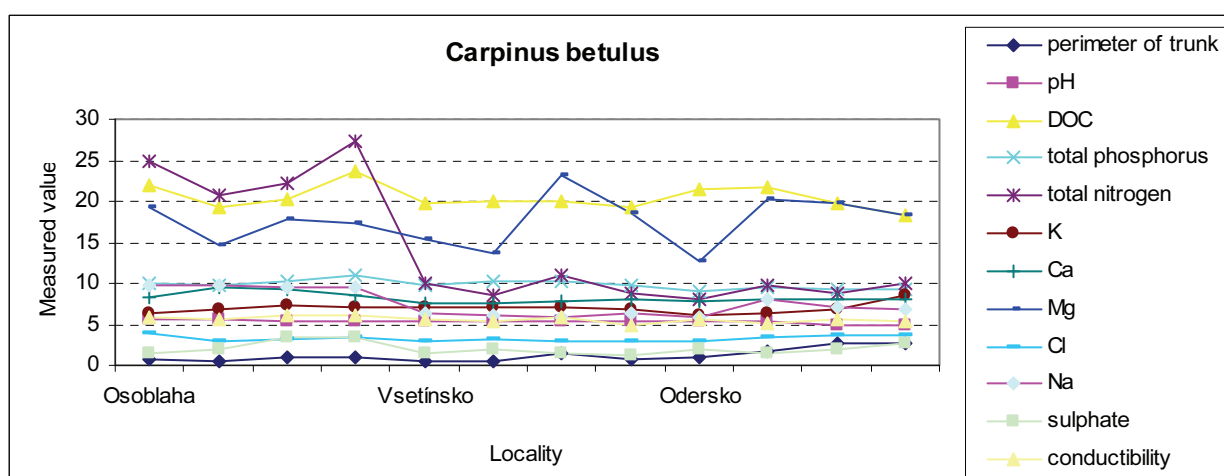
During the first data investigation, it was detected that it was not always possible to stick to the basic presumption and collect four samples in each locality. Therefore the species *Aesculus hippocastanum* was cut out from the next processing – in total only 3 samples from 2 places. Species *Juglans regia* and *Tilia platyphyllos* were collected in one locality only. Other tree species were found in only 2 observed localities (*Fagus sylvatica*, *Acer platanoides*, *Quercus robur*, *Fraxinus excelsior*).

We supposed that the values of all 12 observed quantities (see above) vary in dependence on tree species. This presumption was explicitly confirmed in all cases by the analysis of dispersion made on significance level $\alpha = 0.05$.

From the graphic illustration of individual tree species that shows the value distribution of investigated quantities (we show the graphs of species *Tilia cordata* – Picture No. 1 and *Carpinus betulus* – Picture No. 2 as the examples), the question emerged whether these quantities for individual tree species depend on the region, where the sample was collected.



Picture No. 1. Value illustration of investigated quantities for species *Tilia cordata*



Picture No. 2. Value illustration of investigated quantities for species *Carpinus betulus*

We used the single factor analysis of dispersion to find out, whether the choice of locality, where the tree species is present, is important for individual investigated quantities. The results for all tree species (with exception of species *Juglans regia* and *Tilia platyphyllos* – where the samples were taken from only 1 place of collection) and all observed quantities are shown in [2].

As we had the data about the trunk perimeter, where the samples were taken, another question aroused, whether the chemical composition of tree bark is affected by the age of tree. We admitted the assumption of direct correlation between the age of tree and trunk's perimeter. Then, we searched for the way how to take into account the age of tree in data format. With regard to the samples that were collected from the tree bark, we adjusted the data by dividing the measured and laboratory established data with the appropriate trunk's perimeter.

Tree species	pH	DOC	P	N	K	Ca	Mg	Cl	Na	sulphate	conductibility
Salix fragilis	P	P	Z	P	P	Z/P	P	P	P	Z	P
Larix decidua	P	P	P	P	P	P	P	P	P	Z	P
Tilia cordata	Z/P	P	P	P	Z/P	Z	P	Z/P	P	Z	Z/P
Carpinus betulus	P	P	P	Z	P	P	P	P	P/Z	P	P
Betula pendula	P	P	P	P	P	P	P	P	P	P	P
Picea abies	Z	Z	P	P	Z	Z	Z	Z/P	P	Z	Z
Pinus sylvestris	P	P	P	Z/P	P	Z	P	P	Z	P	P
Acer pseudoplatanus	P	P	P	P	P	P	P	P	P	P	P
Malus sp A	P	P	P	P	P	P	P	P	P	P	P
Fraxinus excelsior	P	P	P	P	P	P	P	P	P	P	P
Acer platanoides	P	P	P	P	P	P	P	P	P	P	P
Quercus robur	P	P	P	P	P	P	P	P	P	P	P
Fagus sylvatica	P	Z	P	P	P	P	P	P	P	P	P

Table No. 1. Results of analysis of dispersion for tree species

On significance level $\alpha = 0.05$ we verified whether the content of in this way modified quantities for individual tree species depend on the location where the samples were collected. In case of rejection of zero hypothesis (H_0 : investigated quantity doesn't depend on the location of sample collection, for the given species) on significance level $\alpha = 0.05$, the significance level $\alpha = 0.025$ was used. The results are shown for all tree species, where the planned 12 samples (4 from each locality) were collected, in Table No. 1, where

P means the acceptance of zero hypothesis on significance level $\alpha = 0.05$,

Z/P means the rejection of zero hypothesis on significance level $\alpha = 0.05$ and its acceptance on significance level $\alpha = 0.025$,

Z means the rejection of zero hypothesis on significance level $\alpha = 0.05$ and $\alpha = 0.025$.

On the basis of introduced analysis the required characteristics were calculated for the individual tree species, see [2].

4 Conclusion

The dependences of presence of biogenous elements in the tree bark on the location of sample collection are investigated in this subscription. On the basis of their conclusions, the required characteristics for individual quantities, tree species and regions, are established. The ascertained results are intended for interpretation of dependences of bryophyte distribution on the chemical characteristics of individual tree species in context of air pollution.

The next step is to compare the data from the region with relatively unpolluted air with the data from Ostrava, from the region with long term air pollution.

Acknowledgement

The paper was prepared thanks to support provided by the project MSM 61989100 – DeCOx.

References

- [1] DAVIES, L., BATES, J.W., BELL, J. N .B., JAMES, P. W., PURVIS, O.W.: *Diversity and sensitivity of epiphytes to oxides of nitrogen in London*. Environmental Pollution. 2007, vol.146, s. 299-310.
- [2] DOLEŽALOVÁ, J., VALOVÁ, M.: *Analýza chemického složení kůry stromů*. Sborník Konf. Moderní matematické metody v inženýrství, s. 26-30, Dolní Lomná 2009, ISBN 978-80-248-2118-4.
- [3] KUČERA, J.: Charakteristiky taxonomické skupiny, mechorosty [online]. [cit. 2007- 4- 5]. Dostupné z URL <<http://www.usbe.cas.cz/cervenakniha/>>.
- [4] POIKOLAINEN, J. MOSSES, epiphytes lichens and tree bark as biomonitors for air pollutants-specifically for heavy metals in regional surveys. University of Oulu, 2004. 64s. ISBN 951-42-7478-4.

Current address

Jarmila Doležalová (doc. RNDr. CSc.),

VŠB – TU Ostrava; Katedra matematiky a deskriptivní geometrie; 17. listopadu 15;

708 33 Ostrava – Poruba; 420 597324185,

e-mail: jarmila.dolezalova@vsb.cz

Marie Valová (Mgr. PhD.),

VŠB – TU Ostrava; Institut environmentálního inženýrství; 17. listopadu 15;

708 33 Ostrava – Poruba; 420 724711569,

e-mail: marie.mac@email.cz

SOLVING LOGISTICS PROBLEMS USING $M|G|\infty$ QUEUE SYSTEMS BUSY PERIOD

FERREIRA Manuel Alberto M., (P), FILIPE José António, (P),

Abstract. In the $M|G|\infty$ queuing systems customers arrive according to a Poisson process at rate λ . Each of them receives immediately after its arrival a service whose length is a positive random variable with distribution function $G(\cdot)$ and mean value α . An important parameter of the system is the traffic intensity $\rho = \lambda\alpha$. The service of a customer is independent of the services of the other customers and of the arrival process. The busy period of a queuing system begins when a customer arrives there, finding it empty, and ends when a customer leaves the system letting it empty. During the busy period there is always at least one customer in the system. Therefore in a queuing system there is a sequence of idle and busy periods. For these systems with infinite servers the busy period length distribution is difficult to derive, except for a few exceptions. But formulae that allow the calculation of some of the busy period length parameters for the $M|G|\infty$ queuing system are presented. These results can be applied in logistics (see, for instance, Ferreira [4,5] and Ferreira, Andrade and Filipe [9]). For instance, they can be applied to the failures which occur in the operation of an aircraft, shipping or trucking fleet. The customers are the failures. And their service time is the time that goes from the instant at which they occur till the one at which they are completely repaired. Here a busy period is a period in which there is at least one failure waiting for reparation or being repaired. The formulae referred allow the determination of measures of the system performance.

Key Words: $M|G|\infty$, Busy Period, Failures

Mathematics Subject Classification: Primary 60E05, 60G07; Secondary 60K25.

1 Introduction

In the $M|G|\infty$ queuing system (see, for example, Ferreira [2,6] and Ferreira and Andrade [7])

- The customers arrive according to a Poisson process at rate λ ,
- Each of them receives a service whose length is a positive random variable with distribution function $G(\cdot)$ and mean value α . So

$$\alpha = \int_0^{\infty} [1 - G(t)] dt \quad (1),$$

- There are infinite servers. So when a customer arrives it always finds a server available,
- The service of a customer is independent of the services of the other customers and of the arrival process.

An important parameter is the traffic intensity called ρ , being

$$\rho = \lambda \alpha \quad (2)$$

It is obvious that in an $M | G | \infty$ queuing system there are neither losses nor waiting. In fact there is no queuing in the formal sense of the word.

For these systems it is not so important to study the population process as for other systems with losses or waiting. Generally it is much more interesting the study of some other processes as, for instance, the busy period.

The busy period of a queuing system begins when a customer arrives there, finding it empty, and ends when a customer leaves the system letting it empty. During the busy period there is always at least one customer in the system.

Therefore in a queuing system there is a sequence of idle and busy periods.

It will be shown in the next section that these concepts can be applied in logistics, particularly to the failures that occur in the operation of a fleet of aircraft, of shipping or of trucking.

The results related to the busy period length of the $M | G | \infty$ queuing system, that is a random variable, allow the evaluation of performance measures of the fleet. In consequence it is possible to identify ways of improving the performance of the fleet.

The theory will be illustrated with a very simple and short numerical example.

2 Results and applications

Let us call B the $M | G | \infty$ queuing system busy period length (see Ferreira and Andrade [8]).

The mean value of B , whatever is $G(\cdot)$, is given by Takács [11]

$$E[B] = \frac{e^{\rho} - 1}{\lambda} \quad (3)$$

But $VAR [B]$, the variance of B , depends largely on the form of B . But Sathe [10] showed that

$$\lambda^{-2} \max \left[e^{2\rho} + e^{\rho} \rho^2 \gamma_s^2 - 2\rho e^{\rho} - 1; 0 \right] \leq VAR[B] \leq \lambda^{-2} \left(2e^{\rho} (\gamma_s^2 + 1) (e^{\rho} - 1 - \rho) - (e^{\rho} - 1)^2 \right) \quad (4),$$

where γ_s is the variation coefficient of $G(\cdot)$. And, after (4), the bounds to SD $[B]$ and the standard deviation of B can be very easily computed.

Being $R(t)$ the mean number of busy periods that begin in $[0, t]$ (being $t = 0$ the beginning of a busy period), see Ferreira [2],

$$e^{-\rho}(1 + \lambda t) \leq R(t) \leq 1 + \lambda t \quad (5)$$

Let us call N_B the mean number of the customers served during a busy period in the $M | G | \infty$ queuing systems. After Ferreira [3],

- If $G(\cdot)$ is exponential

$$N_B^M = e^\rho \quad (6),$$

- For any other distribution function

$$N_B \cong \frac{e^{\rho(\gamma_s^2+1)}(\rho(\gamma_s^2+1)+1) + \rho(\gamma_s^2+1) - 1}{2\rho(\gamma_s^2+1)} \quad (7)$$

These results can be applied to logistics. They are applied, for instance, to the failures that occur in the operation of a fleet of aircraft, of shipping or of trucking. The customers are the failures. And its service time is the time that goes from the instant at which they occur till the one at which they are completely repaired. For examples of applications of this kind see, for instance, Carrillo [1]. So

- A busy period is a period, in which there is at least one failure waiting for a reparation or being repaired,
- An idle period is a period in which there are no failures present.

Here some simple expressions that allow computing the mean and bounds to the variance of the busy period were given. And also simple bounds to the mean number of busy periods that begin in a certain length of time. And finally, expressions to the mean number of failures that occur in a busy period were presented.

These formulae are very simple and of evident application. They only require the knowledge of α , λ , ρ and γ_s that are very easy to compute. The only problem is to test the hypothesis of that the failures occur according to a Poisson process.

Only to conclude note that, calling $I(t)$ the idle period of the $M | G | \infty$ queuing system distribution function,

$$I(t) = 1 - e^{-\lambda t} \quad (8),$$

as it happens with any queue with arrival Poisson process. In this application it gives the probability of that the length of time with no failures is lesser or equal to t .

3 Examples

Suppose a fleet where the failures occur at a rate of 20 per year. So $\lambda = 20/\text{year}$. Suppose too that the mean time to repair a failure is 4 days ($\alpha = 4 \text{ day} = (4/365) \text{ year}$). In consequence $\rho \cong 0.22$.

Possibly ρ maybe decreased to 0.11. It is enough to make $\lambda = 10/\text{year}$, for instance buying more vehicles and decreasing, in consequence, the use intensity of each one. Or making $\alpha = 2 \text{ day}$. For instance increasing the teams that repair the failures.

On other side, if nothing is changed, things can get worse and maybe ρ can increase to 0.44.

If it is supposed that the repair services times are exponential (a very frequent supposition for this kind of services), $\gamma_s = 1$, and after (3), (4), (5) and (6), with $t = 1$ year, being $SD [B] = \sqrt{Var[B]}$,

Table 1

ρ	$E[B]$	SD [B] (Lower Bound)	SD [B] (Upper Bound)	R (1) (Lower Bound)	R (1) (Upper Bound)	N_B^M
0.11	2.12 day	2.16 day	2.2 day	18	21	1.12
0.22	4.5 day	4.65 day	4.82 day	16	21	1.25
0.44	10 day	10.72 day	11.46 day	13	21	1.60

And it is possible to conclude that when ρ increases, less busy periods in one year occur, with more failures in each one, of course in mean values.

The mean of the length of the busy period and its dispersion increase with ρ too.

If it is supposed now that the repair service times are constant ($D = \text{deterministic}$), $\gamma_s = 0$, and after (3), (4) (in this case the lower bound is equal to the upper bound and so the real value of $VAR[B]$ is got), (5) and (7), with $t = 1$ year

Table 2

ρ	$E[B]$	SD [B]	R (1) (Lower Bound)	R (1) (Upper Bound)	N_B^D
0.11	2.12 day	0.41 day	18	21	1.59
0.22	4.5 day	1.22 day	16	21	1.68
0.44	10 day	3.85 day	13	21	1.90

$E[B]$ and the $R(1)$ bounds are the same that in the former case, evidently. The behavior of the parameters with the increase of ρ is similar to the one of the exponential situation. But now the busy period length dispersion is much lesser and the mean value of failures in each busy period is greater.

4 Conclusion

Of course, in the operation of a fleet, one is interested in big idle periods and in little busy periods. And if these busy periods occur it is good that they are as rare as possible, with a short number of failures.

Knowing α , λ , ρ and γ_s the manager of the fleet can evaluate the conditions of the operation, namely:

- The mean length of a period with failures,
- The length dispersion of a period with failures,
- The mean number of periods with failures that will occur in a certain length if time,
- The mean number of failures that occur in a period with failures.

As the expressions depend only on a few parameters and are very simple they show very simple ways to improve the operation, although they may be hard to implement.

References

- [1] CARRILLO, M. J. (1991), “Extensions of Palm’s Theorem: A Review”, *Management Science*, Vol. 37, n.º 6, p.739 – 744.
- [2] FERREIRA, M. A. M. (1995), “Comportamento transeunte e período de ocupação de sistemas de filas de espera sem espera”, PhD Thesis discussed in ISCTE, Supervisor: Prof. Augusto A. Albuquerque.
- [3] FERREIRA, M. A. M. (2001), “Mean number of the Customers Served During a Busy Period in the $M|G|\infty$ Queueing System”, *Statistical Review*, Vol. 3, INE.
- [4] FERREIRA, M. A. M. (2002), “Busy Period of Queueing Systems with Infinite Servers and Logistics”, *IMRL 2002 International Meeting for Research in Logistics. Proceedings (Volume-1)*, p. 270-274.
- [5] FERREIRA, M. A. M. (2003), “M/G/ ∞ Queue Busy Period and Logistics”, *APLIMAT 2003. Proceedings (Part I)*, p. 329-332.
- [6] FERREIRA, M. A. M. (2009), “Statistical Queuing Theory”, entry in Lovric, M., *International Encyclopedia of Statistical Science*, Springer-Verlag. Berlin Heidelberg. Forthcoming.
- [7] FERREIRA, M. A. M. and ANDRADE, M. (2009), $M|G|\infty$ Queue System for a particular Collection of Service Time Distributions. *African Journal of Mathematics and Computer Science Research*. Vol. 2, n.º7. p.138-141.
- [8] FERREIRA, M. A. M. and ANDRADE, M. (2009), “The Ties between the $M|G|\infty$ Queue System Transient Behaviour and the Busy Period”, *International Journal of Academic Research*, Vol. 1, n.º 1, p. 84-92.
- [9] FERREIRA, M. A. M., ANDRADE, M. and FILIPE, J. A. (2009), “Networks of queues with infinite servers in each node applied to the management of a two echelons repair system”, *China-USA Business Review*, Vol. 8, n.º 8, p. 39-45.
- [10] SATHE, Y. S. (1985), “Improved Bounds for the variance of the busy period of the $M|G|\infty$ queue”, *A.A.P.*, p.913 – 914.
- [11] TAKÁCS, L. (1962), “An Introduction to queueing theory”, Oxford University Press, New York.

Current address

Manuel Alberto M. Ferreira, Professor Catedrático

ISCTE – Lisbon University Institute

UNIDE – Unidade de Investigação e Desenvolvimento Empresarial

Av. Forças Armadas 1649-026 Lisboa, Portugal

tel. +351 217 903 000

e-mail: manuel.ferreira@iscte.pt

José António Filipe, Professor Auxiliar

ISCTE – Lisbon University Institute

UNIDE – Unidade de Investigação e Desenvolvimento Empresarial
Av. Forças Armadas 1649-026 Lisboa, Portugal
tel.+351 217 903 000
e-mail: jose.filipe@iscte.pt

M|G| ∞ SYSTEM TRANSIENT BEHAVIOR WITH TIME ORIGIN AT THE BEGINNING OF A BUSY PERIOD MEAN AND VARIANCE

FERREIRA Manuel Alberto M., (P), ANDRADE Marina, (P)

Abstract. The $M|G|\infty$ queue system transient probabilities, with time origin at the beginning of a busy period, are determined. It is highlighted the obtained distribution mean and variance study as time functions. In this study it is determinant the hazard rate function service time. The results obtained are applied in modeling disease and unemployment situations.

Key words and phrases. $M|G|\infty$, hazard rate function, disease, unemployment.

Mathematics Subject Classification. 60K35.

1 Introduction

In the $M|G|\infty$ queue system the customers arrive according to a Poisson process at rate λ , receive a service which time is a positive random variable with distribution function $G(\cdot)$ and mean α and, when they arrive, they find immediately an available server. Each customer service is independent from the other customers' services and from the arrivals process. The traffic intensity is $\rho = \lambda\alpha$.

$N(t)$ is the number of occupied servers (or the number of customers being served) at instant t , in a $M|G|\infty$ system.

From (Takács, 1962), as $p_{0n}(t) = P[N(t) = n | N(0) = 0]$, $n = 0, 1, 2, \dots$,

$$p_{0n}(t) = \frac{\left(\lambda \int_0^t [1 - G(v)] dv\right)^n}{n!} e^{-\lambda \int_0^t [1 - G(v)] dv}, \quad n = 0, 1, 2, \dots \quad (1)$$

So, the transient distribution, when the system is initially empty, is Poisson with mean $\lambda \int_0^t [1 - G(v)] dv$.

The stationary distribution is the limit distribution:

$$\lim_{t \rightarrow \infty} p_{0n}(t) = \frac{\rho^n}{n!} e^{-\rho}, \quad n = 1, 2, \dots \quad (2)$$

This queue system, as any other, has a sequence of busy periods and idle periods. A busy period begins when a customer arrives at the system finding it empty.

Be $p_{1'n} = P[N(t) = n | N(0) = 1']$, $n = 0, 1, 2, \dots$, meaning $N(0) = 1'$ that the time origin is an instant at which a customer arrives at the system jumping the number of customers from 0 to 1.

That is: a busy period begins.

At $t \geq 0$ possibly:

- The customer that arrived at the initial instant either abandoned the system, with probability $G(t)$, or goes on being served, with probability $1 - G(t)$;
- The other servers, that were unoccupied at the time origin, either go on unoccupied or occupied with 1, 2, ... customers, being the probabilities $p_{0n}(t)$, $n = 0, 1, 2, \dots$

Both subsystems, the one of the initial customer and the one of the servers initially unoccupied, are independent and so

$$\begin{aligned} p_{1'0}(t) &= p_{00}(t) G(t) \\ p_{1'n}(t) &= p_{0n}(t) G(t) + p_{0n-1}(t) (1 - G(t)), \quad n = 1, 2, \dots \end{aligned} \quad (3)$$

It is easy to see that

$$\lim_{t \rightarrow \infty} p_{1'n}(t) = \frac{\rho^n}{n!} e^{-\rho}, \quad n = 0, 1, 2, \dots \quad (4)$$

Denoting $\mu(1', t)$ and $\mu(0, t)$ the distributions given by (3) and (1) mean values, respectively,

$$\begin{aligned} \mu(1', t) &= \sum_{n=1}^{\infty} n p_{1'n}(t) = \sum_{n=1}^{\infty} n G(t) p_{00}(t) + \sum_{n=1}^{\infty} n p_{0n-1}(t) (1 - G(t)) = \\ &= G(t) \mu(0, t) + (1 - G(t)) \sum_{j=0}^{\infty} (j+1) p_{0j}(t) = \mu(0, t) + (1 - G(t)), \end{aligned}$$

that is

$$\mu(1', t) = 1 - G(t) + \lambda \int_0^t [1 - G(v)] dv \quad (5)$$

As

$$\begin{aligned}\sum_{n=0}^{\infty} n^2 p_{1'n}(t) &= G(t) \sum_{n=1}^{\infty} n^2 p_{0n}(t) + (1 - G(t)) \sum_{n=1}^{\infty} n^2 p_{0n-1}(t) = \\ &= G(t) (\mu^2(0, t) + \mu(0, t)) + (1 - G(t)) (\mu^2(0, t) + 3\mu(0, t) + 1) = \\ &= \mu^2(0, t) + (3 - 2G(t)) \mu(0, t) + 1 - G(t),\end{aligned}$$

denoting $V(1', t)$ the variance associated to the distribution defined by (3), it is obtained

$$V(1', t) = \mu(0, t) + G(t) - G^2(t). \quad (6)$$

The main target is to study $\mu(1', t)$ and $V(1', t)$ as time functions. It will be seen that, in its behavior as time functions, plays an important role the hazard rate function service time given by, see for instance (Ross, 1983),

$$h(t) = \frac{g(t)}{1 - G(t)} \quad (7)$$

and $g(\cdot)$ is the density associated to $G(\cdot)$.

2 Application in disease and unemployment situations

$M|G|_{\infty}$ systems have great applicability in the modeling of real problems. See, for instance, the works of (Carrillo, 1991), (Ferreira, 1988), (Ferreira, Andrade and Filipe, 2009) (Hershey, Weiss and Morris, 1981) and (Kelly, 1979). The ones presented in this paper are very interesting and the results that will be shown in the sequence are particularly adequate to its study.

2.1 Disease

In this case the customers are the people that have a certain disease. They arrive at the system when they fall sick and their service time is the time during which they are sick. The time the first one falls sick, may be the beginning of an epidemic, is the beginning of a busy period. An idle period is a period of disease absence.

The service hazard rate function is the rate at which they get cured.

2.2 Unemployment

Now the customers are the unemployed in a certain activity. They arrive at the system when they loose their jobs and their service time is the time during which they are unoccupied.

An idle period is a full employment period. A busy period begins with the first worker loosing his job.

The hazard rate function is the rate at which the unemployed workers turn employees.

In both cases (3) is applicable. It must be checked if the people fall sick or loose their jobs according to a Poisson process. The failing of this hypothesis is more expectable in the unemployment situation. In some of this situations may be it is more adequate to consider a mechanism of batch arrivals.

The beginning of the epidemics or of the unemployment periods can be determined today with a great precision.

The results that will be presented can help to forecast the evolution of the situations.

Finally it is necessary to adjust the time distributions adequate to the disease and unemployment periods. In this last case the situation may not be the same for the various activities.

3 $\mu(1', t)$ study as time function

Lemma 3.1 *If $G(t) < 1$, $t > 0$ continuous and differentiable and*

$$h(t) \leq \lambda, t > 0 \quad (8)$$

$\mu(1', t)$ is non- decreasing.

Dem.:

It is enough to note, according to (5), that

$$\frac{d}{dt} \mu(1', t) = (1 - G(t)) (\lambda - h(t)) .$$

Obs.:

- If the rate at which the services end is lesser or equal than the customers arrival rate $\mu(1', t)$ is non- decreasing.
- **Disease:** If the rate at which people get cured is lesser or equal than the rate at which they fall sick, the mean number of sick people is a non- decreasing time function.
- **Unemployment:** If the rate at which the workers loose their jobs is lesser than the rate at which they turn employees, the mean number of unemployed people is a non- decreasing time function.
- For the $M|M|\infty$ system (8) is equivalent to

$$\rho \geq 1 \quad (9)$$

- $\lim_{t \rightarrow \infty} \mu(1', t) = \rho$.
- **Disease:** If an epidemic lasts a very long time, the mean number of sick people will be closer and closer from the traffic intensity.
- **Unemployment:** If an unemployment period lasts a very long time, the mean number of unemployed people will be closer and closer from the traffic intensity.

Making $h(t) - \lambda = \beta(t)$, $\beta(\cdot)$ it is obtained

$$G(t) = 1 - (1 - G(0)) e^{-\lambda t - \int_0^t \beta(u) du}, t \geq 0, \frac{\int_0^t \beta(u) du}{t} \geq -\lambda \quad (10)$$

So

Lemma 3.2 If $\beta = 0$

$$G(t) = 1 - (1 - G(0)) e^{-\lambda t}, \quad t \geq 0 \quad (11)$$

and $\mu(1', t) = 1 - G(0) = \rho, \quad t \geq 0$.

Obs.:

- **Disease:** If the time that a patient is sick is a random variable, with a distribution function given by (11) the mean number of sick people is always equal to the traffic intensity.
- **Unemployment:** If the time of unemployment is a random variable, with a distribution function given by (11) the mean number of unemployed people is always equal to the traffic intensity.

For some particular service time distributions:

- Deterministic with value α

$$\mu(1', t) = \begin{cases} 1 + \lambda t, & t < \alpha \\ \rho, & t \geq \alpha \end{cases} \quad (12)$$

- Exponential

$$\mu(1', t) = \rho + (1 - \rho) e^{-\frac{t}{\alpha}}. \quad (13)$$

-

$$\begin{aligned} G(t) &= 1 - \frac{(1 - e^{-\rho})(\lambda + \beta)}{\lambda e^{-\rho}(e^{(\lambda + \beta)t} - 1) + \lambda}, t \geq 0, -\lambda \leq \beta \leq \frac{\lambda}{e^\rho - 1} \\ \mu(1', t) &= \frac{(1 - e^{-\rho})(\lambda + \beta)}{\lambda e^{-\rho}(e^{(\lambda + \beta)t} - 1) + \lambda} + \rho - \log(1 + (e^\rho - 1)e^{-(\lambda + \beta)t}) \end{aligned} \quad (14)$$

For this collection of service time distributions the busy period (and so also the time that an **epidemic** or an **unemployment period** lasts) is exponentially distributed with an atom at the origin

$$B^\beta(t) = 1 - \frac{\lambda + \beta}{\lambda} (1 - e^{-\rho}) e^{-e^{-\rho}(\lambda + \beta)t}, \quad t \geq 0, \quad -\lambda \leq \beta \leq \frac{\lambda}{e^\rho - 1}. \quad (15)$$

4 $V(1', t)$ study as time function

Lemma 4.1 *If $G(t) < 1$, $t > 0$, continuous and differentiable and*

$$h(t) \geq -\frac{\lambda}{1 - 2G(t)} \quad (16)$$

$V(1', t)$ is non-decreasing.

Dem.:

It is enough to note, according to (6), that

$$\begin{aligned} \frac{d}{dt} V(1', t) &= \lambda(1 - G(t)) + g(t) - 2G(t)g(t) = \lambda(1 - G(t)) + g(t)(1 - 2G(t)) = \\ &= (1 - G(t))(h(t)(1 - 2G(t)) + \lambda). \end{aligned}$$

Obs.:

- Obviously $1 - 2G(t) < 0 \Leftrightarrow G(t) > \frac{1}{2}$, $t > 0$.
- **Disease:** If the rate at which people get cured, the rate at which they fall sick and the sickness duration distribution function hold (16) the variance of the number of sick people is a non-decreasing time function.
- **Unemployment:** If the rate at which the workers loose their jobs, the rate at which they turn employees and the unemployment duration distribution function hold (16) the variance of the number of sick people is a non-decreasing time function.
- $\lim_{t \rightarrow \infty} V(1', t) = \rho$.
- **Disease:** If an epidemic lasts a very long time, the variance of the number of sick people will be closer and closer from the traffic intensity.
- **Unemployment:** If an unemployment period lasts a very long time, the variance of the number of unemployed people will be closer and closer from the traffic intensity.

- **Disease:** If an epidemic lasts a very long time the number of sic people is distributed according to a Poisson distribution with mean ρ , see (4).
- **Unemployment:** If an unemployment period lasts a very long time the mean number of unemployed people is Poisson distributed with mean ρ , see (4).

Making $h(t) + \frac{\lambda}{1-2G(t)} = 0$ the following proposition holds:

Lemma 4.2 *If $G(\cdot)$ is implicitly defined as*

$$\frac{1-G(t)}{1-G(0)} e^{2(G(t)-G(0))} = e^{-\lambda t}, t \geq 0 \quad (17)$$

$$V(1', t) = \rho, t \geq 0.$$

Obs.:

- The density associated to (17) is

$$g(t) = -\frac{\lambda e^{-\lambda t} (1-G(0))}{(1-2G(t)) e^{2(G(t)-G(0))}} \quad (18)$$

- After (18), denoting S the associated random variable, it is easy to see that, with $G(0) > \frac{1}{2}$,

$$\frac{(1-G(0)) n! e^{-2(1-G(0))}}{\lambda^n} \leq E[S^n] \leq \frac{(1-G(0)) n!}{(2G(0)-1) \lambda^n}, n = 1, 2, \dots \quad (19)$$

For some particular service time distributions:

- Deterministic with value α

$$V(1', t) = \begin{cases} \lambda t, t < \alpha \\ \rho, t \geq \alpha \end{cases} \quad (20)$$

- Exponential

$$V(1', t) = \rho \left(1 - e^{-\frac{t}{\alpha}} \right) + e^{-\frac{t}{\alpha}} + e^{-\frac{2t}{\alpha}}. \quad (21)$$

-

$$G(t) = 1 - \frac{(1-e^{-\rho})(\lambda+\beta)}{\lambda e^{-\rho}(e^{(\lambda+\beta)t}-1)+\lambda}, \quad t \geq 0, \quad -\lambda \leq \beta \leq \frac{\lambda}{e^\rho-1}$$

$$V(1', t) = \rho - \log(1 + (e^\rho - 1)e^{-(\lambda+\beta)t}) + \frac{(1-e^{-\rho})(\lambda+\beta)}{\lambda e^{-\rho}(e^{(\lambda+\beta)t}-1)+\lambda} +$$

$$+ \left(\frac{(1-e^{-\rho})(\lambda+\beta)}{\lambda e^{-\rho}(e^{(\lambda+\beta)t}-1)+\lambda} \right)^2. \quad (22)$$

5 Conclusions

With very simple probabilistic reasoning, the $M | G | \infty$ transient probabilities, being the time origin the beginning of a busy period instant, were determined. It is enough to condition to the service lasting of the first costumers.

It was possible to study $\mu(1', t)$ and $V(1', t)$, as time functions, playing here an important role the service time hazard rate function.

This model may be applied in modeling real situations being the difficulties the usual ones when theoretical models are applied to real situations.

Finally note that, in the disease application, the model is not applicable to contagious epidemics. In this situation it would be more realistic to consider arrival rates not constant. But, with this kind of rates, it is not possible to have results as interesting and useful as those presented in this work.

References

- [1] [1] CARRILLO, M.J. (1991), “*Extensions of Palm’s Theorem: A Review*”. Management Science. Vol. 37. N ° 6. 739-744.
- [2] FERREIRA, M.A.M. (1988), “*Redes de Filas de Espera*”. Dissertação de Mestrado apresentada no IST.
- [3] FERREIRA, M.A.M. (1998), “*Aplicação da Equação de Ricatti ao Estudo do Período de Ocupação do Sistema $M | G | \infty$* ”. Revista de Estatística.Vol. 1. 1 ° Quadrimestre. INE 23-28.
- [4] FERREIRA, M.A.M. (2010), “*Statistical Queuing Theory*” in Lovriv, Miodrag (Ed.), “*International Encyclopedia of Statistical Science*”. Springer. ISBN: 978-3-642-04897-5. Forthcoming.
- [5] FERREIRA, M. A. M. and ANDRADE, M. (2009), “ *$M | G | \infty$ Queue System Parameters for a Particular Collection of Service Time Distributions*. AJMCSR-African Journal of Mathematics and Computer Science Research. 2(7) 138-141. ISSN: 2006-9731.
- [6] FERREIRA, M. A. M. and ANDRADE, M. (2009), “*The Ties between the $M—G—\infty$ Queue System Transient Behavior and the Busy Period*”. International Journal of Academic Research 1(1) 84-92. ISSN: 2075-4124.
- [7] Ferreira, M. A. M., Andrade, M. and Filipe, J. A. (2008), “*The Ricatti Equation in the $M | G | \infty$ System Busy Cycle Study*”. Journal of Mathematics, Statistics and Allied Fields 2(1). ISSN: 1556-6757.
- [8] FERREIRA, M. A. M., ANDRADE, M. and FILIPE, J. A. (2009), “*Networks of Queues with Infinite Servers in Each Node Applied to the Management of a Two Echelons Repair System*”. China-USA Business Review. 8 (8) 39-45 and 62. ISSN: 1537-1514.
- [9] Hershey, J. C., WEISS, E. N. and MORRIS, A. C. (1981), “*A Stochastic Service Network Model with Application to Hospital Facilities*”. Operations Research 29 1-22.
- [10] KELLY, F.P. (1979), “*Reversibility and Stochastic Networks*”. New York. John Wiley & Sons.

- [11] ROSS, S. (1983), *“Stochastic Processes”*. Wiley. New York.
- [12] TAKÁCS, L. (1992), *“An Introduction to Queueing Theory”*. Oxford University Press. New York.

Current address

Manuel Alberto Martins Ferreira, Professor Catedrático

ISCTE – Lisbon University Institute

UNIDE – Unidade de Investigação e Desenvolvimento Empresarial

Av. das Forças Armadas 1649-026 Lisboa (Lisboa, Portugal)

Tel. +351 217 903 000

e-mail: manuel.ferreira@iscte.pt

Marina Alexandra Pedro Andrade, Professor Auxiliar

ISCTE – Lisbon University Institute

UNIDE – Unidade de Investigação e Desenvolvimento Empresarial

Av. das Forças Armadas 1649-026 Lisboa (Lisboa, Portugal)

Tel. +351 217 903 000

e-mail: marina.andrade@iscte.pt

SIMULTANEOUS TOLERANCE INTERVALS IN A LINEAR REGRESSION

CHVOSTEKOVÁ Martina, (SK)

Abstract. The simultaneous tolerance intervals are important for many measurement procedures. The most common application for simultaneous tolerance intervals is a multiple-use calibration problem; see e.g. Mee et. al. (Technometrics 1991). In this paper we present a brief overview of the methods for constructing simultaneous tolerance intervals in a linear regression with normal errors. In particular, we describe the Lieberman-Miller method, the Wilson method, the modified Wilson method, the Limam-Thomas method, and the Mee-Eberhardt-Reeve method.

Key words and phrases. predictor, linear regression model, simultaneous tolerance intervals, tolerance factor, confidence-set approach

Mathematics Subject Classification. Primary 62F25, 62J02.

1 Introduction

Statistical calibration is the process whereby the scale of a measuring instrument is determined or adjusted on the basis of a calibration experiment which consists of calibration data (or training set), i.e. $(x_i, Y_i), i = 1, \dots, n$, where x_i 's have been determined by an extremely accurate standard method (they are treated as known constant, i.e. we consider a controlled calibration) and Y_i is outcome from a measurement at a position x_i . We can suppose a linear relationship between a predictor and a response variable, for example. Based on a random sample $\mathbf{Y} = (Y_1, \dots, Y_n)$ we require confidence interval estimate for K future predictors corresponding to the future observed response variables, Y_{n+1}, \dots, Y_{n+K} . If K a natural number is unknown and possibly arbitrary large, this problem can be solved by inverting two-sided simultaneous tolerance intervals, see [6].

A tolerance interval is an interval based on a random sample, that is expected to capture a certain proportion or more of the population, with given confidence level. Unlike a confidence interval, which is used to bound an unknown scalar population parameter (population mean, standard deviation, etc.), tolerance interval provides information on the entire population. We can make statements not only about the expected value, but we can also locate in which range individual will be found given predictor value. In case of a calibration problem we want to be valid over the entire range of possible predictor values, so we must consider simultaneous intervals. It is not relevant to use a prediction interval, since it is only for a single future observation.

The simultaneous tolerance intervals have been recognized and considered in various setting by many authors. These intervals are constructed such that, with confidence $1 - \alpha$, we can claim that at least a specified proportion, $1 - \gamma$, of the population is contained in the tolerance interval, for all possible values of the predictor variables. All known simultaneous tolerance intervals in a regression are conservative.

In section 2 we will define underlying model. Here we consider methods for computing simultaneous tolerance intervals for a linear regression with normal errors. The general form of a tolerance interval depends basically on the tolerance factor, which further depends on the given proportion γ , the confidence level $1 - \alpha$, the predictor variable, and the distribution of random sample. In section 3 we describe several methods for deriving and computing a tolerance factor. In section 4 a possible method for computing the tolerance factor is mentioned.

2 Simultaneous tolerance intervals in a linear regression

Throughout this paper we will assume the linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \sigma\mathbf{Z} \quad (1)$$

with normally distributed errors, where \mathbf{Y} represents an n -dimensional random vector of response variables, \mathbf{X} is the $n \times q$ matrix of non-stochastic explanatory variables (for simplicity, here we assume that \mathbf{X} is a full-rank matrix), β is a q -dimensional vector of regression parameters, \mathbf{Z} is an n -dimensional vector of standard normal errors, i.e. $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$, and σ is the error standard deviation, $\sigma > 0$. Note that a simple linear regression is a special case of (1).

Let $\hat{\beta}$ denote the least squares estimator of β and S^2 denote the residual mean square under the model (1), then

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad \text{and} \quad S^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})}{n - q}. \quad (2)$$

Note that $\hat{\beta} \sim N_q(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ and $(n - q)S^2/\sigma^2 \sim \chi_{n-q}^2$, where χ_{n-q}^2 denotes a central chi-square random variable with $n - q$ degrees of freedom. Random variables $\hat{\beta}$ and S^2 are independent.

Now let $Y(\mathbf{x})$ denote a future observation of a response at the predictor \mathbf{x} , i.e. $Y(\mathbf{x}) = \mathbf{x}^T\beta + \sigma Z \sim N(0, \sigma^2)$, where $Y(\mathbf{x})$ is assumed to be independent of \mathbf{Y} in (1). There, we will

consider the two-sided tolerance interval for $Y(\mathbf{x})$ in the following (general) form

$$\left\langle \mathbf{x}^T \hat{\beta} - \lambda(\gamma, \mathbf{x})S, \mathbf{x}^T \hat{\beta} + \lambda(\gamma, \mathbf{x})S \right\rangle, \quad (3)$$

where $\lambda(\gamma, \mathbf{x})$ is a tolerance factor that depends on the given proportion γ , the predictor \mathbf{x} , the confidence level $1 - \alpha$, and the distribution of \mathbf{Y} . The estimates of β and σ are computed from the observation \mathbf{y} using (2).

The simultaneous tolerance intervals are constructed using the vector of observations (1), so that with the confidence level $1 - \alpha$, at least a proportion γ of the $Y(\mathbf{x})$ -distribution is to be contained in the corresponding tolerance interval, simultaneously for all possible values of predictor variables \mathbf{x} . Let $C(\mathbf{x}^T \hat{\beta}, S)$ denote the content for the tolerance interval (3), given $\hat{\beta}$ and S , then

$$C(\mathbf{x}^T \hat{\beta}, S) = P_{Y(\mathbf{x})}(\mathbf{x}^T \hat{\beta} - \lambda(\gamma, \mathbf{x})S \leq Y(\mathbf{x}) \leq \mathbf{x}^T \hat{\beta} + \lambda(\gamma, \mathbf{x})S | \hat{\beta}, S). \quad (4)$$

The content (4) of the tolerance interval (3) can be expressed in terms of the pivotal quantities

$$\mathbf{b} = \frac{(\hat{\beta} - \beta)}{\sigma} \sim N(0, (\mathbf{X}^T \mathbf{X})^{-1}), \quad u = \frac{S}{\sigma}, \quad (n - q)u^2 \sim \chi_{n-q}^2, \quad (5)$$

where \mathbf{b} and u are independently distributed. Now, we can rewrite (4) using the cumulative distribution function of the standard normal distribution

$$C(\mathbf{x}^T \mathbf{b}, \lambda(\gamma, \mathbf{x})u) = \Phi(\mathbf{x}^T \mathbf{b} + \lambda(\gamma, \mathbf{x})u) - \Phi(\mathbf{x}^T \mathbf{b} - \lambda(\gamma, \mathbf{x})u), \quad (6)$$

which is an even and monotonically decreasing function of $\mathbf{x}^T \mathbf{b}$. To obtain the actual limits of the tolerance interval, it is necessary to compute the value of the tolerance factor. Tolerance factor $\lambda(\gamma, \mathbf{x}) = \lambda$ should satisfy the condition

$$P_{\mathbf{b},u}(C(\mathbf{x}^T \mathbf{b}, \lambda(\gamma, \mathbf{x})u) \geq \gamma \quad \forall \mathbf{x} \in \mathcal{R}^{q \times 1}) = 1 - \alpha. \quad (7)$$

The interpretation of this equation is that $(1 - \alpha)100\%$ of the tolerance intervals estimated from different samples will contain the proportion γ of the proportion of $Y(\mathbf{x})$ for any \mathbf{x} . Denote the set G of \mathbf{b}, u satisfying the inequality in (7); it is called a $(1 - \alpha)$ -pivotal set. A confidence set for both true β, σ can be described in terms of the $(1 - \alpha)$ -pivotal set G

$$\{(\beta, \sigma) = (\hat{\beta} - \mathbf{b}\hat{\sigma}/u, \hat{\sigma}/u) : (\mathbf{b}, u) \in G\}. \quad (8)$$

Expression (7) can be rewritten as

$$P_{\mathbf{b},u}(\min_{\mathbf{x}} C(\mathbf{x}^T \mathbf{b}, \lambda(\gamma, \mathbf{x})u) \geq \gamma) = 1 - \alpha, \quad (9)$$

in which form it is used for finding $\lambda(\gamma, \mathbf{x})$.

3 Methods for computing tolerance factors

Here we will present a brief overview of five methods for computing tolerance factors of the simultaneous tolerance intervals in a linear regression. In particular, we will describe the Lieberman-Miller method, [4], the Wilson method, [9], the Limam-Thomas method, [5], the modified Wilson method, [5], the Mee-Eberhardt-Reeve method, [6].

3.1 The Lieberman-Miller method

Wallis in [8] was the first who considered the problem of finding a tolerance interval in a linear regression model. He derived a formula for the tolerance factor for the non-simultaneous tolerance interval (i.e. an interval for a fixed value \mathbf{x}), using approximation of the content function

$$C(\mathbf{x}^T \mathbf{b}, \lambda u) \approx C(\delta(\mathbf{x}), \lambda(\gamma, \mathbf{x})u), \quad (10)$$

where $\delta(\mathbf{x}) = \sqrt{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}$ is the standard error of $\mathbf{x}^T \mathbf{b}$. Lieberman and Miller [4] extended the Wallis method for the simultaneous tolerance intervals (i.e. to the case where predictor is allowed to vary). Naturally, it was proposed to formulate tolerance factor in the form

$$\lambda(\gamma, \mathbf{x}) = \lambda^* \delta(\mathbf{x}). \quad (11)$$

Unknown constant λ^* is sought so that it is satisfied

$$P_u(\min_{\mathbf{x}} C(\delta(\mathbf{x}), \lambda^* \delta(\mathbf{x})u) \geq \gamma) = 1 - \alpha. \quad (12)$$

The function $\min_{\mathbf{x}} C(\delta(\mathbf{x}), h\delta(\mathbf{x}))$ is monotonically non-decreasing function of h , there exists a constant h_0 satisfying equation

$$\min_{\mathbf{x}} C(\delta(\mathbf{x}), h_0 \delta(\mathbf{x})) = \gamma. \quad (13)$$

Thus the function $\min_{\mathbf{x}} C(\delta(\mathbf{x}), \lambda^* \delta(\mathbf{x})u) > \gamma$ for $\lambda^* u > h_0$, and constant λ^* is chosen so that

$$P(u > h_0 / \lambda^*) = 1 - \alpha. \quad (14)$$

Hence

$$\lambda^* = h_0 \sqrt{(n - q) / \chi_{n-q}^2(\alpha)}, \quad (15)$$

where h_0 is computed numerically as was described above.

3.2 The Wilson method

Several authors considered the confidence-set approach for constructing the simultaneous tolerance intervals in a linear regression. It means they proposed a certain form of the $(1 - \alpha)$ -pivotal set G and computed tolerance factor based on G as

$$\lambda(\gamma, \mathbf{x}) = \min\{\lambda : C(\mathbf{x}^T \mathbf{b}, \lambda u) \geq \gamma \text{ for all } (\mathbf{b}, u) \in G\}. \quad (16)$$

Wilson [9] used a familiar approximation $(2\chi_{n-q}^2)^{1/2} \sim N([2(n - q) - 1]^{1/2}, 1)$ (see [3]) to approximate distribution of the pivotal quantity u , recall $(n - q)u^2 \sim \chi_{n-q}^2$. Let $v = n - q$, Wilson constructed an $(q + 1)$ -dimensional ellipsoidal pivotal set for the pivotal quantities \mathbf{b}, u with approximate confidence $1 - \alpha$ in the form

$$G_W = \{(\mathbf{b}, u) : \mathbf{b}^T (\mathbf{X}^T \mathbf{X}) \mathbf{b} + 2v(u - k)^2 \leq c\}, \quad (17)$$

where $c = \chi_{q+1}^2(1 - \alpha)$ is $(1 - \alpha)$ -quantile of the chi-squared distribution with $q + 1$ degrees of freedom. Since, matrix $(\mathbf{X}^T \mathbf{X})^{-1}$ is positive definite and regular, we get (see [7])

$$\max_x \frac{(\mathbf{x}^T \mathbf{b})^2}{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}} = \mathbf{b}^T (\mathbf{X}^T \mathbf{X}) \mathbf{b}. \quad (18)$$

Inequality in (17) can be rewritten by $G_W = \{(\mathbf{b}, u) : \mathbf{b}^T (\mathbf{X}^T \mathbf{X}) \mathbf{b} \leq c - 2v(u - k)^2\}$ and using the result (18), we get

$$|\mathbf{x}^T \mathbf{b}| \leq \sqrt{c - 2v(u - k)^2} \delta(\mathbf{x}) \quad \text{for all } \mathbf{x}. \quad (19)$$

Let us denote $A_{\mathbf{x}}(u) = \sqrt{c - 2v(u - k)^2} \delta(\mathbf{x})$. This function of u is defined over the interval $u \in [k - \sqrt{c/2v}, k + \sqrt{c/2v}]$. Wilson proved that if $a = \mathbf{x}^T \mathbf{b}$ and $r = \lambda u$, then

$$G_W \subset H(\lambda, \mathbf{x}) = \{a^2 / \delta^2(\mathbf{x}) + 2v(r/\lambda - k)^2 \leq c\}, \quad (20)$$

for all \mathbf{x} and all $\lambda > 0$. Wilson proposed to bound the curve $S_\gamma = \{(a, r) : C(a, r) = \gamma\}$ defined in R^2 from below by the upper branch of the hyperbola $(r - r_0)^2 - a^2 = h^2$, where $r_0 = \Phi^{-1}(\gamma)$ and h^2 is chosen to achieve a good approximation. He tabulated approximate values of h^2 for certain selected γ , for example $h^2 = 0.0244$ and $h^2 = 0.0438$ for $\gamma = 0.99$ and $\gamma = 0.95$, respectively. Optimal tolerance factor $\lambda = \lambda(\gamma, \mathbf{x})$ based on the pivotal set G_W will lie on the intersection of the set $H(\lambda, \mathbf{x})$ with the hyperbola. Substituting expression a^2 by $(r - r_0)^2 - h^2$ and the inequality sign by the equality sign gives

$$(r - r_0)^2 - h^2 - [c - 2v(r/\lambda - k)^2] \delta^2(\mathbf{x}) = 0. \quad (21)$$

By setting the discriminant of this quadratic equation in r equal to 0 and by solving the resulting quadratic equation in λ , we obtain two real roots, of which the larger is Wilson's tolerance factor.

3.3 The Limam-Thomas method

Limam and Thomas [5] derived the $(1 - \alpha)$ -pivotal set from the product set of $(1 - \alpha/2)$ -level confidence sets for β and σ . The confidence set for the parameter β is a q -dimensional ellipsoid, for σ a one-sided upper confidence interval is used. Applying Bonferroni inequality, the pivotal set is of the form

$$G_{LT} = \{(\mathbf{b}, u) : \mathbf{b}^T (\mathbf{X}^T \mathbf{X}) \mathbf{b} \leq u^2 k_1^2 \quad \text{and} \quad u \geq k_2\}, \quad (22)$$

where $k_1^2 = qF_{q, n-q}(1 - \alpha/2)$ and $F_{q, n-q}(1 - \alpha/2)$ is $(1 - \alpha/2)$ -quantile of F-distribution with q and $n - q$ degrees of freedom, $k_2 = \sqrt{\chi_{n-q}^2(\alpha/2)/(n - q)}$. Like Willson (19), they shifted the problem from the $(q + 1)$ dimensional space to the 2-dimensional using the bound for a linear combination $\mathbf{x}^T \mathbf{b}$, $|\mathbf{x}^T \mathbf{b}| \leq uk_1 \delta(\mathbf{x})$. Then, Limam and Thomas obtained

$$C(\mathbf{x}^T \mathbf{b}, \lambda u) \geq C(uk_1 \delta(\mathbf{x}), \lambda u) \quad \text{for all } (\mathbf{b}, u) \in G_{LT}. \quad (23)$$

The result (23) can be bounded from below using $C(uk_1\delta(\mathbf{x}), \lambda u) \geq C(k_2k_1\delta(\mathbf{x}), \lambda k_2)$ for all $u \geq k_2$, if we consider the restriction $C(uk_1\delta(\mathbf{x}), \lambda u) \geq 1/2$.

The tolerance factor is computed from the root r of the equation $C(k_1k_2\delta(\mathbf{x}), r) - \gamma = 0$ as $\lambda = r/k_2$. The authors suggested as a good initial estimate for the root r point from a hyperbola $r^0 = \Phi^{-1}(\gamma) + \{(\Phi^{-1}[(\gamma + 1)/2] - \Phi^{-1}[\gamma])^2 + (k_1k_2\delta(\mathbf{x}))^2\}^{1/2}$. This method is acceptable only, when the value of γ is at least $1/2$.

3.4 The modified Wilson method

In [5] Limam and Thomas modified Wilson's method. They realized that a smaller value, denoted c_m , of constant c in (17), would consequently lead a smaller tolerance factor. Since the content function of $|\mathbf{x}^T\mathbf{b}|$ is decreasing, using result (19) gives $C(\mathbf{x}^T\mathbf{b}, \lambda u) \geq C(A_x(u), \lambda u)$ for $(\mathbf{b}, u) \in G_W$. The function $A_x(u)$ and λu increases over range $u \in [k, k + \sqrt{c/2v}]$. Thus, the function $C(A_x(u), \lambda u)$ as a function of u is increasing over this interval. Therefore only a subset of G_W corresponding to the interval $[k - \sqrt{c/2v}, k]$, is needed for the determination of λ .

They defined two pivotal sets. First, G_{MW1} , was of the same form as (22) for $u \in [k - \sqrt{c_m/2v}, k]$. The second was constructed in the form

$$G_{MW2} = \{\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} \leq u^2c_m/k^2 \text{ for } u \geq k\}, \quad (24)$$

it intersects the first at $u = k$. It holds $P(G_W) \subset P(G_{MW1} \cup G_{MW2})$ for $c = c_m$ and it is straightforward, that $P(G_W) = P(G_{MW1} \cup G_{MW2})$ implies $c_m < c$. The coefficient c_m can be numerically computed as

$$\int_{\sqrt{v}(k - \sqrt{c_m/(2v)})}^{\sqrt{vk}} P_{\chi_q^2}(c_m - 2(x - \sqrt{vk})^2) f_{\chi_v}(x) dx + \int_{vk^2}^{\infty} P_{\chi_q^2}((c_mx)/(k^2v)) f_{\chi_v^2}(x) dx = 1 - \alpha, \quad (25)$$

where $P_{\chi_q^2}(\cdot)$ denotes the cumulative distribution function of the chi-square random variable with q degrees of freedom, $f_{\chi_v^2}(\cdot)$ denotes the density function of the chi-square distribution with q degrees of freedom and f_{χ_v} denotes the density of the chi distribution with v degree of freedom. In case $n = 15$ and $q = 2$, the value of c_m for $\alpha = 0.01$ and $\alpha = 0.05$, is $c_m = 9.656$ and $c_m = 6.432$, respectively. If coefficient is evaluated, the tolerance factor is determined by Wilson's procedure.

3.5 The Mee-Eberhardt-Reeve method

Likewise Lieberman and Miller [4], Mee et al. [6] considered the form of tolerance factors as a linear function in $d = \delta(\mathbf{x})$ (it depends on \mathbf{x} only through d , which introduces a useful simplification) and proposed simple form

$$\lambda(d, \gamma) = \lambda^*[\Phi^{-1}((1 + \gamma)/2) + \sqrt{q + 2d}], \quad (26)$$

where λ^* is a constant that must be determined for the given α, γ, n, q , and $\Phi^{-1}((1 + \gamma)/2)$ denote the $((1 + \gamma)/2)$ -quantile of a standard normal distribution.

Mee et. al. investigated the distribution of the content $C(\mathbf{x}^T \mathbf{b}, \lambda(\gamma, d)u)$, which depends on independently distributed random variables $|\mathbf{x}^T \mathbf{b}|$ and $u = S/\sigma$, $u \sim \chi_v/\sqrt{v}$. Note that the content function is decreasing in $|\mathbf{x}^T \mathbf{b}|$ for fixed $\lambda(\gamma, d)u$, thus $C(\mathbf{x}^T \mathbf{b}, \lambda(\gamma, d)u)$ is minimized at the maximum for $|\mathbf{x}^T \mathbf{b}|$. Using Cauchy-Schwartz inequality they obtained

$$\max_{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x} = d^2} |\mathbf{x}^T \mathbf{b}| = d \sqrt{\mathbf{b}^T (\mathbf{X}^T \mathbf{X}) \mathbf{b}}. \quad (27)$$

Note that $\mathbf{b}^T (\mathbf{X}^T \mathbf{X}) \mathbf{b} \sim \chi_q^2$, thus $\max_{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x} = d^2} |\mathbf{x}^T \mathbf{b}| \sim d \chi_q$. The width of the tolerance simultaneous intervals was narrowed by limiting the range of possible values of d . For example, in a simple linear regression the smallest possible value of d is $1/\sqrt{n}$, where n is the number of observations. Let $[d_{min}, d_{max}]$ denote the range of d , then the tolerance factor satisfies the following probability statement

$$P(\Phi(d \chi_q + \lambda(d, \gamma) \chi_v/\sqrt{v}) - \Phi(d \chi_q - \lambda(d, \gamma) \chi_v/\sqrt{v})) \geq \gamma \quad d_{min} \leq d \leq d_{max}) = 1 - \alpha. \quad (28)$$

In [6] there are tabulated values of the tolerance factor for different combinations of selected values of α, γ, n and τ , where d_{max} depends on τ , $d_{max} = [(1 + \tau^2)/n]^{1/2}$. In all the stated tables $q = 2$, so $d_{min} = 1/\sqrt{n}$. Mee et. al. also suggested a procedure for the case, when first n_1 elements of all \mathbf{x} are the same. This reduction is possible in the case when $q - r > 1$.

4 Discussion

In section 2 we have described all known methods for computing tolerance intervals in a linear regression. The narrowest tolerance intervals are achieved by the Mee-Eberhardt-Reeve method, but it computes the tolerance factor only for a bounded range of the possible values of $\delta(\mathbf{x})$. The other mentioned methods are conservative, i.e. their actual confidence level exceeds the nominal level $(1 - \alpha)$. In addition, the Wilson method can exceed the captured proportion, because of the approximation hyperbola, and the Limam-Thomas method can be used only in the case, where the proportion at least $1/2$.

The exact likelihood ratio test [1] for testing the simple null hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ against the alternative $H_1 : (\beta, \sigma) \neq (\beta_0, \sigma_0)$ could be used to construct the exact simultaneous confidence region for all parameters of the linear regression model. In particular, the exact $(1 - \alpha)$ -confidence region for the parameters β and σ is given as

$$\mathcal{C}_{1-\alpha}(\mathbf{Y} | \mathbf{X}) = \{(\beta, \sigma) : \lambda(\mathbf{Y} | \mathbf{X}) \leq \lambda_{1-\alpha}\}, \quad (29)$$

where $\lambda(\mathbf{Y} | \mathbf{X})$ is the LRT statistic, whose distribution depends only on the number of observations n and on $q = \text{rank}(\mathbf{X})$ -the rank of the matrix \mathbf{X} and $\lambda_{1-\alpha}$ is the critical value of the LRT. For different number of explanatory variables $q = 1, \dots, 10$, selected small sample sizes, $n = q + 1 : (1) : 40$, $n = 45 : (5) : 100$ and ∞ , and for the usual significance levels $\alpha = 0.1, \alpha = 0.05, \alpha = 0.01$ critical values are presented in [1]. A symmetric simultaneous tolerance intervals based on this confidence set was suggested in [2]. These intervals are constructed such that, no more than a proportion $(1 - \gamma)/2$ of the $Y(\mathbf{x})$ distribution is greater than

$\mathbf{x}^T \hat{\beta} + \lambda(\gamma, \mathbf{x})$ and no more than a proportion $(1 - \gamma)/2$ of the $Y(\mathbf{x})$ distribution is less than $\mathbf{x}^T \hat{\beta} - \lambda(\gamma, \mathbf{x})$. An efficient algorithm for these intervals was suggested in [10]. The approximate values of the simultaneous tolerance intervals are computed based on the Monte Carlo simulations method. The method for computing the tolerance factor for non-symmetric simultaneous tolerance interval is searched for.

Acknowledgement

The paper was supported by the Slovak Research and Development Agency (APVV), grant SK-AT-0003-08 and by the Scientific Grant Agency of the Slovak Republic (VEGA), grant 1/0077/09 and 2/0019/10.

References

- [1] CHVOSTEKOVÁ M., WITKOVSKÝ, V.: *Exact Likelihood Ratio Test for the Parameters of the Linear Regression Model with Normal Errors*. In MEASUREMENT 2009. Proceedings of the International Conference on Measurement, Institute of Measurement Science, SAS, Bratislava, pp. 53-56, 2009.
- [2] CHVOSTEKOVÁ, M., WITKOVSKÝ, V.: *Exact Likelihood Ratio Test for the Parameters of the Linear Regression Model with Normal Errors*. Measurement Science Review, Vol. 9, No. 1, pp 1-8, 2009.
- [3] FISHER R. A.: *Statistical Methods for Research Workers*. 2nd Edition, pp. 96-7, 1928
- [4] LIEBERMAN, G. J., MILLER, R. G., Jr.: *Simultaneous Tolerance Intervals in Regression*. Biometrika, Vol. 50, No. 1/2, pp. 155-168, 1963.
- [5] LIMAM, M. M. T., THOMAS, R.: *Simultaneous Tolerance Intervals for the Linear Regression Model*. Journal of the American Statistical Association, Vol. 83, No. 403, pp. 801-804, 1988.
- [6] MEE, R. W., EBERHARDT, K. R., REEVE, C. P.: *Calibration and Simultaneous Tolerance Intervals for Regression*. Technometrics, Vol. 33, No. 2, pp. 211-219, 1991.
- [7] RAO, R. C.: *Lineární metody statistické indukce a jejich aplikace*. Academia Praha, 1978.
- [8] WALLIS, W. A.: *Tolerance Intervals for Linear Regression*. In Second Berkeley Symposium on Mathematical Statistics and Probability, ed. J. Neyman, Berkeley: University of California Press, pp. 43-51, 1951.
- [9] WILSON, A. L.: *An Approach to Simultaneous Tolerance Intervals in Regression*. The Annals of Mathematical Statistics, 38, pp. 1536-1540, 1967.
- [10] WITKOVSKÝ, V., CHVOSTEKOVÁ M.: *Simultaneous tolerance intervals for the linear regression model*. In MEASUREMENT 2009. Proceedings of the International Conference on Measurement, Institute of Measurement Science, SAS, Bratislava, pp. 28-31, 2009.

Current address

Chvosteková Martina Mgr.

Institute of Measurement Science, Slovak Academy of Sciences

Dúbravská cesta 9, 841 04 Bratislava, Slovakia

e-mail:chvosta@gmail.com

CONFIDENCE INTERVAL FOR COMMON MEAN - A COMPARISON OF TWO METHODS

JANKOVÁ Mária, (SK)

Abstract. In this article, the common mean problem is considered. Based on inter-laboratory comparison of observations from each laboratory, common mean is estimated. In metrology the estimator of the common mean is referred to as the key comparison reference value (KCRV). We concentrate on interval estimation of the common mean. Two methods are compared, metrological approach suggested by Witkovský and Wimmer [2] and generalized pivotal quantity approach suggested by Wang and Iyer [4]. Monte Carlo simulations are used to compare frequentist properties of confidence intervals for the common mean of both methods.

Key words and phrases. common mean, key comparison reference value, confidence interval, metrological approach, generalized pivotal approach

Mathematics Subject Classification. Primary 62F25, 62J10

1 Introduction

In inter-laboratory comparison, where number of observations of the measurand are provided by each participating laboratory, common mean problem arises, when trying to assess the true value of the measurand. Each laboratory has its own systematic error (type B), depending on local conditions. Character of this error is supposed to be known. Uncertainty in each measurement of one laboratory is besides the systematic error formed also by the measurement error (type A). The full character of this error is unknown.

To formally put down this problem, we consider the following model:

$$Y_{ij} = \mu + B_i + E_{ij} \tag{1}$$

for $i = 1, \dots, k$ and $j = 1, \dots, n_i$, where k denotes the number of laboratories involved and n_i number of observations of the i^{th} laboratory. In the above notation μ stands for the unknown value the measurand, B_i for $i = 1, \dots, k$ represent systematic laboratory effects and E_{ij} the measurement errors. We assume that distribution of B_i is known, with mean β_i and variance $\sigma_{B,i}$. B_i are assumed to be independent random variables. We consider E_{ij} to be also independent for $j = 1, \dots, n_i, i = 1, \dots, k$, $E_{ij} \sim N(0, \sigma_{A,i})$, with unknown $\sigma_{A,i}$. E_{ij} and B_i are mutually independent for all $j = 1, \dots, n_i, i = 1, \dots, k$.

For the given model, various methods for approximation of $(1-\alpha) \times 100\%$ confidence interval for the common mean are used. Frequentist approach is applied in [3]. Bayesian approach assuming prior distribution of type A error distributions can be applied. The expected value of the posterior distribution can be used as the estimator of μ . An approach related to the Bayesian methods can be found in [2]. Another frequently used approach is exploitation of generalized pivotal quantities, (e.g. [4]). The two latter approaches will be discussed in this article.

We will study a special case of the model, where we assume that B_i are normally distributed. Our aim is to compare two methods, which both provide different confidence interval estimator for the common mean: interval estimator for μ by Witkovský and Wimmer and generalized interval method by Wang and Iyer.

The comparison will be based on exploring the frequentist properties of two methods, particularly the empirical coverage of the true value of the measurand μ and length of the coverage interval.

We will proceed by introducing the compared methods. In both methods, following notation for sample mean and sample standard deviation is used, respectively: $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$, $S_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$.

2 Methods

2.1 Method based on metrological approach

In method based on metrological approach suggested by Witkovský and Wimmer [2] (from now on referred to as WW method), the approximate confidence interval is computed as an $(1-\alpha) \times 100\%$ confidence interval for the expected value of random variable $\tilde{\mu}$, where $\tilde{\mu}$ has the following form:

$$\tilde{\mu} = \sum_{i=1}^k w_i \bar{y}_i - \sum_{i=1}^k w_i \sqrt{\frac{s_i^2}{n_i}} T_i - \sum_{i=1}^k w_i B_i, \quad (2)$$

where \bar{y}_i and s_i^2 are realizations of Y_i and S_i^2 and $\sum_{i=1}^k w_i = 1$, are weights chosen as:

$$w_i = \left(1 / \left(\sqrt{\frac{s_i^2}{n_i}} \sqrt{\frac{s_p^2}{n_i} \frac{n_i-1}{n_i-3}} + \sigma_{(B),i}^2 \right) \right) / \sum_{l=1}^k \left(1 / \left(\sqrt{\frac{s_l^2}{n_l}} \sqrt{\frac{s_p^2}{n_l} \frac{n_l-1}{n_l-3}} + \sigma_{(B),l}^2 \right) \right), \quad (3)$$

where s_p^2 is the pooled variance estimate $s_p^2 = \sum_{i=1}^k (n_i - 1)s_i^2 / \sum_{i=1}^k (n_i - k)$.

The expected value of random variable $\tilde{\mu}$ is considered as the key comparison reference value (KCRV). Then the approximate confidence interval for μ has the form:

$$(\mu_{KCRV} + q_{\alpha/2}, \mu_{KCRV} + q_{1-\alpha/2}). \quad (4)$$

Quantile q_α is $\alpha\%$ quantile of distribution of $\tilde{\mu} - \mu_{KCRV}$ and can be computed using package *t-dist* in Matlab or R.

The derivation of the formula based on metrological approach is described in detail in [2]. The weights w_i for $i = 1, \dots, k$ are chosen in accordance with approach in Faiweather [1]. Therefore, in a specific case when $\sigma_{B,i} = 0$ for all $i = 1, \dots, k$, the constructed interval would be exact $(1 - \alpha) \times 100\%$ interval for μ .

In [2], the empirical coverage property of the proposed form of confidence interval is examined, taking normal, uniform and triangular distribution of B_i into account.

2.2 Method based on generalized confidence intervals

The generalized confidence interval constructed by Wang and Iyer method [4] (from now on referred to as WI method) is given by $(R_{\alpha/2}, R_{1-\alpha/2})$, where R_β is a $\beta \times 100\%$ quantile of distribution of variable R_μ .

Here R_μ is given by:

$$R_\mu = \frac{\sum_{i=1}^k (\bar{y}_i - B_i) n_i W_i / [(n_i - 1) s_i^2]}{\sum_{i=1}^k n_i W_i / [(n_i - 1) s_i^2]} - Z \sqrt{\frac{1}{\sum_{i=1}^k n_i W_i / [(n_i - 1) s_i^2]}}. \quad (5)$$

W_i is a random variable $W_i \sim \chi_{n_i-1}^2$ and \bar{y}_i and s_i^2 are realizations of \bar{Y}_i and S_i^2 . The derivation of the formula is based on exploitation of generalized pivotal quantity of model parameters.

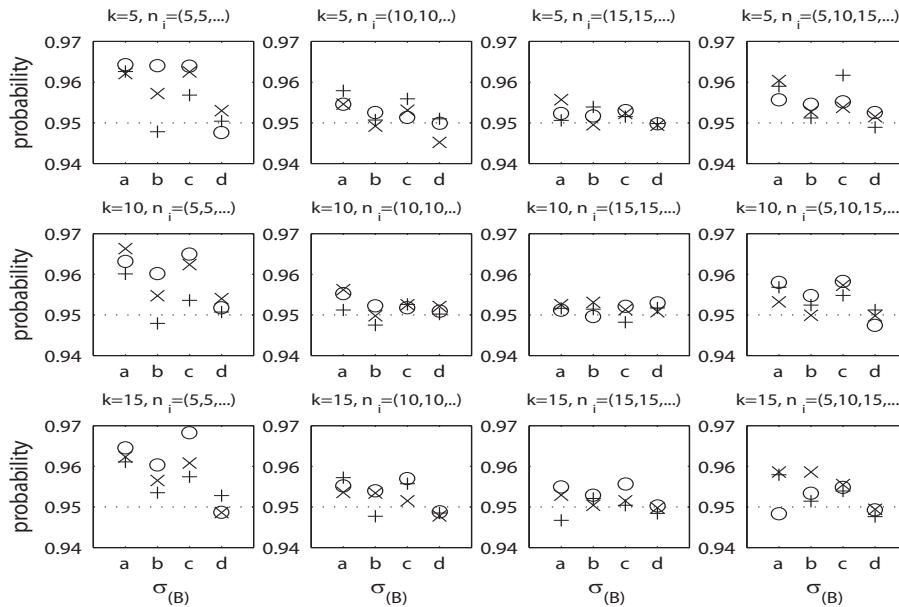
In [4], the frequentist properties, e.g. the empirical coverage probabilities and the lengths of the coverage intervals, which would support the appropriateness of the method, are discussed only for single measurement experiment, where only one laboratory is involved. Thus the coverage properties and the lengths of the coverage intervals for $k > 1$ would be an interesting attribute to gain.

3 Comparison of two confidence interval estimators for the common mean - Simulation study

3.1 Comparison of empirical coverage probabilities

Empirical coverage probabilities of the interval approximations described in 2.1 and 2.2 were gained from 10000 Monte Carlo simulations for each method. These coverage probabilities

Figure 1: Empirical coverage probabilities of $(1 - \alpha) \times 100\%$ confidence interval estimates for μ , where $\alpha = 0.05$, for WW method.

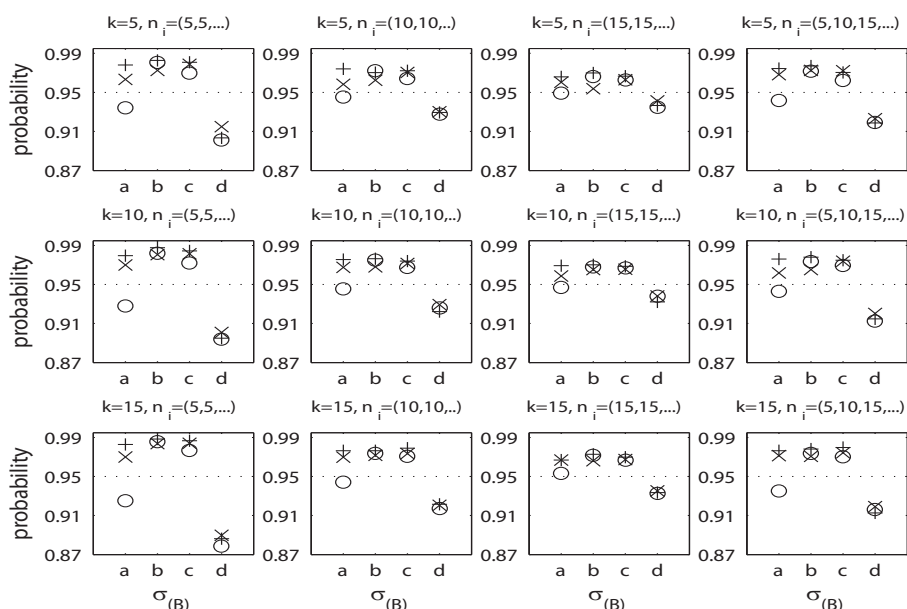


were studied for different combinations of model parameters. For coherence they were chosen the same as in [2]. We suppose that $\mu = 0$ without loss of generality. Testing is performed at significance level $\alpha = 0.05$. The number of participating laboratories is either 5, 10 or 15, i.e. $k \in \{5, 10, 15\}$. As for number of observations in i^{th} laboratory, $n_i = 5$, $n_i = 10$, $n_i = 15$ or $n_i \in \{5, 10, 15, 5, 10, 15, 5, 10, 15, 5, 10, 15, 5, 10, 15\}$, $i = 1, \dots, k$. Designs chosen for $\sigma_{B,i}$ are denoted subsequently: $\sigma_{B,i} = 1$ denoted a , $\sigma_{B,i} = 5$ denoted b , $\sigma_{B,i} \in \{1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5\}$, $i = 1, \dots, k$ denoted c , $\sigma_{B,i} = 0$ denoted d . For chosen designs of $\sigma_{A,i}$ following notation is used: $+$ stands for $\sigma_{A,i} = 1$, o for $\sigma_{A,i} = 5$ and \times for $\sigma_{B,i} \in \{1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5\}$, $i = 1, \dots, k$. Output of the simulations is visualized in Figure 1 for WW method and in Figure 2 for WI method.

3.2 Comparison of length of the intervals

For all designs from section 3.1 the average length of interval for each method was compared to a reference value. As this reference value we have chosen an $(1 - \alpha) \times 100\%$ interval for μ in case where all the parameters of the model except μ are known. This reference confidence interval is constructed by generalized least squares method. By the generalized least squares method, we obtain a BLUE estimate of μ , denoted by $\hat{\mu}_{GLS}$. The BLUE property of $\hat{\mu}_{GLS}$ results from the normal distribution of the systematic error B_i . The distribution of $\hat{\mu}_{GLS}$ is normal, with mean value μ and variance $1 / \sum_1^k x_i$, where $x_i = 1 / (\sigma_{B,i}^2 + \frac{\sigma_{A,i}}{n_i})$. The $(1 - \alpha) \times 100\%$ interval

Figure 2: Empirical coverage probabilities of $(1 - \alpha) \times 100\%$ confidence interval estimates for μ , where $\alpha = 0.05$, for WI method.



takes the form:

$$\left(\hat{\mu}_{GLS} - q_{\alpha/2} \sqrt{\frac{1}{\sum_{i=1}^k x_i}}, \hat{\mu}_{GLS} - q_{1-\alpha/2} \sqrt{\frac{1}{\sum_{i=1}^k x_i}} \right), \quad (6)$$

where q_{β} is $\beta \times 100\%$ quantile of standard normal distribution and $\hat{\mu}_{GLS}$ is given by:

$$\hat{\mu}_{GLS} = \sum_{i=1}^k \left(\frac{\bar{y}_i}{\frac{\sigma_{A,i}^2}{n_i} + \sigma_{B,i}^2} \right) / \sum_{i=1}^k \left(\frac{1}{\frac{\sigma_{A,i}^2}{n_i} + \sigma_{B,i}^2} \right). \quad (7)$$

Figure 3 represents the relative lengths of average intervals for all designs gained by WW method compared to the reference interval, Figure 4 represents the relative length of average intervals gained by WI method.

3.3 Results

For $n_i = 5$, $n_i = 10$, $n_i = 15$ or $n_i \in \{5, 10, 15, 5, 10, 15, 5, 10, 15, 5, 10, 15, 5, 10, 15\}$, $i = 1, \dots, k$ the empirical coverage probabilities appear to be more satisfactory for the WW method according to Figure 1 and Figure 2. Figure 2 indicates that increasing n_i would improve efficiency of WI method. These indications are confirmed in Figure 5, where the designs and notation from 3.1 are preserved, except the number of observations in each laboratory which is for all designs increased by 10. Thus, $n_i = 15$, $n_i = 20$, $n_i = 25$ or $n_i \in \{15, 20, 25, 15, 20, 25, 15, 20, 25, 15, 20, 25, 15, 20, 25\}$, $i = 1, \dots, k$. The same change is done for the length of the average coverage intervals and the output is provided in Figure 6.

Figure 3: Relative lengths of $(1 - \alpha) \times 100\%$ confidence interval estimates for μ , where $\alpha = 0.05$, by WW method, as compared to exact confidence interval with all parameters known.

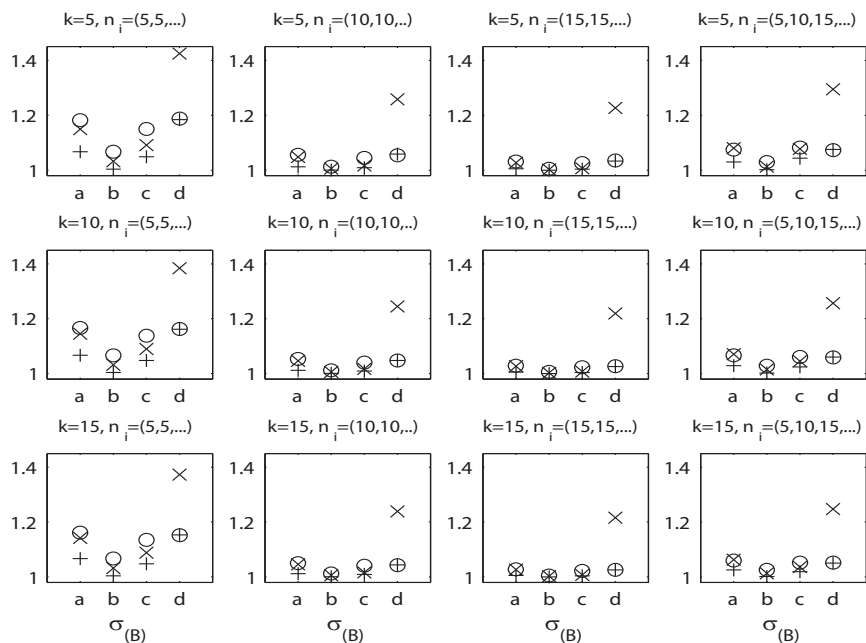


Figure 4: Relative lengths of $(1 - \alpha) \times 100\%$ confidence interval estimates for μ , where $\alpha = 0.05$, by WI method, as compared to exact confidence interval with all parameters known.

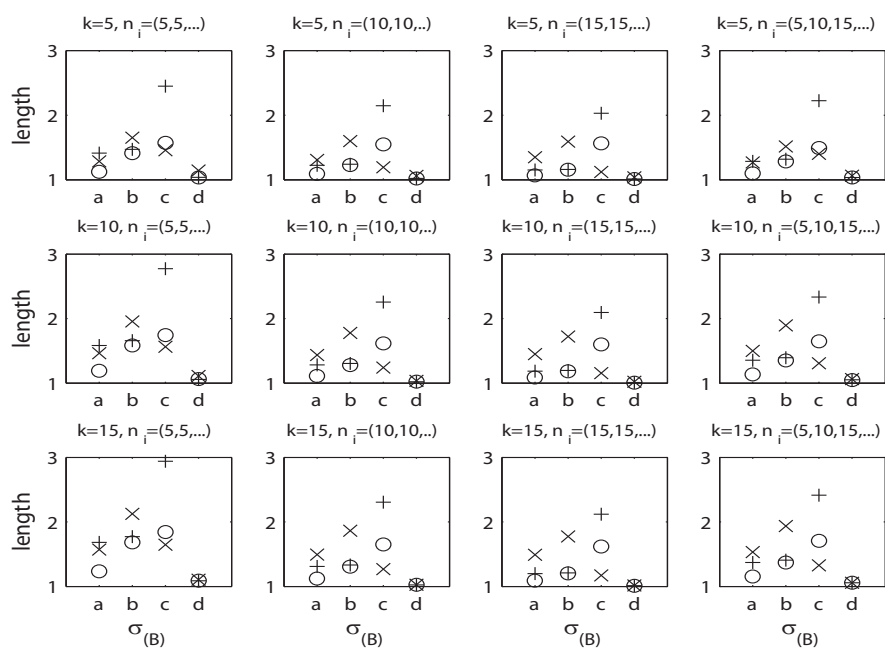


Figure 5: Empirical coverage probabilities of $(1 - \alpha) \times 100\%$ interval estimates for μ , where $\alpha = 0.05$. Here n_i is increased by 10 for all designs. WI method.

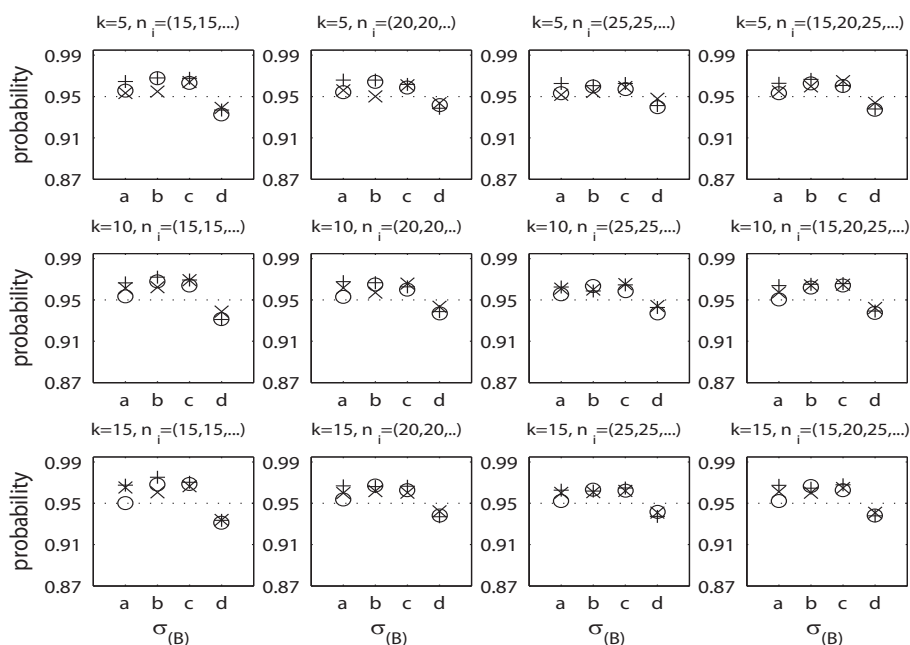
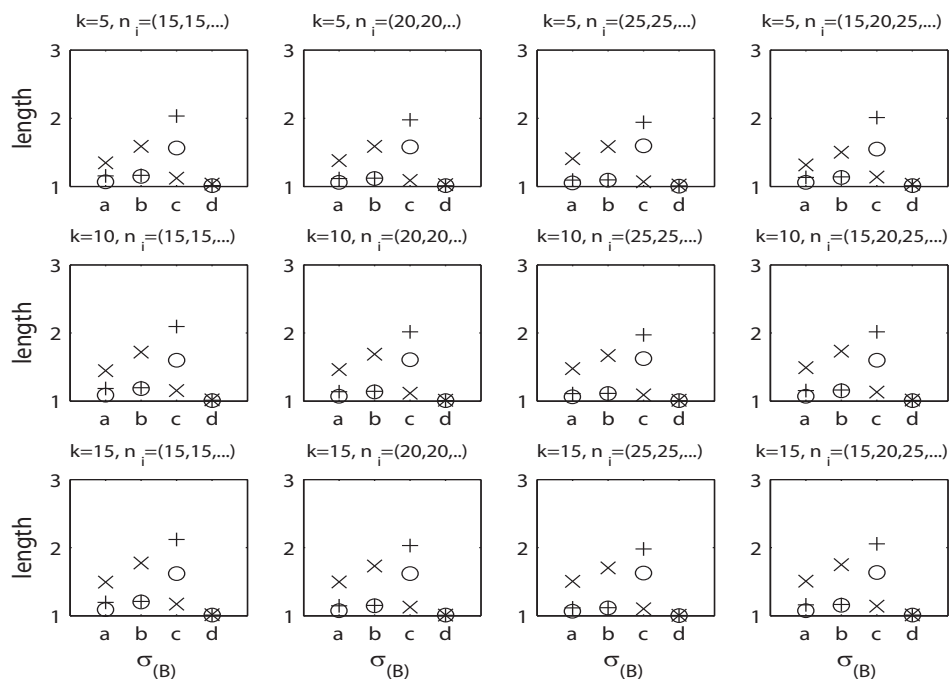


Figure 6: Relative lengths of $(1 - \alpha) \times 100\%$ confidence interval estimates for μ , where $\alpha = 0.05$. Here n_i is increased by 10 for all designs. WI method.



4 Conclusion

Monte Carlo simulations show that method proposed by Wang and Iyer satisfactorily works for higher number of observations provided by the laboratories. However, in case of smaller number of observations, the preferred method is method proposed by Witkovský and Wimmer, due to more appropriate empirical coverage and shorter empirical intervals (relative length as compared to exact confidence interval, when all the parameters are known).

This comparison was executed under the assumption that the systematic error is normally distributed. For more complete information on the topic, a wider range of distributions of B_i should be examined.

Acknowledgment

The paper was supported by the Slovak Research and Development Agency (APVV), grant SK-AT-0003-08 and by the Scientific Grant Agency of the Slovak Republic (VEGA), grant 1/0077/09 and 2/0019/10.

References

- [1] FAIRWEATHER, W. R.: *A method of obtaining an exact confidence interval for the common mean of several normal populations*. Appl. Stat. 21, pp. 229-233, (1972)
- [2] WITKOVSKÝ, V., WIMMER, G.: *Confidence Interval for Common Mean in Interlaboratory Comparisons with Systematic Laboratory Biases*. MEASUREMENT SCIENCE REVIEW, Vol. 7, Section 1, No. 6, pp. 64-73, 2007
- [3] WITKOVSKÝ, V., WIMMER, G.: *Estimation of the common mean and determination of the comparison reference value*. Tatra Mt. Math., Publ. 39, pp. 53-60, 2009
- [4] WANG, C.M., Iyer, H.K.: *A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties*. Measurement, Vol. 39, No. 9, pp. 856-863, 2006

Current address

Mária Janková, Mgr.

Institute of Measurement Science

Slovak Academy of Sciences

Dúbravská cesta 9

842 19 Bratislava, Slovakia

e-mail: majka.jankova@gmail.com

ESTIMATION WITH THE LINEAR MIXED EFFECTS MODEL

JAROŠOVÁ Eva, (CZ)

Abstract. The paper deals with methods of construction confidence intervals based on the linear mixed effects model. Conventional confidence intervals do not take into account the fact that unknown covariance parameters describing the random part of the model must be estimated. Two methods for adjusting these intervals that are implemented in the statistical software product SAS were examined. A small simulation study revealed that estimates of covariance parameters are high unstable and unreliable when they are based on a small data set and that is why adjusted confidence intervals are often worthless.

Key words. Growth curves, confidence and prediction intervals, Satterthwaite's approximation, Kenward-Roger's method

Mathematics Subject Classification: 62P10, 62-07

1 Introduction

The linear mixed effects model is used in situations where the covariance structure of observations is complex. Growth curves representing changes in a response variable observed repeatedly for each experimental unit are a typical example. Observations belonging to the same unit are correlated and moreover, a higher correlation of adjacent observations is expected. Especially when datasets are unbalanced, a generalized model with unrestricted covariance structure does not perform well. On the other hand, the linear mixed effects model enables to include a complex covariance structure in a relatively simple way.

The most important use of the model consists in predicting future values of the response on a given unit. The prediction is based not only on observations obtained from this unit but the information on all the other units is made use of. It is assumed that the curves belonging to different units have a similar shape and that some parameters describing the shape vary randomly across units. The mean profile is described by fixed parameters (fixed effects), random effects distinguish growth curves of different experimental units.

Methods of estimation of both fixed and random effects and of variance components have been described in many works, see e.g. [3], [5] and they are implemented e.g. in SAS and S Plus. Originally, properties of the fixed effects estimators and random effects predictors were derived

under assumption that all covariance parameters are known. Only later on methods reflecting the uncertainty arising from using estimates of the unknown covariance parameters were proposed. The procedure *MIXED* in SAS uses two methods of adjusting. They are referred to as the Satterthwaite's method and the Kenward – Roger's method [6].

In this paper properties of the linear mixed effects model are briefly recapitulated and the principle of the methods of adjusting described. The aim of the study is to examine impacts of the adjustment on the width of confidence intervals for a response value on a given experimental unit in small data sets.

2 Linear mixed effects model

The linear mixed effects model is usually expressed in form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \quad (1)$$

where \mathbf{y} ($N \times 1$) is a vector of responses, \mathbf{X} ($N \times p$) is a design matrix linking $\boldsymbol{\beta}$ to \mathbf{y} , $\boldsymbol{\beta}$ ($p \times 1$) is a vector of unknown parameters (fixed effects), \mathbf{Z} ($N \times q$) is a design matrix linking \mathbf{b} to \mathbf{y} , \mathbf{b} ($q \times 1$) is a vector of unknown random effects and \mathbf{e} ($N \times 1$) is a vector of random errors. Random effects \mathbf{b} and random errors \mathbf{e} are assumed to be independent and normally distributed, $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$. Under these assumptions the mean structure has a form $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and the covariance structure is described by the covariance matrix $\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$ which depends on matrices \mathbf{G} and \mathbf{R} . Usually the elements of \mathbf{G} and \mathbf{R} are assumed to be functions of several parameters $\boldsymbol{\theta}$ ($h \times 1$).

Under the assumption that parameters $\boldsymbol{\theta}$ are known the estimates of $\boldsymbol{\beta}$ and \mathbf{b} obtained by solving the mixed model equations are (see e.g. [5], [6])

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \quad (2)$$

$$\tilde{\mathbf{b}} = \mathbf{G}\mathbf{Z}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \quad (3)$$

In particular the prediction of future values of the response on a given unit $\boldsymbol{\lambda}^T \boldsymbol{\beta} + \boldsymbol{\delta}^T \mathbf{b}$ is requested and is estimated by

$$\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}} + \boldsymbol{\delta}^T \tilde{\mathbf{b}}, \quad (4)$$

what is called the best linear unbiased predictor (BLUP). Usually the unknown parameters $\boldsymbol{\theta}$ are replaced by their estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{V}} = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}^T + \hat{\mathbf{R}}$ and $\hat{\mathbf{G}}$ are substituted in Eq. (2) and (3) and then the attribution BLUP no longer applies.

A variety of procedures were developed to estimate $\boldsymbol{\theta}$. The maximum likelihood method (ML) and the restricted maximum likelihood method (REML) are the best known. Other methods yielding unbiased estimators which are quadratic functions of the data and satisfy some other condition (minimum norm or minimum variance or minimum mean square) are called MINQUE, MIVQUE, MIMSQUE, MIVQUE0 or MINQUE0, I-MINQUE. Comparison of all these methods can be found in [5].

Providing SAS or S Plus is used to apply the linear mixed effects model, only three of them are of interest: ML, REML, and MIVQUE0. There are two substantial differences between ML or REML on one hand and MIVQUE0 on the other hand. The first two are iterative methods, the third one is not. ML and REML require normality assumptions, MIVQUE0 does not. It can be said that MIVQUE0 (MINQUE0 in [5]) is the first iterative solution of REML when the initial value for $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_0^T = (0, 0, \dots, 1)$, 1 corresponding to the variance of random errors. MIVQUE0 is commonly used to compute initial values of $\hat{\boldsymbol{\theta}}$ in REML estimation.

Both ML and REML yield consistent and asymptotically normal estimators with known asymptotic sampling dispersion matrices so that confidence intervals can be constructed and hypotheses about parameters tested. ML estimation takes no account of the degrees of freedom that are involved in estimating fixed effects. To overcome this problem, REML estimation is based on residuals obtained after fitting the fixed part of the model by ordinary least squares. ML estimators are neither unbiased nor of minimum variance. REML estimators have optimal minimum variance properties when the data are balanced [5]. Although it could seem that REML is definitely better for balanced data, a problem may arise in fitting the model. The likelihood ratio test cannot be used to compare models that differ both in fixed and random part. Either information criteria AIC or BIC must be used or ML instead of REML.

3 Confidence intervals

The covariance matrix of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \tilde{\mathbf{b}} - \mathbf{b})$ is given by [4]

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \quad (5)$$

Conventionally, $\hat{\mathbf{C}}$ is obtained by putting $\hat{\mathbf{G}} = \mathbf{G}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{R}} = \mathbf{R}(\hat{\boldsymbol{\theta}})$. An estimator of \mathbf{C} can be written in form (see e.g. [2], [6])

$$\hat{\mathbf{C}} = \begin{bmatrix} \hat{\mathbf{C}}_{11} & \hat{\mathbf{C}}_{21}^T \\ \hat{\mathbf{C}}_{21} & \hat{\mathbf{C}}_{22} \end{bmatrix} \quad (6)$$

where

$$\begin{aligned} \hat{\mathbf{C}}_{11} &= (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \\ \hat{\mathbf{C}}_{21} &= -\hat{\mathbf{G}} \mathbf{Z}^T \hat{\mathbf{V}}^{-1} \mathbf{X} (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \\ \hat{\mathbf{C}}_{22} &= (\mathbf{Z}^T \hat{\mathbf{R}}^{-1} \mathbf{Z} + \hat{\mathbf{G}}^{-1})^{-1} - \hat{\mathbf{C}}_{21} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{Z} \hat{\mathbf{G}}. \end{aligned}$$

Confidence intervals can be expressed generally in the form

$$\mathbf{k}^T \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\mathbf{b}} \end{bmatrix} \pm t_{1-\alpha/2}(\nu) \sqrt{\mathbf{k}^T \hat{\mathbf{C}} \mathbf{k}}, \quad (7)$$

where \mathbf{k} is a $(p+q) \times 1$ vector, $t_{1-\alpha/2, \nu}$ is $1-\alpha/2$ quantile of the t-distribution with ν degrees of freedom, $\nu = N - \text{rank}(\mathbf{XZ})$. In this approach (further called the conventional approach) no account is made for uncertainty in estimating \mathbf{G} and \mathbf{R} . Underestimation of the true variability tends to shorten confidence intervals or in other words, the true confidence level of such intervals will be lower than the claimed one. As a matter of fact, $\nu \frac{\mathbf{k}^T \hat{\mathbf{C}} \mathbf{k}}{E(\mathbf{k}^T \hat{\mathbf{C}} \mathbf{k})}$ has not the χ^2 distribution with ν

degrees of freedom as there are more estimated variance components in $\hat{\mathbf{C}}$. To take this fact into account, approximation by Satterthwaite is done. From matching the first two moments of $\frac{\nu_{SAT} M}{E(M)}$,

where $M = g_1 MS_1 + \dots + g_K MS_K$ and $\nu_k \frac{MS_k}{\sigma_k^2} \sim \chi_{\nu_k}^2$ to those of the χ^2 distribution with ν_{SAT}

degrees of freedom Satterthwaite derived that $\frac{\nu_{SAT} M}{E(M)}$, where $\nu_{SAT} = 2 \frac{E(M)^2}{V(M)}$, has approximately χ^2 distribution with ν_{SAT} degrees of freedom.

Using a Taylor series expansion about $\boldsymbol{\theta}$ with only first-degree terms $M = \mathbf{k}^T \hat{\mathbf{C}} \mathbf{k}$ can be expressed as

$$M = \mathbf{k}^T \hat{\mathbf{C}} \mathbf{k} \cong \mathbf{k}^T \mathbf{C} \mathbf{k} + \sum_{j=1}^h (\hat{\theta}_j - \theta_j) \frac{\partial M}{\partial \theta_j}. \quad (8)$$

It follows that

$$E(M) = \mathbf{k}^T \mathbf{C} \mathbf{k} \quad \text{and} \quad V(M) = \mathbf{g}^T \text{var}(\hat{\boldsymbol{\theta}}) \mathbf{g}, \quad (9)$$

where \mathbf{g} is the gradient of $M = \mathbf{k}^T \hat{\mathbf{C}} \mathbf{k}$ with respect to $\boldsymbol{\theta}$. After the unknown parameters are replaced with their estimates $\hat{\boldsymbol{\theta}}$, the degrees of freedom of the t-quantile in Eq. (7) are adjusted according to

$$\nu_{SAT} = \frac{2(\mathbf{k}^T \hat{\mathbf{C}} \mathbf{k})^2}{\mathbf{g}^T \text{var}(\hat{\boldsymbol{\theta}}) \mathbf{g}} \quad (10)$$

Another procedure consisting in applying an adjusted estimator of \mathbf{C} was proposed by Kenward and Roger [1]. The estimator of $\boldsymbol{\Phi} = \text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}$ is adjusted based on the Taylor series expansion with second-order terms

$$\hat{\boldsymbol{\Phi}} \cong \boldsymbol{\Phi} + \sum_{j=1}^h (\hat{\theta}_j - \theta_j) \frac{\partial \boldsymbol{\Phi}}{\partial \theta_j} + \frac{1}{2} \sum_{i=1}^h \sum_{j=1}^h (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j) \frac{\partial^2 \boldsymbol{\Phi}}{\partial \theta_i \partial \theta_j}. \quad (11)$$

From

$$E(\hat{\boldsymbol{\Phi}}) \cong \boldsymbol{\Phi} + \frac{1}{2} \sum_{i=1}^h \sum_{j=1}^h W_{ij} \frac{\partial^2 \boldsymbol{\Phi}}{\partial \theta_i \partial \theta_j}, \quad (12)$$

where $\mathbf{W} = \text{var}(\hat{\boldsymbol{\theta}})$, the adjusted estimator of $\boldsymbol{\Phi}$ is

$$\hat{\Phi}_A = \hat{\Phi} - \frac{1}{2} \sum_{i=1}^h \sum_{j=1}^h W_{ij} \frac{\partial^2 \Phi}{\partial \theta_i \partial \theta_j}. \quad (13)$$

Beside the conventional method of prediction either the Satterthwaite's approximation itself or Kenward-Roger's method together with the Satterthwaite's approximation can be chosen within the *MIXED* procedure. In S Plus neither confidence limits for the mean value of the response variable nor prediction limits are the part of the output of the procedure *LME*. Even the conventional confidence limits have to be computed additionally afterwards matrix $\hat{\mathbf{C}}$ had been got out according to Eq. (5).

4 Simulation study

To examine impacts of adjustments by Satterthwaite and by Kenward and Roger in case of small samples, some simulations were performed. Balanced data representing six linear growth curves with five repeated observations at the same time points on each of them were considered. The model had the form

$$y_{ijk} = \beta_0 + b_{0,i} + \beta_1 t_j + e_{ijk}, \quad (14)$$

or

$$y_{ijk} = \beta_0 + b_{0,i} + (\beta_1 + b_{1,i}) t_j + e_{ijk}, \quad (15)$$

$$i = 1, 2, \dots, 6; \quad j = 1, 2, \dots, 5; \quad k = 1, 2, \dots, 30.$$

For given parameters β_0, β_1 , random effects $b_{0,i}$ and $b_{1,i}$, and time points t_1, t_2, \dots, t_5 a vector of five random errors e_{ijk} ($j = 1, 2, \dots, 5$) having the multivariate normal distribution $N(\mathbf{0}_5, \mathbf{R}_5)$ with \mathbf{R}_5 corresponding to AR(1) scheme was generated and values of y_{ijk} according to Eq. (14) or (15) were computed. Values of random effects $(b_{0,1}, b_{0,2}, \dots, b_{0,6})$ and $(b_{1,1}, b_{1,2}, \dots, b_{1,6})$ relating to the six growth curves were generated from $N(0, \sigma_0^2)$ and $N(0, \sigma_1^2)$ only once; multiples of these values representing increase in σ_0^2 and σ_1^2 were used in subsequent simulations. 100 simulations were performed for each combination. During each simulation 95% conventional confidence interval and two types of adjusted confidence intervals for the response prediction were computed. For check of coverage of the true value corresponding to the known parameters and random effects in the model the value at t_3 belonging to one of growth curves was chosen. For each combination of covariance parameters the empirical confidence level was determined as a relative frequency of cases when the true value lay within the interval. Besides, the median and the standard deviation of 100 widths of intervals were computed.

Values of β_0, β_1 , parameters of \mathbf{R} , i.e. σ^2 and Φ , and the starting value of σ_0^2 for model (14) were chosen based on the real data set. Values of σ^2 and σ_1^2 for model (15) were chosen so that increasing character of growth curves would be retained for different values of σ_1^2 .

5 Results

At first only random intercepts ($\sigma_1^2 = 0$) were considered. Parameters of the model according to Eq. (14) were $\beta_0 = 1.0895$, $\beta_1 = 0.0934$, $\sigma = 0.037$, $\Phi = 0.8598$. Then six values from the distribution $N(0, \sigma_0^2)$ with $\sigma_0 = \sigma = 0.037$ representing random effects were generated. Their sample standard deviation was 0.029. Multiples of these random effects were considered in further simulations (Table 1).

Table 1. Covariance parameters of the model with random intercepts

		I	II	III	IV	V
$\sigma = 0.037$	$\Phi = 0.8598$	$\sigma_0 = 0.029$	$\sigma_0 = 0.058$	$\sigma_0 = 0.087$	$\sigma_0 = 0.117$	$\sigma_0 = 0.146$

Simulation confirmed the empirical confidence level after adjustment being nearer to the claimed value 0.95 than in case of conventional intervals (Table 2). Confidence levels of both methods of adjustment were similar. The width of all three types of intervals differed only slightly for the most part of data sets. Kenward-Roger's intervals were mostly shorter than Satterthwaite's intervals and conventional intervals were the shortest. In considerable number of cases, however, adjusted intervals were several times longer than conventional intervals. The examples of these dissimilar cases are displayed in Figure 1 (three types of confidence limits are not distinguishable on the left).

Table 2. Empirical confidence levels and characteristics of interval width, model with random intercepts

	Emp. conf. level			Median width			St. dev. of width		
	CON	SAT	KEN	CON	SAT	KEN	CON	SAT	KEN
I	0.74	0.88	0.88	0.0543	0.0624	0.0694	0.0124	0.0796	0.2417
II	0.85	0.90	0.90	0.0649	0.0802	0.0785	0.0217	0.0682	0.4692
III	0.87	0.92	0.92	0.0661	0.0832	0.0797	0.0319	0.1466	0.9658
IV	0.87	0.91	0.90	0.0663	0.0841	0.0803	0.0427	0.2897	1.4216
V	0.90	0.95	0.94	0.0667	0.0849	0.0802	0.0455	0.3284	1.2067

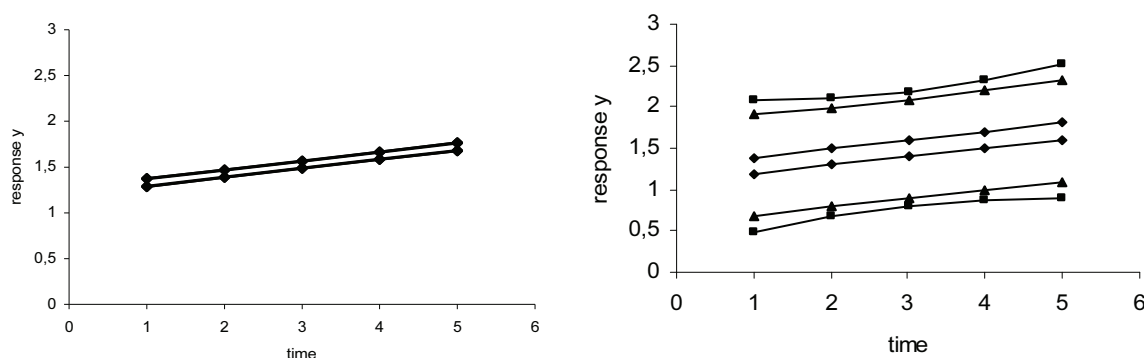


Figure 1. Examples of two dissimilar confidence intervals for the combination V

To find the cause of such a different width the covariance parameter estimates together with approximate standard errors and Wald tests on the output of the *MIXED* procedure in SAS were checked. Whenever the adjusted intervals were strikingly long, P-values of the Wald test of σ^2 were greater than 0.05. An example of such results is in the right part of Table 3. Besides, estimates of covariance parameters differed substantially from the values in Table 1. As noted in [6], Wald tests are valid only asymptotically and they can be unreliable in small samples. It is obvious that covariance estimates are high unstable and that adjusted intervals are unreliable, too.

Table 3. Two examples of covariance parameter estimates and Wald test (*MIXED* proc. in SAS) corresponding to Figure 1 for the combination V

Cov. Par.	Est.	Std Error	Z Value	P Value	Est.	Std Error	Z Value	P Value
σ_0	0.0240	0.0153	1.57	0.0577	0.0159	0.0144	1.11	0.1335
Φ	-0.3116	0.2005	-1.55	0.1202	0.7600	0.4280	1.78	0.0758
σ	0.0008	0.0003	3.37	0.0004	0.0035	0.0063	0.55	0.2898

Further both random intercepts and slopes were considered. Fixed parameters in Eq. (15) were the same as before, covariance parameters are given in Table 4.

Table 4. Covariance parameters of the model with random intercepts and slopes

			VI	VII	VIII
$\sigma = 0.007$	$\Phi = 0.8598$	$\sigma_0 = 0.146$	$\sigma_1 = 0.006$	$\sigma_1 = 0.013$	$\sigma_1 = 0.019$

The results were even worse than in the case of the model with only random intercepts. As for the empirical confidence level, its value worsened with increasing σ_1 (Table 5). At first sight the adjustment improved the coverage of the known true value. But very often either the adjusted intervals were too wide or the calculation of the Kenward-Roger's interval failed. These problems were of course connected with estimation of covariance parameters. In cases when both adjusted intervals were strikingly long the standard error of the estimate of σ was equal to zero. Whenever Kenward-Roger's interval were not calculated it appeared that standard errors of covariance estimates were too large as was revealed by insignificant Wald tests.

Table 5. Empirical confidence levels and characteristics of interval width, model with random intercepts and slopes

	Emp. conf. level			Median width			St. dev. of width		
	CON	SAT	KEN	CON	SAT	KEN	CON	SAT	KEN
VI	0.82	0.95	0.95*	0.0123	0.0215	0.0175*	0.0652	0.4053	0.8494*
VII	0.73	0.95	0.94*	0.0142	0.0349	0.0216*	0.1098	0.6600	58.0929*
VIII	0.58	0.89	0.89*	0.0241	0.1459	0.0548*	0.1193	0.6933	1.6101*

* Values are calculated based on cases with successful adjustment.

6 Conclusion

Simulation study revealed that quality of estimation of covariance parameters and consequently adjustment of confidence intervals in small samples can be quite poor. The estimates of covariance parameters are high unstable and their standard errors are unreliable. This finding corresponds to the note about the Wald tests of covariance parameters in small samples in [6]. The situation worsens with increasing number of covariance parameters. A poor estimation may influence not only the true confidence level of conventional intervals but it can also make the adjustment impossible. It is obvious that even for balanced data three types of intervals may differ substantially. When a striking difference between the conventional and the adjusted intervals is observed, standard errors of covariance parameter estimates should be checked. On the other hand, when the differences between tree types of intervals are small, the adjustment seems to be unnecessary.

References

- [1] KENWARD, M.G., ROGER, J.H.: Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. In *Biometrics*, 53, pp. 983 – 997, 1997.
- [2] McLEAN, R.A., SANDERS, W.L., and STROUP, W.W.: A Unified Approach to Mixed Linear Models. In *The American Statistician*, 45, pp. 54 – 64, 1991.
- [3] PINHEIRO, J.C., BATES, D.M.: *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000.
- [4] SEARLE, S. R.: *Linear Models for Unbalanced Data*, John Wiley & Sons, Inc., New York, 1987.
- [5] SEARLE, S. R., CASELLA, G., McCULLOCH, Ch.E.: *Variance Components*, John Wiley & Sons, Inc., New York, 1992.
- [6] SAS 9.1. Help and Documentation. SAS Institute Inc., Cary, NC, USA, 2002-2003

Current address

doc. Ing. Eva Jarošová, CSc.

Skoda Auto University
 Tr. Vaclava Klementa 864
 293 60 Mlada Boleslav
 Czech Republic
 Phone Number 732469892
 e-mail: jarosova@is.savs.cz

NEW LTPD PLANS FOR INSPECTION BY VARIABLES - CALCULATION AND ECONOMICAL ASPECTS

KLÚFA Jindřich, (CZ)

Abstract. In this paper we shall deal with the acceptance sampling plans when the remainder of rejected lots is inspected. We shall consider two types of LTPD plans – for inspection by variables and for inspection by variables and attributes (all items from the sample are inspected by variables, remainder of rejected lots is inspected by attributes) – see [5]. These plans we shall compare with the corresponding Dodge-Romig LTPD plans for inspection by attributes. We shall report on an algorithm allowing the calculation of these plans when the non-central t distribution is used for the operating characteristic. The calculation is considerably difficult, we shall use an original method and software R. From the results of numerical investigations it follows that under the same protection of consumer the LTPD plans for inspection by variables are in many situations more economical than the corresponding Dodge-Romig attribute sampling plans.

Key words. Sampling inspection by variables, LTPD plans, non-central t distribution, software R.

1. Introduction

Under the assumption that each inspected item is classified as either good or defective (acceptance sampling by attributes) Dodge and Romig in [2] introduced sampling plans which minimize the mean number of items inspected per lot of process average quality

$$I_s = N - (N - n) \cdot L(\bar{p}; n, c) \quad (1)$$

under the condition

$$L(p_i; n, c) = 0.10 \quad (2)$$

(LTPD single sampling plans), where N is the number of items in the lot (the given parameter), \bar{p} is the process average fraction defective (the given parameter), p_i is the lot tolerance fraction defective (the given parameter, $P_i = 100p_i$ is the lot tolerance per cent defective, denoted LTPD), n

is the number of items in the sample ($n < N$), c is the acceptance number (the lot is rejected when the number of defective items in the sample is greater than c), $L(p)$ is the operating characteristic (the probability of accepting a submitted lot with fraction defective p).

Condition (2) protects the consumer against the acceptance of a bad lot – the probability of accepting a submitted lot of tolerance quality p_r (consumer's risk) shall be 0.10. The LTPD plans for inspection by attributes are in [2] extensively tabulated.

2. LTPD plans by variables and attributes

The problem to find LTPD plans for inspection by variables has been solved in [2] under the following assumptions:

Measurements of a single quality characteristic X are independent, identically distributed normal random variables with unknown parameters μ and σ^2 . For the quality characteristic X is given either an upper specification limit U (the item is defective if its measurement exceeds U), or a lower specification limit L (the item is defective if its measurement is smaller than L). It is further assumed that the unknown parameter σ is estimated from the sample standard deviation s .

The inspection procedure is as follows (see [1]):

Draw a random sample of n items and compute

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Accept the lot if

$$\frac{U - \bar{x}}{s} \geq k, \text{ or } \frac{\bar{x} - L}{s} \geq k. \quad (3)$$

The problem is to determine the sample size n and the critical value k . There are different solutions of this problem. In paper [5] similar conditions to the approach of Dodge and Romig in [2] were used for determination of n and k .

Now we shall formulate this problem. Let us consider *LTPD plans for inspection by variables and attributes* – all items from the sample are inspected by variables, but the remainder of rejected lots is inspected only by attributes. Let us denote

c_s^* - the cost of inspection of one item by attributes,

c_m^* - the cost of inspection of one item by variables.

Inspection cost per lot, assuming that the remainder of rejected lots is inspected by attributes (the inspection by variables and attributes), is $n \cdot c_m^*$ with probability $L(p; n, k)$, and $[n \cdot c_m^* + (N - n) \cdot c_s^*]$ with probability $[1 - L(p; n, k)]$. The mean inspection cost per lot process average quality is therefore

$$C_{ms} = n \cdot c_m^* + (N - n) \cdot c_s^* \cdot [1 - L(\bar{p}; n, k)] \quad (4)$$

Now we shall look for the acceptance plan (n, k) minimizing the mean inspection cost per lot of process average quality C_{ms} under the condition

$$L(p_t; n, k) = 0.10. \quad (5)$$

The condition (5) is the same one as used for protection of the consumer by Dodge and Romig in [2]. Let us introduce a function

$$I_{ms} = n \cdot c_m + (N - n) \cdot [1 - L(\bar{p}; n, k)], \quad (6)$$

where

$$c_m = c_m^* / c_s^*. \quad (7)$$

Since

$$C_{ms} = I_{ms} \cdot c_s^*, \quad (8)$$

both functions C_{ms} and I_{ms} have a minimum for the same acceptance plan (n, k) . Therefore, we shall look for the acceptance plan (n, k) minimizing (6) instead of (4) under the condition (5).

For these LTPD plans for inspection by variables and attributes *the new parameter* c_m was defined – see (7). This parameter must be estimated in each real situation. Usually is

$$c_m > 1. \quad (9)$$

Putting formally $c_m = 1$ into (6) (I_{ms} in this case is denoted I_m) we obtain

$$I_m = N - (N - n) \cdot L(\bar{p}; n, k), \quad (10)$$

i.e. the mean number of items inspected per lot of process average quality, assuming that both the sample and the remainder of rejected lots is inspected by variables. Consequently *the LTPD plans for inspection by variables* are a special case of *the LTPD plans by variables and attributes* for $c_m = 1$. From (10) is evident that for the determination of LTPD plans by variables it is not necessary to estimate c_m ($c_m = 1$ is not real value of this parameter).

Summary: For the given parameters p_t , N , \bar{p} and c_m we must determine the acceptance plan (n, k) for inspection by variables and attributes, minimizing I_{ms} in (6) under the condition (5).

In the first place we shall deal with the solution of the equation (5). The operating characteristics is (e.g. [6])

$$L(p; n, k) = \int_{k/\sqrt{n}}^{\infty} g(t; n-1, u_{1-p}\sqrt{n}) dt, \quad (11)$$

where $g(t; n-1, u_{1-p}\sqrt{n})$ is probability density function of non-central t distribution with $(n-1)$ degrees of freedom and noncentrality parameter $\lambda = u_{1-p}\sqrt{n}$.

Instead of (11), using the normal distribution as an approximation of the non-central t distribution (see [4]) we have

$$L(p; n, k) = \Phi\left(\frac{u_{1-p} - k}{A}\right), \quad (12)$$

where

$$A = \sqrt{\frac{1}{n} + \frac{k^2}{2(n-1)}}. \quad (13)$$

The function Φ in (13) is a standard normal distribution function and u_{1-p} is a quantile of order $1-p$, i.e. $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u \exp(-x^2/2) dx$, $u_{1-p} = \Phi^{-1}(1-p)$ (the unique root of the equation $\Phi(u) = 1-p$). The approximation (13) holds both for an upper specification limit U and for a lower specification limit L .

If we use (12) for operating characteristics, the equation $L(p_t; n, k) = 0.10$ has one and only one solution (see [5])

$$k = \frac{u_{1-p_t} - u_{0.10} \cdot h}{g}, \quad (14)$$

where

$$g = 1 - \frac{u_{0.10}^2}{2(n-1)}, \quad h = \sqrt{\frac{g}{n} + \frac{u_{1-p_t}^2}{2(n-1)}}. \quad (15)$$

This is an approximate solution of the equation (5). Exact solution of the equation (5) is

$$k = \frac{t_{0.9}(n-1, u_{1-p_t} \sqrt{n})}{\sqrt{n}}, \quad (16)$$

where $t_{0.9}(n-1, u_{1-p_t} \sqrt{n})$ is a quantile of order 0.9 of non-central t distribution with $(n-1)$ degrees of freedom and noncentrality parameter $\lambda = u_{1-p_t} \sqrt{n}$.

Inserting (14) or (15) into (6) we obtain a function of one variable n

$$I_{ms}(n) = n \cdot c_m + (N - n) \cdot \alpha(n), \quad (17)$$

where $\alpha(n)$ is the producer's risk¹ (the probability of rejecting a lot of process average quality). Now we shall look for the sample size n minimizing (17).

Theorem 1. (Relation between lot size and sample size)

Let \bar{p} , p_t and c_m be given parameters, $0 < \bar{p} < p_t < \frac{1}{2}$, $c_m \geq 1$. Let us denote

¹ Producer's risk is not given for these plans, in [5] it is proved that the producer's risk is non-increasing function of lot size N .

$$F(n) = \frac{c_m - \alpha(n+1)}{\alpha(n) - \alpha(n+1)} + n. \quad (18)$$

If the lot size $N > F(6)$, then there is one and only one $n \in \{7, 8, \dots, N-1\}$ for which holds

$$F(n-1) < N \leq F(n), \quad (19)$$

where $F(n)$ is given by (18). For this sample size n the function (6) has an absolute minimum.

Proof: See [5]

Remark. From (19) it follows (the inverse function F^{-1} to the function F is for $N \geq F(6)$ increasing – see [5]) that

$$n-1 < F^{-1}(N) \leq n, \quad (20)$$

i.e. when N increases, then n does not vary or increases (the sample size n is nondecreasing function of the lot size N).

For the comparison these plans with the corresponding Dodge-Romig LTPD attribute sampling plans from an economical point of view we used parameters E (inspection by variables) and e (inspection by variables and attributes), defined by relations

$$E = \frac{I_m}{I_s} \cdot 100, \quad e = \frac{I_{ms}}{I_s} \cdot 100. \quad (21)$$

The LTPD plans for inspection by variables and attributes are more economical than the corresponding Dodge-Romig plans when

$$e < 100, \quad (22)$$

similarly, if c_m is statistically estimated and the following inequality holds

$$E \cdot c_m < 100, \quad (23)$$

then the LTPD plans for inspection by variables are more economical than the corresponding Dodge-Romig LTPD plans – see [6].

It was shown that under the same protection of the consumer the LTPD plans for inspection by variables and attributes are in many situations **more economical** than the corresponding Dodge-Romig LTPD attribute sampling plans. This conclusion is valid especially for the large lots and for the small values of the lot tolerance fraction defective – see [6].

Similar conclusions were obtained also for the LTPD plans for inspection by variables² with the corresponding Dodge-Romig LTPD plans.

3. Calculation of the LTPD plans by variables and attributes in R

For the calculation of the LTPD plans by variables and attributes we shall use software R – see [7].

Example. Let $N = 450$, $p_t = 0.01$, $\bar{p} = 0.0015$, and $c_m = 1.7$ (the cost of inspection of one item by variables is by 70% higher than the cost of inspection of one item by attributes). We shall look for the LTPD plan for inspection by variables and attributes. Furthermore we shall compare this plan and the corresponding Dodge-Romig LTPD plan for inspection by attributes.

Given parameters ($N = \text{nbig}$, $\bar{p} = \text{pbar}$, b denotes consumer's risk):

```
> nbig=450
> pt=0.01
> pbar=0.0015
> cm=1.7
> b=0.1
```

Approximate solution is (according to (15), (14), (12) and Theorem 1):

```
> k0= function(n_,pt_,b_) {
g= function(n_,b_) 1-qnorm(b_, mean = 0, sd = 1)^2/(2*n_-2);
h= function(n_,pt_,b_) (g(n_,b_)/n_ + qnorm(1-pt_, mean = 0, sd =
1)^2/(2*n_-2))^0.5;
return((qnorm(1-pt_, mean = 0, sd = 1)-qnorm(b_, mean = 0, sd =
1)*h(n_,pt_,b_))/g(n_,b_));
}

> alpha0=function(n_,pt_,b_,pbar_) pnorm((k0(n_,pt_,b_)-qnorm(1-
pbar_))/((1/n_)+(k0(n_,pt_,b_)^2/(2*n_-2)))^0.5)

> nAPPROX=function(nbig_,pt_,b_,pbar_,cm_) {
fF= function(n_,pt_,b_,pbar_,cm_) (cm_-
alpha0(n_+1,pt_,b_,pbar_))/(alpha0(n_,pt_,b_,pbar_)-
alpha0(n_+1,pt_,b_,pbar_))+n_;
return(ceiling(uniroot(function(n_) fF(n_,pt_,b_,pbar_,cm_)-
nbig_,c(5,nbig_))$root)))}
```

² The LTPD plans for inspection by variables and attributes are always more economical than the corresponding LTPD plans for inspection by variables.


```
> planAPPROX=function(nbig_,pt_,b_,pbar_,cm_) {
n_=nAPPROX(nbig_,pt_,b_,pbar_,cm_);
return(list(n=n_,k=k0(n_,pt_,b_))) }
```

```
> planAPPROX(nbig,pt,b,pbar,cm)
```

```
$n
```

```
[1] 67
```

```
$k
```

```
[1] 2.662032
```

Approximate solution is $n = 67$, $k = 2.662032$. For this plan, consumer's risk is only approximately 0.10.

Exact solution is (see (11), (16) and (17), half-intervals method is used and we make use of the approximate solution for specification of interval to be searched for minimum):

```
> lambda= function(n_,p_) qnorm(1-p_, mean = 0, sd = 1)*n_^0.5
```

```
> k=function(n_,pt_,b_) qt(p=1-b_, df=n_-1,
ncp=lambda(n_,pt_))/n_^0.5
```

```
> alpha={function(n_,pt_,b_,pbar_)
pt(k(n_,pt_,b_)*(n_)^0.5,ncp=lambda(n_,pbar_),df=n_-1)};
```

```
> Ims={function(n_,nbig_,pt_,b_,pbar_,cm_) n_*cm_+(nbig_-
n_)*alpha(n_,pt_,b_,pbar_) }
```

```
> fOptimn=function(nbig_,pt_,b_,pbar_,cm_) {
Imsf=function(n_) Ims(n_,nbig_,pt_,b_,pbar_,cm_);
fMinSearch=function(nl_,nu_) ifelse(nl_==nu_,nl_,
ifelse(Imsf(nl_+floor((nu_-nl_)/2)) <= Imsf(nl_+floor((nu_-
nl_)/2)+1),
fMinSearch( nl_,nl_+floor((nu_-nl_)/2)),
fMinSearch(nl_+floor((nu_-nl_)/2)+1,nu_)));
init_=nAPPROX(nbig_,pt_,b_,pbar_,cm_);
return(fMinSearch(init_-10,init_+10))
};
```

```
> planExact=function(nbig_,pt_,b_,pbar_,cm_) {
n_=fOptimn(nbig_,pt_,b_,pbar_,cm_);
return(list(n=n_,k=k(n_,pt_,b_))) }
```

```
> planExact(nbig, pt, b, pbar, cm)
```

\$n

[1] 67

\$k

[1] 2.670840

The LTPD plans for inspection by variables and attributes is

$$n = 67, \quad k = 2.67084.$$

The implementation of our algorithm in R software is very efficient - computation of the result obtained for the problem solved above is almost immediate. Using R, it is thus possible efficiently calculate tables of LTPD acceptance sampling plans for inspection by variables and attributes.

The corresponding LTPD plan for inspection by attributes we find in a book written by Dodge and Domig [2]. For given parameters N , p_i and \bar{p} we have

$$n_2 = 180, c = 0.$$

Comparison of the operating characteristics of these plans (see (11) and e.g. [3] – the operating characteristic for inspection by attributes):

```
> Lt=function(p_,n_,k_) 1-pt(q = k_*n_^0.5, df= n_ - 1,ncp =
qnorm(1 - p_) * n_^0.5)

>Lh=function(nbig_,p_,n_,c_) {(function(i_) sum(
choose(p_*nbig_,i_)*choose((1-p_)*nbig_,n_-i_
)/choose(nbig_,n_))) (seq(0,c_)) )
}
```

For example we get $Lt(\bar{p}, n, k) = Lt(0.0015, 67, 2.67084) = 0.878356$, i.e. the producer's risk³ for the LTPD plan for inspection by variables and attributes is

$$\alpha = 1 - Lt(0.0015, 67, 2.67084) = 0.121644.$$

The producer's risk for the corresponding Dodge-Romig plan is

$$\alpha = 1 - Lh(450, 0.0015, 180, 0) = 0.291528.$$

Graphic comparison of the operating characteristics of the LTPD plans

- for inspection by variables and attributes (67, 2.67084)

³ The consumer's risk is exactly 0.1 ($Lt(p_i, n, k) = Lt(0.01, 67, 2.67084) = 0.1$).

- for inspection by attributes (180, 0):

```
>plot(seq(0:100000)/100000,Lt(seq(0:100000)/100000,67,2.67840),
type="l",xlab="",ylab="",xlim=c(0,0.022),ylim=c(0,1.05),xaxs="i",y
axs="i",las=1,xaxp=c(0,0.02,4),yaxp=c(0,1,5),bty="l",lwd=1.6);

lines(seq(0:100000)/100000,
sapply(seq(0:100000)/100000,function(p_)
Lh(450,p_,180,0)),lty=3,lwd=3)

mtext("p",at=c(0.022),side=1,line=1);
  mtext("L(p)",at=1.07,side=2,las=1,line=1);

legend(0.015,0.9,legend = c("L(p, 180, 0)", "L(p, 67, 2.67840)"),
col = c("black", "black"), lty = c(3,1),lwd=c(3,1.6), merge = TRUE,
trace = FALSE, horiz = FALSE, ncol = 1, bty="n",cex=0.8);
```

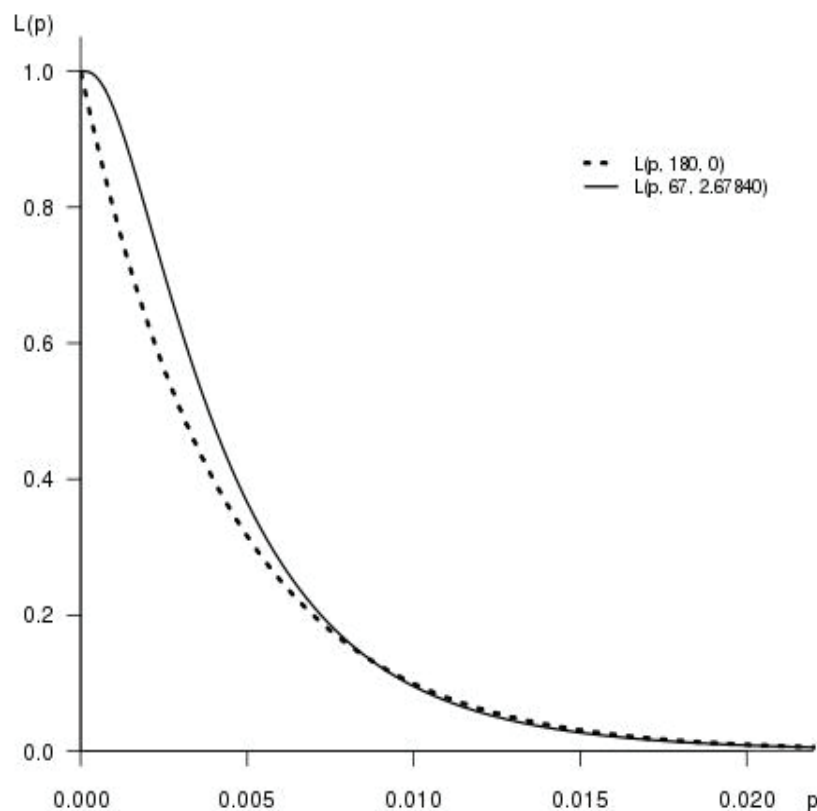


Figure 1

Comparison from an economical point of view (see (21), (6) and (1)):

```
> nbig=450;cm=1.7;pbar=0.0015;n=67;k=2.67084;n2=180;c=0
```

```
> e=100*(n*cm+(nbig-n)*(1-Lt(pbar,n,k)))/(nbig-(nbig-
n2)*Lh(nbig,pbar,n2,c))
```

```
> e
```

```
[1] 62.03395
```

Conclusion

From the results obtained it follows that under the same protection of the consumer the LTPD plan for inspection by variables and attributes (67, 2.67084) is more economical than the corresponding Dodge-Romig attribute sampling plan (180, 0). Since $e = 62.034$ (see the last output), approximately **38% saving of the inspection cost** can be expected.

Furthermore the OC curve for the LTPD plan by variables and attributes (67, 2.67084) is better than the corresponding OC curve for the LTPD plan by attributes – see Figure 1. For example the producer's risk for the LTPD plan by variables and attributes $\alpha = 0.12$ is considerably less than for the corresponding Dodge-Romig plan $\alpha = 0.29$.

Calculation of the plans using our approach is very fast and implementation of our algorithm in R software provides framework sufficiently efficient even for sampling inspection tables calculation.

References

- [1] COWDEN D J: Statistical Methods of Quality Control. Prentice-Hall, Englewood Cliffs, New Jersey, 1957
- [2] DODGE H.F.–ROMIG H.G.: Sampling Inspection Tables. John Wiley, New York, 1998.
- [3] HALD A: Statistical Theory of Sampling Inspection by Attributes. Academic Press, London, 1981
- [4] JOHNSON, N., L. - WELCH, B. L.: Applications of the Non-Central t-Distribution, Biometrika, Vol. 31, No. 3/4 (Mar., 1940), pp. 362-389.
- [5] KLÚFA J. (1994): Acceptance sampling by variables when the remainder of rejected lots is inspected. Statistical Papers 35: 337 – 349.
- [6] KLÚFA J.: Economical Aspects of Acceptance Sampling. Ekopress, Prague, 1999.
- [7] R DEVELOPMENT CORE TEAM (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

Current address

Prof. RNDr. Jindřich Klůfa, CSc.

University of Economics, Department of Mathematics,
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
e-mail: klufa@vse.cz

THE SIMPLE ALTERNATIVES OF THE CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO BINOMIAL PROPORTIONS

POBOČÍKOVÁ Ivana, (SK)

Abstract. The confidence intervals for the difference of two independent binomial proportions are an important problem in biomedical research. Confidence intervals are often used in clinical trials to compare a new treatment with a standard treatment. The most commonly used Wald interval performs poorly. In this paper we propose and evaluate the alternatives of the confidence intervals for the difference of two independent binomial proportions, which have better performance and are attractive for their computational simplicity. We compare the performance of the confidence intervals in terms of the coverage probability and the interval length. We consider these methods: the Newcombe's hybrid-score interval, the Agresti-Caffo interval, the Jeffreys interval.

Key words and phrases. binomial distribution, two binomial proportions, confidence interval, coverage probability, interval length, Wald interval, Newcombe's hybrid-score interval, Agresti-Caffo interval, Jeffreys interval.

Mathematics Subject Classification. Primary 60A05, 62F25.

1 Introduction

The confidence intervals for the difference of two independent binomial proportions are an important problem in a biomedical research. Confidence intervals are often used in clinical trials to compare a new treatment with a standard treatment.

Let $X \sim Bi(n_1, \pi_1)$ and $Y \sim Bi(n_2, \pi_2)$ be two independent binomial random variables. The joint probability mass function is the product of the binomial mass functions of X and Y

$$P(X = x, Y = y) = \binom{n_1}{x} \pi_1^x (1 - \pi_1)^{n_1 - x} \binom{n_2}{y} \pi_2^y (1 - \pi_2)^{n_2 - y}, \quad (1)$$

for $x = 0, 1, \dots, n_1$, $y = 0, 1, \dots, n_2$, $\pi_i \in \langle 0, 1 \rangle$, $n_i \in \mathcal{N}$, $i = 1, 2$.

In practice the values of the parameters π_1 and π_2 are usually unknown and must be estimated from the samples. Let X be a number of successes in a random sample of size n_1 and let Y be a number of successes in a random sample of size n_2 . The maximum likelihood estimators for π_1 and π_2 , respectively, are $p_1 = \frac{X}{n_1}$ and $p_2 = \frac{Y}{n_2}$.

Let $\delta = \pi_1 - \pi_2$ is the difference of two binomial proportions, $-1 \leq \delta \leq 1$. We want to find the $100 \times (1 - \alpha) \%$ two-sided confidence interval $\langle L, U \rangle$ for difference of the two binomial proportions $\delta = \pi_1 - \pi_2$, where $(1 - \alpha)$ is the desired confidence level, $\alpha \in (0, 1)$.

The most commonly used confidence interval is the Wald interval (Wald). This interval is based on the standard normal approximation. The lower and upper bounds of $100 \times (1 - \alpha) \%$ Wald interval are

$$\begin{aligned} L &= (p_1 - p_2) - k_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \\ U &= (p_1 - p_2) + k_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \end{aligned} \quad (2)$$

where k_α is the α -quantile of standard normal distribution $N(0, 1)$.

The Wald interval is simple to compute, but it is known that this interval performs poorly (Agresti, Caffo, 2000, Newcombe, 1998b).

In this paper we propose and evaluate the alternatives of confidence intervals for the difference of two independent binomial proportions, which have much better performance than the Wald interval and are attractive for their computational simplicity: the Newcombe's hybrid-score interval (Newcombe), the Agresti-Caffo interval (Agresti-Caffo) and the Jeffreys interval (Jeffreys). We compare the performance of confidence intervals in terms of the coverage probability and the interval length. We consider the small sample sizes $n_1, n_2 = 5$ to 20.

2 Alternatives of Confidence Intervals

In this section we introduce three alternatives of confidence intervals for the difference of two independent binomial proportions. All these alternatives have following advantages: they have closed-form, are simple to compute and have better performance than the Wald interval.

Newcombe's hybrid-score interval

Newcombe (1998 b) combines the score intervals for the single proportions π_1 and π_2 . Let the lower and upper bounds for π_i be denoted $\langle l_i, u_i \rangle$, $i = 1, 2$, where $\langle l_i, u_i \rangle$ are roots obtained for π_i in the equation

$$|p_i - \pi_i| = k_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi_i(1-\pi_i)}{n_i}}, \quad i = 1, 2. \quad (3)$$

The lower and upper bounds of $100 \times (1 - \alpha) \%$ Newcombe's hybrid-score interval are

$$\begin{aligned} L &= (p_1 - p_2) - k_{1-\frac{\alpha}{2}} \sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}}, \\ U &= (p_1 - p_2) + k_{1-\frac{\alpha}{2}} \sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}}, \end{aligned} \quad (4)$$

where k_α is the α -quantile of standard normal distribution $N(0, 1)$.

Agresti - Caffo interval

Agresti and Caffo (2000) introduced slight modification of the Wald interval by adding one success and one failure into each sample. The point estimators of π_1 and π_2 , respectively, then are $\tilde{p}_1 = \frac{X+1}{n_1+2}$ and $\tilde{p}_2 = \frac{Y+1}{n_2+2}$. The lower and upper bounds of $100 \times (1 - \alpha) \%$ Agresti-Caffo interval are

$$\begin{aligned} L &= (\tilde{p}_1 - \tilde{p}_2) - k_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}, \\ U &= (\tilde{p}_1 - \tilde{p}_2) + k_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}, \end{aligned} \quad (5)$$

where k_α is the α -quantile of standard normal distribution $N(0, 1)$.

Jeffreys interval

This interval is inspired with the Jeffreys interval in one sample situation, that is based on the Bayesian approach (Brown, Cai, DasGupta, 2001). This interval is a pseudo Bayesian confidence interval, that only uses the Bayesian estimators to construct the confidence interval (Brown, Li, 2005). The point estimators of π_1 and π_2 , respectively, are $\tilde{p}_1 = \frac{X + \frac{1}{2}}{n_1 + 1}$ and $\tilde{p}_2 = \frac{Y + \frac{1}{2}}{n_2 + 1}$. The lower and upper bounds of $100 \times (1 - \alpha) \%$ Jeffreys interval are

$$\begin{aligned} L &= (\tilde{p}_1 - \tilde{p}_2) - k_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}, \\ U &= (\tilde{p}_1 - \tilde{p}_2) + k_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}, \end{aligned} \quad (6)$$

where k_α is the α -quantile of standard normal distribution $N(0, 1)$.

3 Criteria for Comparing of Confidence Intervals

In this section we introduce the criteria that are used for comparing of confidence intervals.

Coverage probability

The coverage probability of confidence interval $\langle L, U \rangle$ is for fixed n_1, n_2 and π_1, π_2 defined as

$$C_{n_1, n_2}(\pi_1, \pi_2) = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} I(x, y, \pi_1, \pi_2) \binom{n_1}{x} \pi_1^x (1 - \pi_1)^{n_1-x} \binom{n_2}{y} \pi_2^y (1 - \pi_2)^{n_2-y}, \quad (7)$$

where $I(x, y, \pi_1, \pi_2)$ is an indicator function, defined as $I(x, y, \pi_1, \pi_2) = \begin{cases} 1 & \text{if } \delta \in \langle L, U \rangle \\ 0 & \text{otherwise} \end{cases}$, where $\delta = \pi_1 - \pi_2$.

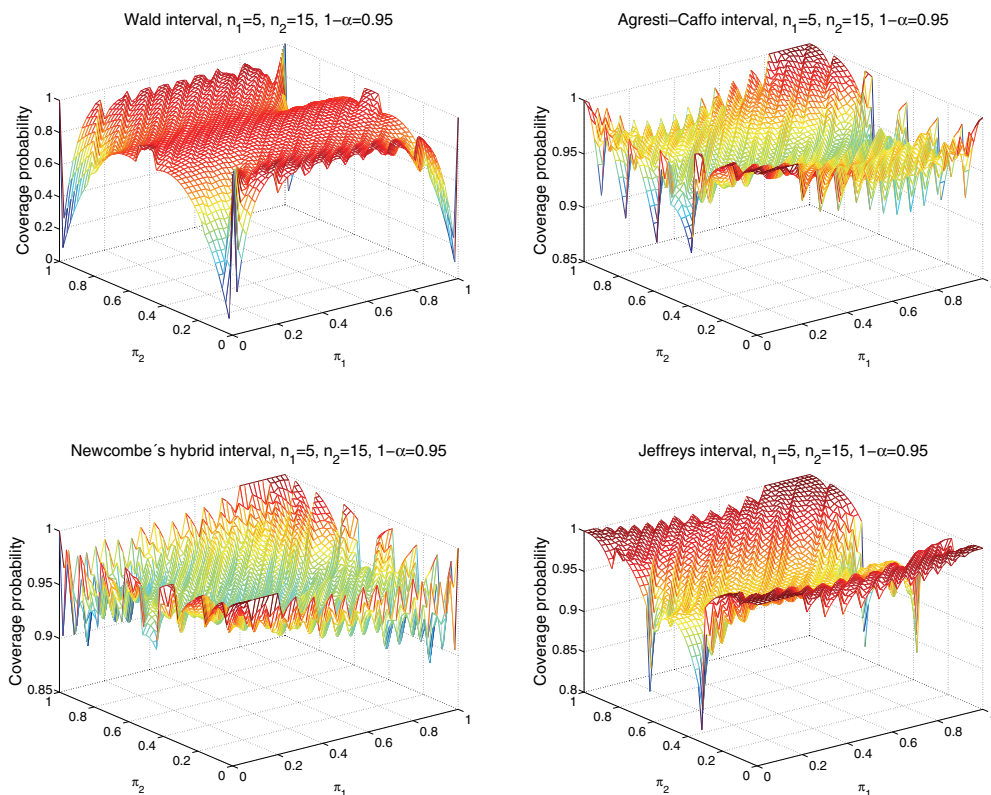


Figure 1: Coverage probability of 95 % confidence intervals for $(n_1, n_2) = (5, 15)$

Average coverage probability

The average coverage probability (AVEC) is defined as

$$AVEC(n_1, n_2) = \int_0^1 \int_0^1 C_{n_1, n_2}(\pi_1, \pi_2) d\pi_1 d\pi_2. \quad (8)$$

Expected length

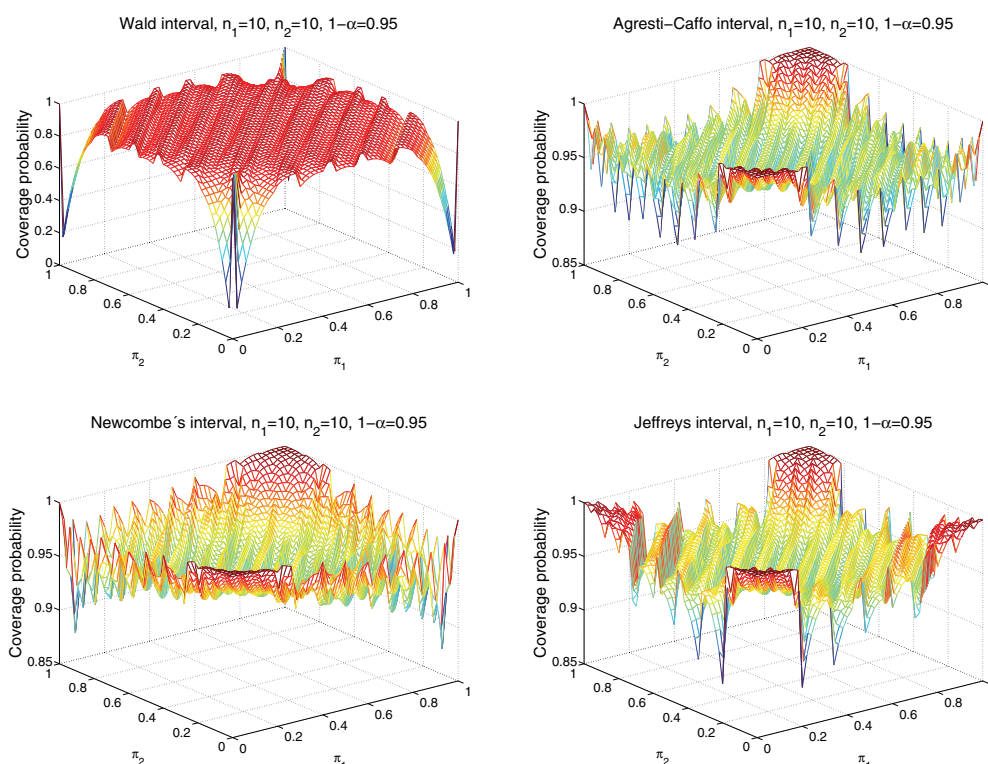
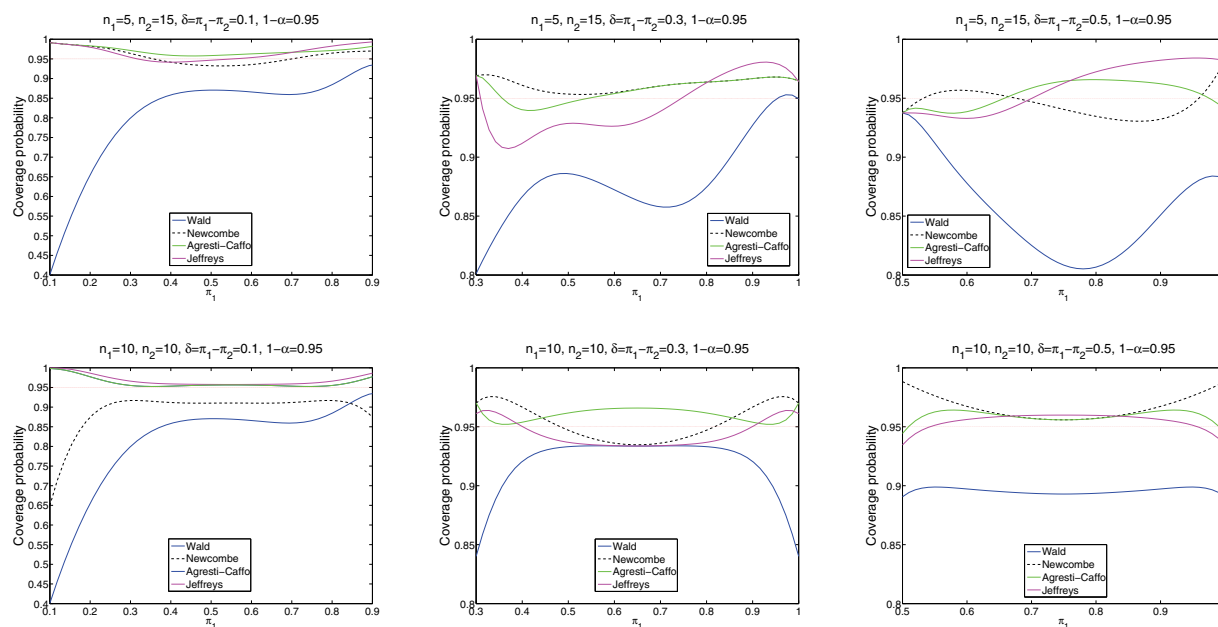


Figure 2: Coverage probability of 95 % confidence intervals for $(n_1, n_2) = (10, 10)$



The expected length of confidence interval is defined as

$$E_{n_1, n_2}(\pi_1, \pi_2) = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} [U(x, y) - L(x, y)] \binom{n_1}{x} \pi_1^x (1 - \pi_1)^{n_1-x} \binom{n_2}{y} \pi_2^y (1 - \pi_2)^{n_2-y}, \quad (9)$$

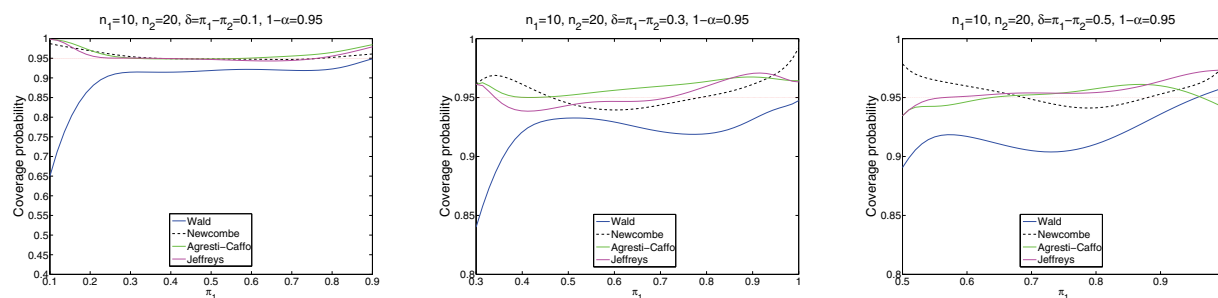


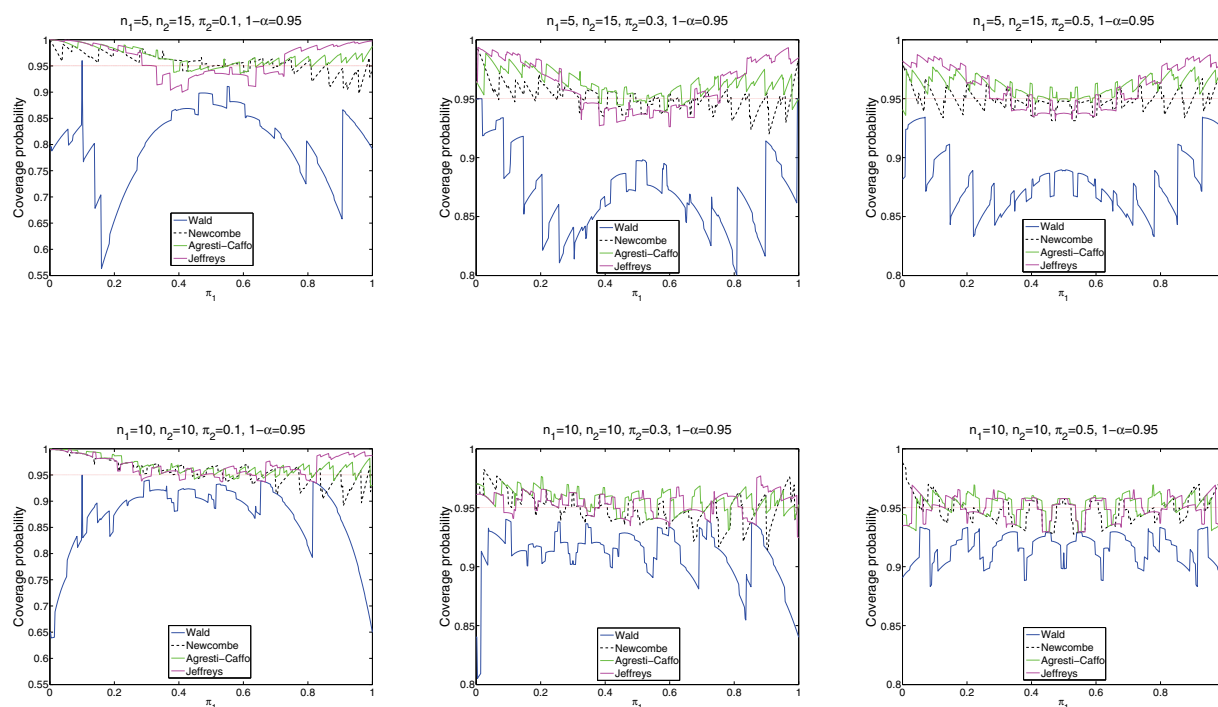
Figure 3: Coverage probability of 95 % confidence intervals for $\delta = \pi_1 - \pi_2 = 0.1; 0.3; 0.5$ and $(n_1, n_2) = (5, 15), (10, 10), (10, 20)$

where $L(x, y)$, $U(x, y)$ are lower and upper bounds of a particular confidence interval.

Average expected length

The average expected length (AVEL) of confidence interval is defined as

$$AVEL(n_1, n_2) = \int_0^1 \int_0^1 E_{n_1, n_2}(\pi_1, \pi_2) d\pi_1 d\pi_2. \quad (10)$$



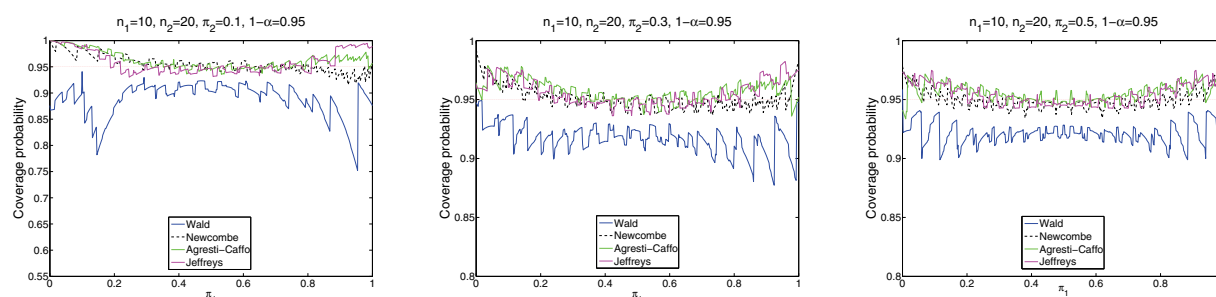


Figure 4: Coverage probability of 95 % confidence intervals for $\pi_2 = 0.1; 0.3; 0.5$ and $(n_1, n_2) = (5, 15), (10, 10), (10, 20)$

4 Comparison of Confidence Intervals

In this section we demonstrate the performance of the confidence intervals which we compare in terms of the coverage probability and the interval length.

Coverage probability. To evaluate and compare the performance of alternatives confidence intervals the coverage probability has been computed. The calculations were performed in Matlab. We consider small sample sizes $n_1, n_2 = 5$ to 20 and $\alpha = 0.05$. For each pair (n_1, n_2) we consider a grid of (π_1, π_2) values from the $\langle 0, 1 \rangle \times \langle 0, 1 \rangle$, which are given by $(\pi_1, \pi_2) = (0.02i; 0.02j)$, $i, j = 0, 1, \dots, 50$. Figures 1 and 2 show the coverage probabilities of 95 % confidence intervals for $(n_1, n_2) = (5, 15)$ and $(10, 10)$.

The coverage probability of the Wald is much below to the nominal level especially when π_1 and π_2 are close to 0 or 1. The Newcombe, the Agresti-Caffo and the Jeffreys perform much better than the Wald. Their coverage probability is generally above the nominal level. The Agresti-Caffo, the Newcombe and the Jeffreys are conservative when π_1 and π_2 are close to 0 or 1, but the Newcombe and the Jeffreys are less conservative.

For fixed $\delta = \pi_1 - \pi_2$ the Newcombe, the Agresti-Caffo and the Jeffreys are very conservative when π_1 is close to 0 and 1.

Figure 3 shows the coverage probabilities of 95 % confidence intervals as a function of π_1 for $(n_1, n_2) = (5, 15), (10, 10)$ and $(10, 20)$. When $\delta = \pi_1 - \pi_2 = 0.1$, π_1 varying over the points given by $\pi_1 = 0.1 + 0.018i$; $i = 0, 1, \dots, 50$. When $\delta = \pi_1 - \pi_2 = 0.3$, π_1 varying over the points given by $\pi_1 = 0.3 + 0.012i$; $i = 0, 1, \dots, 50$. When $\delta = \pi_1 - \pi_2 = 0.5$, π_1 varying over the points given by $\pi_1 = 0.5 + 0.01i$; $i = 0, 1, \dots, 50$.

Figure 4 shows the coverage probabilities of 95 % confidence intervals as a function of π_1 , when $\pi_2 = 0.1; 0.3$ and 0.5 for $(n_1, n_2) = (5, 15), (10, 10)$ and $(10, 20)$. π_1 varying over the points given by $\pi_1 = 0.002i$; $i = 0, 1, \dots, 500$.

Minimum coverage probability. The Wald has minimum coverage probability much below to the nominal level. The Agresti-Caffo has the minimum coverage probability better than other intervals and it is closely followed by the Newcombe and by the Jeffreys.

Figure 5 shows the minimum coverage probabilities of 95 % confidence intervals for $n_1 = n_2$, $n_1 = 10$ and $n_1 = 20$.

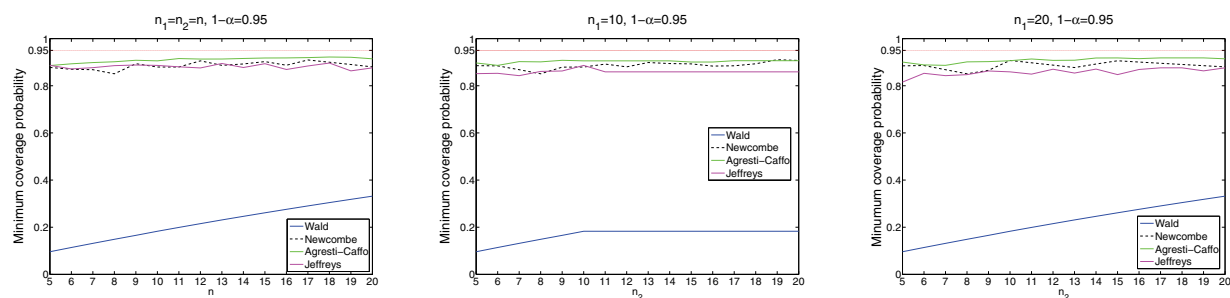


Figure 5: Minimum coverage probability of 95 % confidence intervals

Average coverage probability. The AVEC of the Wald tends to be under others intervals and much below the nominal level. The Newcombe, the Agresti-Caffo and the Jeffreys are slightly conservative on average. The Newcombe has the best AVEC, which is slightly above to the nominal level. The Agresti-Caffo and the Jeffreys are comparable intervals.

Figure 6 shows the average coverage probabilities of 95 % confidence intervals for $n_1 = n_2$, $n_1 = 10$ and $n_1 = 20$.

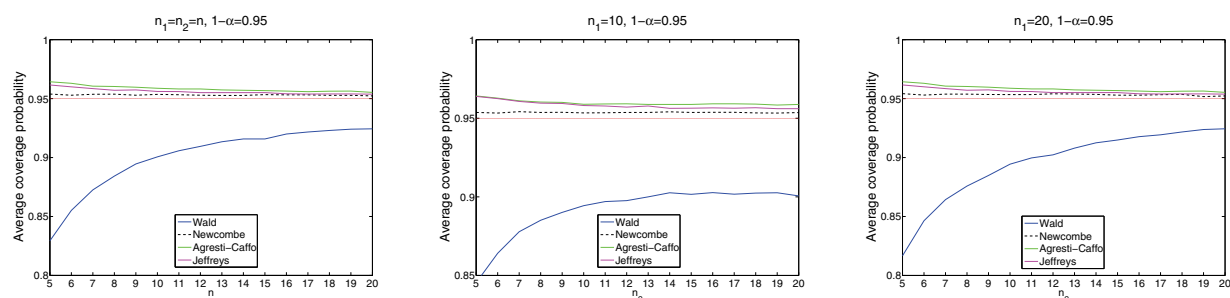


Figure 6: Average coverage probability of 95 % confidence intervals

Expected length. Figure 7 shows the expected lengths of 95 % confidence intervals as a function of π_1 for $\delta = \pi_1 - \pi_2 = 0.1; 0.3; 0.5$ and $(n_1, n_2) = (5, 15), (10, 10)$ and $(10, 20)$. For fixed $\delta = \pi_1 - \pi_2$ the Wald is the shortest when π_1 is close to 0 and 1, while the Newcombe is the widest. For π_1 moderate the Newcombe is the shortest, followed by the Agresti-Caffo, by the Wald and by the Jeffreys. The Jeffreys is wider than others intervals.

Average expected length. The AVEL of the Jeffreys is larger than others intervals. The Wald and the Newcombe are comparable intervals. Their AVEL is smaller. In comparison to them the AVEL of the Agresti-Caffo is larger.

Figure 8 shows the average expected lengths of 95 % confidence intervals for $n_1 = n_2$, $n_1 = 10$ and $n_1 = 20$.

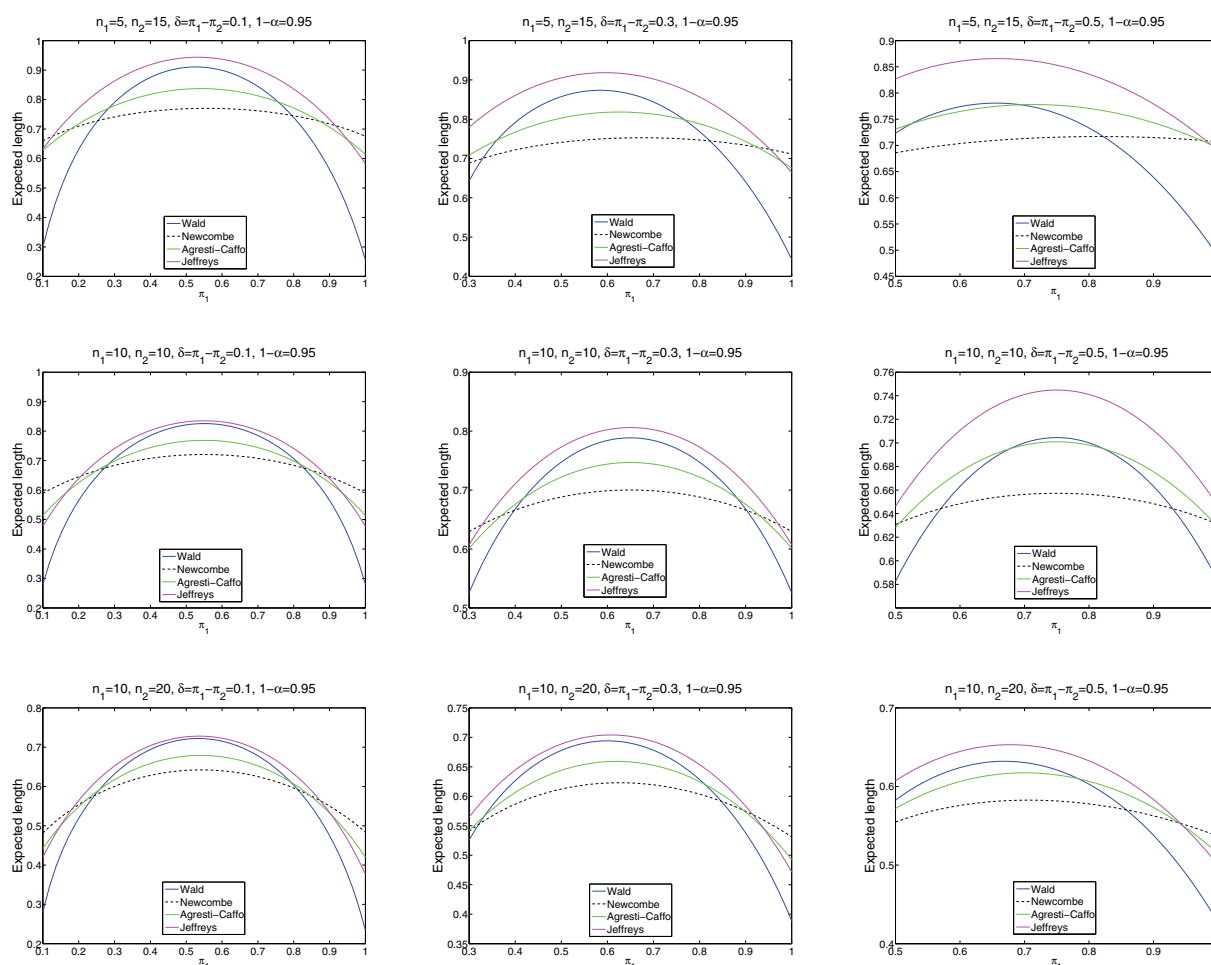


Figure 7: Expected length of 95 % confidence intervals for $\pi_2 = 0.1; 0.3; 0.5$ and $(n_1, n_2) = (5, 15), (10, 10), (10, 20)$

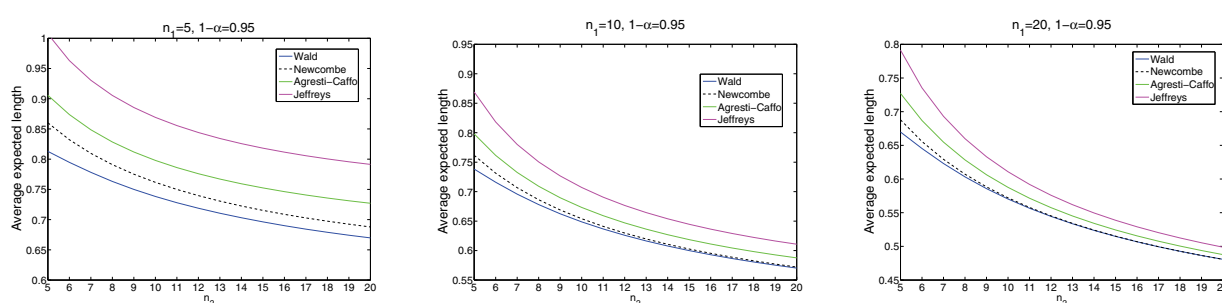


Figure 8: Average expected length of 95 % confidence intervals

Table 1 shows the comparison of 95 % confidence intervals for selected pairs of (n_1, n_2) in terms of the minimum coverage probability (MCP), the average coverage probability (AVEC), the average expected length (AVEL) and the proportion of cases with $C_{n_1, n_2}(\pi_1, \pi_2) < 0.93$.

Table 1.

(n_1, n_2)	Methods	MCP	AVEC	AVEL	< 0.93
(5, 5)	Wald	0.0960	0.8167	0.8129	0.9516
	Newcombe	0.8778	0.9542	0.8601	0.1661
	Agresti-Caffo	0.8857	0.9658	0.9057	0.0285
	Jeffreys	0.8866	0.9681	0.9823	0.0300
(5, 10)	Wald	0.0960	0.8454	0.7383	0.9662
	Newcombe	0.8851	0.9537	0.7616	0.0392
	Agresti-Caffo	0.8967	0.9642	0.7980	0.0177
	Jeffreys	0.8513	0.9639	0.8556	0.0454
(5, 15)	Wald	0.0960	0.8386	0.6963	0.9446
	Newcombe	0.8933	0.9537	0.7153	0.0354
	Agresti-Caffo	0.8967	0.9641	0.7524	0.0161
	Jeffreys	0.8156	0.9628	0.8074	0.0746
(5, 20)	Wald	0.0960	0.8294	0.6697	0.9516
	Newcombe	0.8844	0.9539	0.6879	0.0185
	Agresti-Caffo	0.9005	0.9643	0.7271	0.0115
	Jeffreys	0.8156	0.9616	0.7818	0.1207
(10, 10)	Wald	0.1829	0.8944	0.6488	0.8631
	Newcombe	0.8794	0.9534	0.6540	0.0723
	Agresti-Caffo	0.9057	0.9589	0.6732	0.0169
	Jeffreys	0.8863	0.9581	0.7023	0.0185
(10, 15)	Wald	0.1829	0.9016	0.6000	0.8858
	Newcombe	0.8928	0.9537	0.6026	0.0169
	Agresti-Caffo	0.9007	0.9588	0.6186	0.0115
	Jeffreys	0.8589	0.9564	0.6416	0.0231
(15, 20)	Wald	0.2612	0.9185	0.5149	0.5905
	Newcombe	0.9059	0.9534	0.5152	0.0131
	Agresti-Caffo	0.9180	0.9568	0.5244	0.0123
	Jeffreys	0.8473	0.9546	0.5376	0.0154
(10, 20)	Wald	0.1829	0.9006	0.5703	0.9193
	Newcombe	0.9073	0.9536	0.5720	0.0092
	Agresti-Caffo	0.9062	0.9588	0.5876	0.0077
	Jeffreys	0.8589	0.9561	0.6085	0.0254
(20, 20)	Wald	0.3318	0.9244	0.4807	0.4087
	Newcombe	0.8801	0.9523	0.4807	0.0361
	Agresti-Caffo	0.9143	0.9553	0.4878	0.0108
	Jeffreys	0.8760	0.9538	0.4981	0.0154

5 Concluding Remarks

The better confidence interval is such confidence interval, which coverage probability is close to the nominal level and the AVEC is a little above to the nominal level. The shorter interval and the smaller AVEL are preferred.

The Wald is very simple to compute and use, but performs poorly in terms of the coverage probability. This interval should not be used.

The Newcombe, the Agresti-Caffo and the Jeffreys perform much better than the Wald. These intervals have closed form and are attractive for their computational simplicity. Based on the comparison of these confidence intervals we recommend the Newcombe. The Newcombe perform very well and is the best choice. The Agresti-Caffo and the Jeffreys are also valid choice. They perform also well. The Agresti-Caffo is simpler to compute than the Newcombe and the Jeffreys. His advantage is a presentation.

Acknowledgement

This paper was supported by the Grant VEGA No. 1/0249/09.

References

- [1] AGRESTI, A., CAFFO, B.: *Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures*. In American Statistician, Vol. 54, p. 280-288, 2000.
- [2] BROWN, D.L., CAI, T.T., DAS GUPTA, A.: *Interval estimation for a binomial proportion*. Statistical Science 16, p. 101-133, 2001.
- [3] BROWN, L., LI, X.: *Confidence intervals for two sample binomial distribution*. In Journal of Statistical Planning and Inference. Vol. 130, p. 359-375, 2005.
- [4] CHAN, I. S. F., ZHANG, Z.: *Test-based exact confidence intervals for the difference of two binomial proportions*. In Biometrics, Vol. 55, p. 1202-1209, 1999.
- [5] NEWCOMBE, R.G.: *Two-sided confidence intervals for the single proportion; comparison of several methods*. Statistics in Medicine 17, p. 857-872, 1998a.
- [6] NEWCOMBE, R. G.: *Interval estimate for the difference between independent proportions: comparison of eleven methods*. In Statistics in Medicine, Vol. 17, pp. 873-890, 1998b.
- [7] ZHOU, X.H., TSAO, M., QIN, G.: *New intervals for the difference between two independent binomial proportions*. Journal of Statistical Planning and Inference 123, p. 97-115, 2004.

Current address

Ivana Pobočíková (Mgr.),

Univerzitná 1, Department of Applied Mathematics, Faculty of Mechanical Engineering,
University of Žilina, 010 26 Žilina,
e-mail: ivana.pobocikova@fstroj.uniza.sk

APPLICATION OF NEURAL NETWORKS IN FINANCE

TREŠL Jiří, (CZ)

Abstract. The possibility of application of neural networks for the prediction of both stock and exchange rate returns was investigated. First, the capability of neural networks to reveal specific underlying process was studied using different simulated time series. Second, actual weekly returns from Czech financial markets were analyzed and predicted. Particularly, the problems connected with capturing of outliers and structural breaks were discussed. The predictive power of neural networks was investigated both as a function of network architecture and the length of training set.

Key words. Neural networks, financial time series, predictive power

1 Introductory Remarks to Neural Networks

Artificial neural networks (ANN) are now frequently used in many modelling and forecasting problems, mainly thanks to the possibility of the use of computer intensive methods. Recently, they have been increasingly applied in financial time series analysis as well [1], [2]. The main advantage of this tool is the ability to approximate almost any nonlinear function arbitrarily close. Particularly in financial time series with complex nonlinear dynamical relationships, the ANN can provide a better fit compared with parametric linear models. On the other hand, usually it is difficult to interpret the meaning of parameters and ANN are often treated as „black box“ models constructed for the pattern recognition and prediction. Further, excellent in-sample fit does not guarantee satisfactory out-of-sample forecasting.

Generally, the ANN is supposed to consist of several layers. The *input layer* is formed by individual inputs (explanatory variables). These inputs are multiplied by *connection strengths* which are called *weights* in statistical terminology. Further, there is one or more *hidden layers*, each consisting of certain number of *neurons*. In the hidden layer, the linear combinations of inputs are created and transformed by the *activation functions*. Finally, the *output* is obtained as a weighted mean of these transformed values. Usually, this kind of ANN is referred to as *multilayered feedforward network* and we restrict ourselves to the models with one or two hidden layers. It is useful to realize, information flows only in one direction here, from inputs to output. In time series problems, variables are measured over a time interval and we suppose to exist relationships among

variables at successive times. In this case, our objective is to predict future values of a variable at a given time either from the same or other variables at earlier times. We restrict here to the case, when single numeric variable is observed and its next values is predicted using number of lagged values.

The mathematical representation of the feedforward network with one hidden layer and logsigmoid activation functions is given by the following system [1]

$$\begin{aligned}n_{k,t} &= w_{k,0} + \sum_{i=1}^I w_{k,i} x_{i,t} \\N_{k,t} &= 1 / \left[1 + \exp(-n_{k,t}) \right] \\Y_t &= \gamma_0 + \sum_{k=1}^K \gamma_k N_{k,t} + \sum_{i=1}^I \beta_i x_{i,t}\end{aligned}\tag{1}$$

The first equation describes the creation of linear combination of input variables, whereas second one expresses the transform by logsigmoid activation function. The third equation explains that output value can be obtained either from neurons or from inputs directly. Clearly, if there are no hidden layers, the model reduces to purely linear one.

2 Predictions with Simulated Data

First, the various types of ANN were trained and applied to three kinds of simulated time series. The main aim was to investigate prediction ability with respect to the length of time series (250 or 500), the number of lagged explanatory values (10 or 20) and the number of hidden layers (1 or 2). In each case, 10 last values were used for prediction, so that either 240 or 490 values were left as training data. To quantify the prediction ability, the following goodness of fit measures were computed: *Mean Prediction Error* (MPE), *Mean Deviation of Prediction Errors* (MDPE) and *Mean Absolute Prediction Error* (MAPE)

$$\begin{aligned}MPE &= \frac{1}{h} \sum_{j=1}^h (y_{n+j} - Y_{n+j}) = \frac{1}{h} \sum_{j=1}^h e_{n+j} \\MDPE &= \frac{1}{h} \sum_{j=1}^h |e_{n+j} - \bar{e}| \\MAPE &= \frac{1}{h} \sum_{j=1}^h |e_{n+j}|\end{aligned}\tag{2}$$

In all formulas, n denotes the number of training data and h the prediction length. Further, the following structural notation will be used: *length of time series – number of lagged values – number of neurons in the first hidden layer – number of neurons in the second hidden layer*. For example, the notation 250–10–03–01 specifies 250 data in time series, 10 lagged explanatory values and 3 (resp. 1) neurons in the first (resp. second) hidden layer. All computations were performed with the use of *STATISTICA* software, version 7.

Simulation 1: Deterministic Chaos. Even some simple nonlinear *deterministic* systems can under certain conditions pass to chaotic states due to extremely sensitivity both to initial conditions and control parameters [3]. As an example, let us consider discrete time system described by *logistic difference equation*

$$y_{t+1} = Ay_t(1 - y_t) \quad (3)$$

with control parameter $A=4$, because the most chaotic behaviour is observed just for this value. Clearly, the values from the interval $<0,1>$ will be mapped again into this interval. As for modelling, it is obvious from the following table, longer time series provided better results. On the other hand, the number of lagged values and hidden layers are of minor importance here.

Table 1. Results of ANN Modelling: Deterministic Chaos

Network Type	MPE	MDPE	MAPE
500-20-07-00	-0.048	0.100	0.080
500-20-10-08	-0.013	0.220	0.217
500-10-02-00	-0.139	0.203	0.188
500-10-02-02	-0.079	0.121	0.111
250-20-04-00	-0.042	0.325	0.325
250-20-10-10	+0.065	0.252	0.251
250-10-05-00	-0.066	0.295	0.305
250-10-05-04	-0.082	0.234	0.246

Simulation 2: Bilinear Process. The simplest diagonal form of this process can be written as [4]

$$y_t = \alpha y_{t-1}u_{t-1} + u_t \quad u_t \approx N(0, \sigma^2) \quad (4)$$

Clearly, the first term on right hand side leads to process nonlinearity. On contrary to the previous case, there is no preferred model and the results are comparable.

Table 2. Results of ANN Modelling: Bilinear Process

Network Type	MPE	MDPE	MAPE
500-20-01-00	0.740	1.326	1.431
500-20-02-01	0.428	1.282	1.337
500-10-04-00	0.716	1.257	1.328
500-10-07-02	0.619	1.291	1.346
250-20-01-00	0.746	1.324	1.479
250-20-01-01	0.501	1.348	1.457

250-10-04-00	0.456	1.265	1.320
250-10-05-05	0.409	1.273	1.277

Simulation 3: Kesten Process. This process is a natural generalization of classical AR(1) process to the form [5]

$$y_t = \alpha y_{t-1} + u_t \quad u_t \approx N(0, \sigma^2) \quad \alpha \approx R(a, b) \quad (5)$$

where R denotes regular distribution. Again, MDPE and MAPE exhibit relatively slow variations and the best results are achieved with 250-10-03-00 model.

Table 3. Results of ANN Modelling: Kesten Process

Type	MPE	MDPE	MAPE
500-20-03-00	0.511	1.211	1.144
500-20-06-05	0.551	1.279	1.205
500-10-02-00	0.441	1.062	1.022
500-10-03-02	0.363	1.077	1.005
250-20-01-00	0.314	1.243	1.191
250-20-04-03	0.728	1.402	1.337
250-10-03-00	0.215	0.937	0.887
250-10-01-01	0.595	1.131	1.131

3 Predictions with Financial Data

The main aim was the testing of ANN predictive power in financial applications. We employed weekly logarithmic stock returns (CEZ, KB, TEL, UNIP) and exchange rate returns (CZK/EUR, CZK/GBP, CZK/CHF, CZK/USD) during 2005-2008, i.e. 200 weekly values for each time series. In all cases, 25 preceeding values were used and last 12 values were left for prediction testing. Further, both linear models and neural networks with one and two hidden layers were applied.

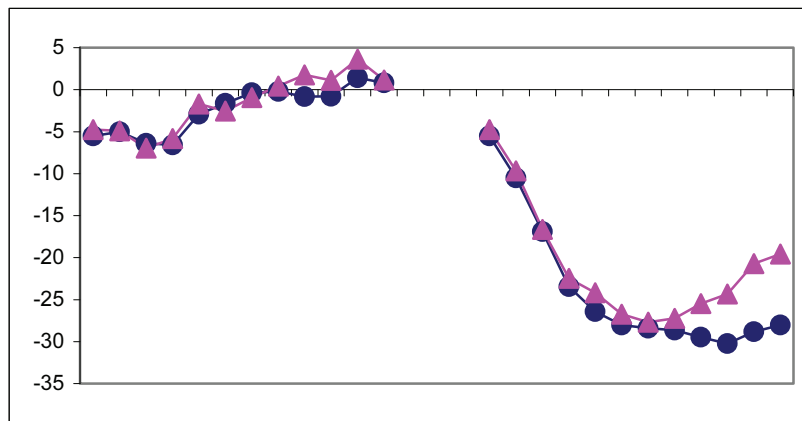


Figure 1. CEZ Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-05-00

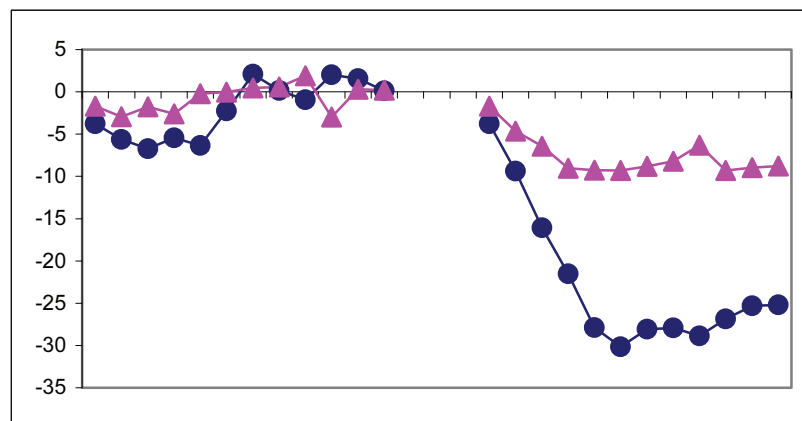


Figure 2. KB Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-12-10

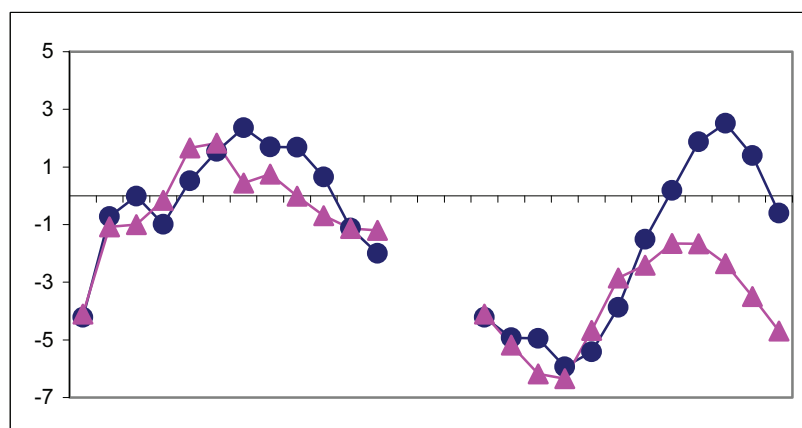


Figure 3. TEL Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-12-06

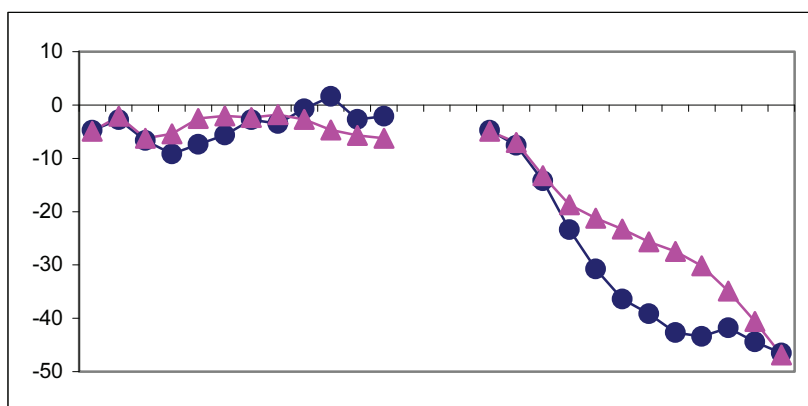


Figure 4. UNIP Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-03-04

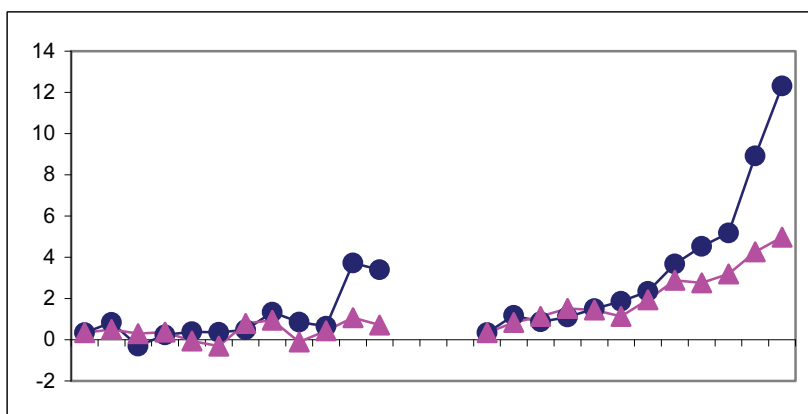


Figure 5. CZK/EUR Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-06-00

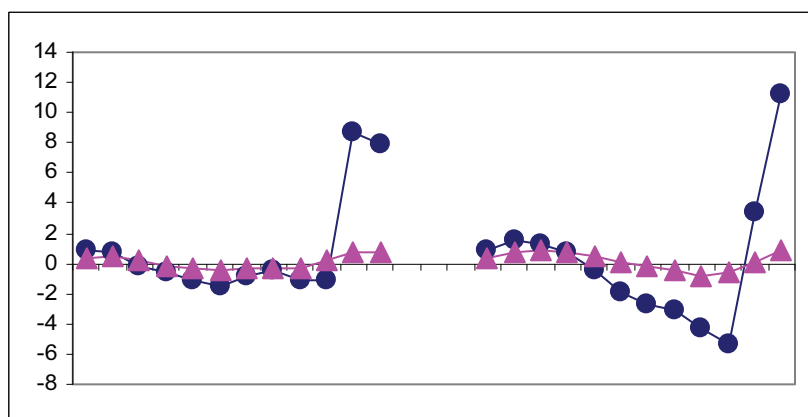


Figure 6. CZK/GBP Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-06-00

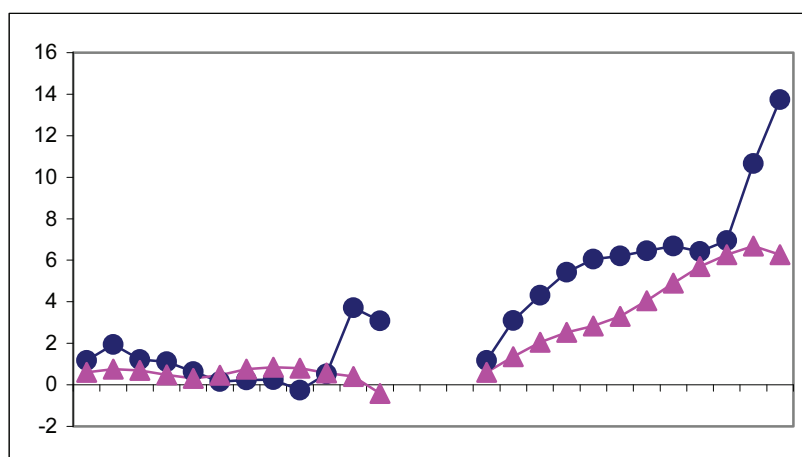


Figure 7. CZK/CHF Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-12-06

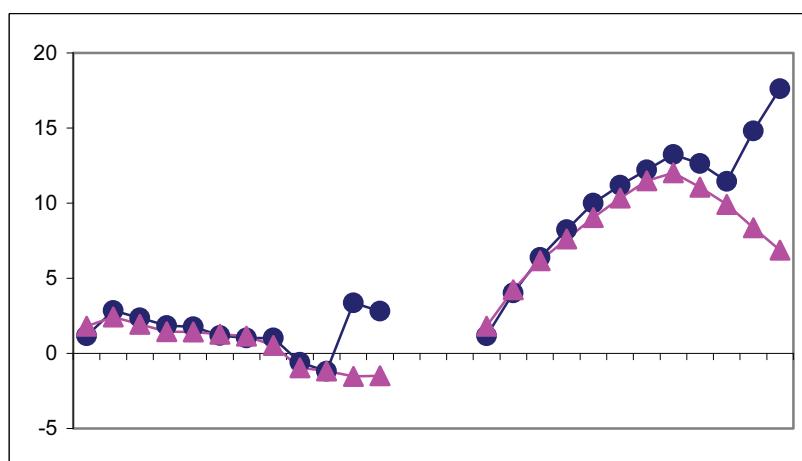


Figure 8. CZK/USD Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-05-00

Table 4. Results of ANN Modelling: Stock Returns

Stock	Model	Type	MPE	MDPE	MAPE
CEZ	Linear		-1.550	1.227	1.910
	One Layer	200-25-05-00	-0.704	0.869	1.054
	Two Layers	200-25-06-02	-1.565	1.281	2.075
KB	Linear		-2.197	3.211	3.460
	One Layer	200-25-04-00	-1.833	2.964	3.138
	Two Layers	200-25-12-10	-1.368	2.356	2.672
TEL	Linear		+0.906	1.322	1.296
	One Layer	200-25-01-00	+0.527	1.572	1.538

	Two Layers	200-25-12-06	+0.342	0.865	0.868
UNIP	Linear		-2.611	1.892	3.044
	One Layer	200-25-02-00	-0.153	2.897	2.943
	Two Layers	200-25-03-04	+0.028	2.597	2.592

Table 5. Results of ANN Modelling: Exchange Rate Returns

Stock	Model	Type	MPE	MDPE	MAPE
CZK/EUR	Linear		+1.087	0.823	1.121
	One Layer	200-25-06-00	+0.610	0.753	0.786
	Two Layers	200-25-02-01	+1.273	0.830	1.290
CZK/GBP	Linear		+0.730	2.389	2.059
	One Layer	200-25-06-00	+0.870	2.237	1.787
	Two Layers	200-25-04-02	+1.090	2.343	1.801
CZK/CHF	Linear		+1.537	0.959	1.537
	One Layer	200-25-08-00	+0.152	1.293	1.247
	Two Layers	200-25-12-06	+0.623	1.021	1.048
CZK/USD	Linear		+0.913	1.104	1.519
	One Layer	200-25-05-00	+0.896	1.236	1.044
	Two Layers	200-25-03-04	-0.642	1.253	1.395

4 Conclusion

The first group of findings is related to artificial data. Undoubtedly, the best results were obtained for deterministic chaotic process, because there is relatively simple relation between neighbouring values. Second, the process itself is bounded between zero and one and the notion of outliers is meaningless here. On the other hand, the results for both bilinear and Kesten processes are markedly worse due to ability to create sudden random excursions. Further, the results seem to be similar for the lengths of time series used (500 and 250), number of lagged values (20 and 10) and number of hidden layers (1 and 2).

As for stock returns, different kinds of individual behaviour were revealed. Both for CEZ and TEL, there has been good agreement between real data and predictions till 7th week and some deviations occurred after this time. On the other hand, predictions of UNIP returns balanced out with respect to their sign, so that corresponding mean predicted error was very small. The worst results were achieved in the case of KB returns, where both actual values and predictions exhibited negative signs up to sixth week, but absolute values of predictions were systematically lower. In most cases, neural networks with two hidden layers turned out to be the best alternative.

Exchange rate returns exhibited similar behaviour, but there were strongly manifested outliers. In all cases, 11th and 12th actual values were strong positive outliers with markedly worse predictions. Thus, the corresponding deviations occurred, but the general agreement between actual values and predictions has been observed till 10th week for CZK/USD and CZK/EUR returns, whereas CZK/CHF one manifested some kind of compensation. Further, the signs of actual values

and predictions in 11th and 12th weeks were the same for CZK/EUR and CZK/GBP and opposite in the case of CZK/USD. In most cases, neural networks with one hidden layers turned out to be the best alternative.

References

- [1] McNELIS, P.D.: *Neural Networks in Finance*. Elsevier Academic Press, Amsterdam, 2005.
- [2] FRANCES, P.H., van DIJK, D.: *Non-Linear Time Series Models in Empirical Finance*. Cambridge University Press, Cambridge, 2000.
- [3] HILBORN, R.: *Chaos and Nonlinear Dynamics*. Oxford University Press, Oxford, 2001.
- [4] TSAY, R.S.: *Analysis of Financial Time Series*. Wiley, New York, 2002.
- [5] SORNETTE, D.: *Critical Phenomena in Natural Sciences*. Springer, 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

