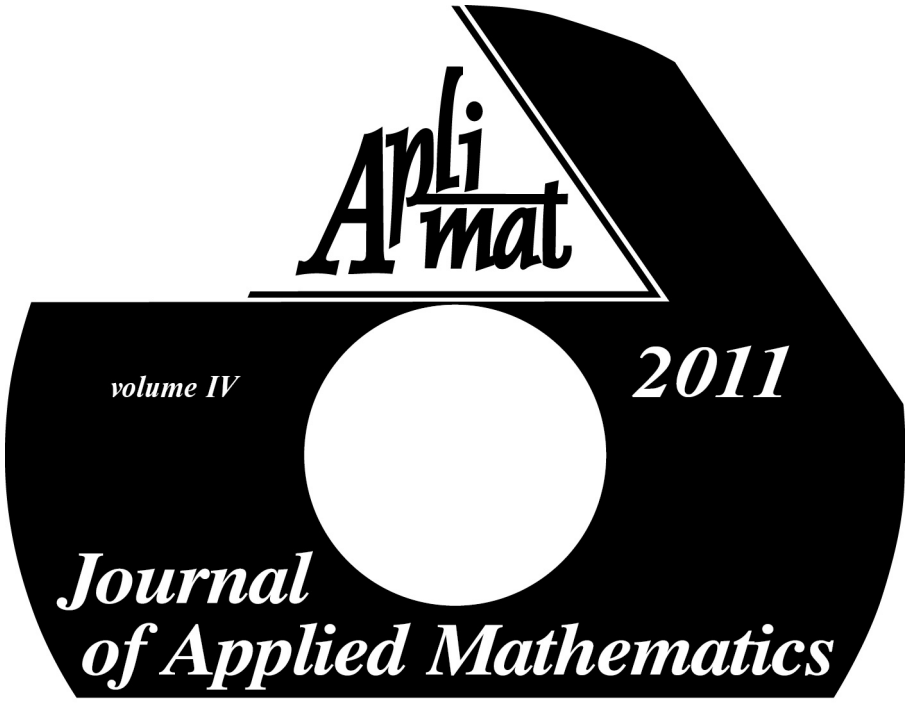


ISSN: 1337-6365
© Slovak University of Technology in Bratislava
All rights reserved

APLIMAT - JOURNAL OF APPLIED MATHEMATICS

VOLUME 4 (2011), NUMBER 3

APLIMAT - JOURNAL OF APPLIED MATHEMATICS

VOLUME 4 (2011), NUMBER 3

Edited by: Slovak University of Technology in Bratislava

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

Editorial Board: CARKOVS Jevgenijs (Latvia)
CZANNER Gabriela (Great Britain)
CZANNER Silvester (Great Britain)
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)
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

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

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

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

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

Copyright © STU 2007-2011, 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 4 (2011), NUMBER 3

FINANCIAL AND ACTUARY MATHEMATICS

APLIMAT - JOURNAL OF APPLIED MATHEMATICS
VOLUME 4 (2011), NUMBER 3

FUZZY MATHEMATICS AND ITS APPLICATIONS

APLIMAT - JOURNAL OF APPLIED MATHEMATICS
VOLUME 4 (2011), NUMBER 3

STATISTICAL METHODS IN TECHNICAL AND ECONOMIC SCIENCES AND PRACTICE

LIST OF REVIEWERS

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

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

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

Pospíšil Jiří, Prof., Ing., CSc.	Czech Technical University in Prague, Prague, Czech Republic
Potuzak Tomas, Ing., PhD.	University of West Bohemia, Plzen, |Czech Republic
Pulpan Zdenek, Prof. RNDr., PhDr., CSc.	University of Hrasdec Kralove, Hradec Kralove, Czech Republic
Řezanková Hana, Prof.	University of Economics, Praha, Czech Republic
Růžičková Miroslava	Žilina University, Žilina, Slovak Republic
Slavík Jan Josef, doc. PaedDr., CSc.	University of West Bohemia, Pilsen, Czech Republic
Smetanová Dana, RNDr., PhD.	Palacky University, Olomouc, Czech Republic
Sousa Cristina Alexandra, Master	Universidade Portucalense Infante D. Henrique, Porto, Portugal
Stachová Maria, Mgr., PhD.	Matej Bel University, Banská Bystrica, Slovak Republic
Stankovičová Iveta, Ing., PhD.	Comenius University, Bratislava, Slovak Republic
Svoboda Zdeněk, RNDr., CSc.	Brno University of Technology, Brno, Czech Republic
Swaczyna Martin, RNDr., PhD.	Ostravská Univerzita, Ostrava, Czech Republic
Sýkorová Irena, RNDr	University of Economics, Praha, Czech Republic
Šamšula Pavel, doc., PaedDr., CSc.	Charles University, Prague, Czech Republic
Šír Zbyněk, RNDr., PhD.	Charles University in Prague, Prague, Czech Republic
Tomáš Jiří, doc. RNDr., PhD.	Brno University of Technology, Brno, Czech Republic
Trešl Jiří, doc. Ing., CSc.	University of Economics, Prague, Czech Republic
Trokanová Katarina, doc.	Slovak Technical University, Bratislava, Slovak Republic
Tvrdík Josef, Assoc. Prof.	University of Ostrava, Ostrava, Czech Republic
Uddin Md. Sharif, PhD	Khulna University, Khulna, Bangladesh
Vaníček Jiří, PhD.	University of South Bohemia, Ceske Budejovice, Czech Republic
Vanžurová Alena, doc. RNDr., CSc.	Palacký University, Olomouc, Czech Republic
Velichová Daniela, doc. RNDr., PhD.	Slovak University of Technology, Bratislava,Slovak Republic
Volna Eva, doc. RNDr. PaedDr., PhD.	University of Ostrava, Ostrava, Czech Republic
Volný Petr, RNDr., PhD.	VŠB - Technical University of Ostrava, Ostrava, Czech Republic

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

Journal of Applied Mathematics

volume IV (2011), number III

THE IMPACT OF SERIAL CORRELATION ON RISK HEDGING

EGLE Aigars, (LV)

Abstract. There are a number of publications that documents the predictability of financial asset returns. In our paper we develop a continuous diffusion model for the case of serially correlated stock returns. We obtain European call option pricing formula written on a stock with autocorrelated returns and show that even small levels of predictability due to serial correlation can give substantial deviation from results obtain by Black-Sholes formula. Finally, we derive formulas for sensitivities of the value of European call option and show how in risk management widely used option hedging parameters depend on assumptions made about correlation in underlying asset returns.

Key words and phrases. serial correlation, diffusion approximation, option pricing

Mathematics Subject Classification. Primary 60A05, 08A72; Secondary 28E10.

1 Introduction

1.1 Empirical Evidence of the Predictability of Asset Returns

There is a wide list of financial research that documents the predictability of financial asset returns. Auto correlation in short term stock index returns has been analyzed by Lo and MacKinley in [1], Jokivuolle in [2] and Stoll and Whaley in [3]. They argue that positive autocorrelation shows up in index returns due to presence of stale prices of stocks included into the index. Above mentioned happens when the increase in the number of stocks comes from inclusion of small capitalization stocks, which are known to trade less frequently than large ones. Due to infrequent trading in small capitalization stocks the observed index value do not reflect the true market value of the underlying stock portfolio as the index value is calculated using the last observed stock transaction prices.

Conrad and Kaul in [4] avoiding the nonsynchronos trading problem analyze autocorrelation of Wednesday-Wednesday returns for size grouped portfolios and find first order autocorrelation of weekly returns varying between 0.09 to 0.30. For longer time periods Fama and French in [5] find that autocorrelation of returns of diversified portfolios of NYSE stocks becomes strongly negative.

The evidence of serial autocorrelation in stock and stock index returns contradicts assumptions made in widely accepted stock return model used by Black and Sholes in [6] and Merton in [7] to derive call option pricing formula. They assume that asset returns are distributed independently of each other.

1.2 Convergence of stochastic difference equations to stochastic differential equations

In this section we would like to present general conditions for a sequence of finite-dimensional discrete time Markov processes $\{{}_hX_t\}_{h\downarrow 0}$ to converge weakly to an Ito process. These are drawn from Nelson [10].

The formal set-up is as follows: Let $D([0,\infty), R^n)$ be the space of mappings from $[0,\infty)$ into R^n that are continuous from the right with finite left limits, and let $B(R^n)$ denote the Borel sets on R^n. D is a metric space when endowed with Skorohod metric. For each $h > 0$, let M_{kh} be the σ-algebra generated by kh, ${}_hX_0$, ${}_hX_h$, ${}_hX_{2h}$, ..., ${}_hX_{kh}$, and let ν_h be a probability measure on $(R^n, B(R^n))$. For each $h > 0$ and each $k = 0, 1, 2, 3, ...$, let $\Pi_{h,kh}(x, \cdot)$ be a transition function on R^n, i.e.

(a) $\Pi_{h,kh}(x, \cdot)$ is a probability measure on $(R^n, B(R^n))$ for all $x \in R^n$,

(b) $\Pi_{h,kh}(\cdot, \Gamma)$ is $B(R^n)$ measurable for all $\Gamma \in B(R^n)$.

For each $h > 0$, let P_h be the probability measure on $D([0,\infty), R^n)$ such that

$$P_h[{}_hX_0 \in \Gamma] = \nu_h(\Gamma) \tag{1}$$

for any $\Gamma \in B(R^n)$,

$$P_h[{}_hX_t =_h X_{kh}, kh \le t < (k+1)h] = 1 \tag{2}$$

and

$$P_h[{}_hX_{(k+1)h} \in \Gamma \mid M_{kh}] = \Pi_{h,kh}({}_hXkh, \Gamma) \tag{3}$$

almost surely under P_h for all $k \ge 0$ and $\Gamma \in B(R^n)$.

For each $h > 0$, (1) specifies the distribution of the random starting point and (3) the transition probabilities of n-dimentional discrete time markov process ${}_hX_{kh}$. We form the continuous time process ${}_hX_t$ from the discrete time process ${}_hX_{kh}$ by (2), making ${}_hX_t$ a step function with jumps at times h, $2h$, $3h$ and so on.

Now if for each $h > 0$ and each $\varepsilon > 0$ we define

$$a_h(x,t) \equiv h^{-1} \int_{\|y-x\|\le 1} (y-x)(y-x)'\Pi_{h,h[t/h]}(x, dy), \tag{4}$$

$$b_h(x,t) \equiv h^{-1} \int_{\|y-x\|\le 1} (y-x)\Pi_{h,h[t/h]}(x, dy), \tag{5}$$

$$\Delta_{h,\varepsilon}(x,t) \equiv h^{-1} \int_{\|y-x\|\leq\varepsilon} \Pi_{h,h[t/h]}(x,dy), \tag{6}$$

where $[t/h]$ is the integer part of t/h, i.e. the largest integer $k \leq t/h$, it is possible under some assumtions which are in detail specified by Nelson in [10] to prove the following theorem:

Theorem 1.1 *Under some assumptions, the sequence of* ${}_hX_t$ *process defined by* (1)-(3) *converges weakly (i.e. in distribution) as* $h \downarrow 0$ *to the* X_t *process defined by the stochastic integral equation*

$$X_t = X_0 + \int_0^t b(X_s,s)ds + \int_0^t \sigma(X_s,s)dW_{n,s}, \tag{7}$$

where $W_{n,t}$ *is an n-dimentional standard Brownian motion, independent of* X_0*, and where for any* $\Gamma \in B(R^n)$*,* $P(X_0 \in \Gamma) = \nu(\Gamma)$*. Such an* X_t *process exists and is distributionally unique. This distribution does not depend on the choice of* $\sigma(\cdot,\cdot)$*. Finally,* X_t *remains finite in finite time intervals almost surely, i.e. for all* $T > 0$*,*

$$P[\sup_{0\leq t\leq T} \|X_t\| < \infty] = 1. \tag{8}$$

Carkovs in [8] follows a similar approach as Nelson [10] and analysis a discrete Markov dynamic system given in a form of stochastic difference equation

$$x_{t+1} = x_t + \varepsilon f_1(x_t, y_{t+1}) + \varepsilon^2 f_2(x_t, y_{t+1}) \tag{9}$$

where $\{y_t\}$ is an ergodic Markov chain with transition probability $p(y,dz)$, invariant measure μ and potential operator Π. Using interpolation

$$s \in [t\varepsilon^2, (t+1)\varepsilon^2] \tag{10}$$

and

$$X_{\varepsilon^2}(s) := (x_{t+1} - x_t)(s\varepsilon^{-2} - t) + x_t \tag{11}$$

Carkovs in his paper [8] is able to prove that for any $\{t_1, t_2, ..., t_n\}$ distribution of vector $\{X_{\varepsilon^2}(t_1), X_{\varepsilon^2}(t_2), ..., X_{\varepsilon^2}(t_n)\}$ for sufficiently small ε^2 may be approximated by distribution of vector $\{X(t_1), X(t_2), ..., X(t_n)\}$ defined by solution of stochastic Ito differential equation

$$dX(s) = a(X(s))ds + \sigma(X(s))dw(s), \tag{12}$$

where

$$a := \bar{f}_2 + [\mathcal{P}\Pi f_1']f_1, \tag{13}$$

$$\sigma^2 := \bar{f}_1^2 + 2\overline{f_1\mathcal{P}\Pi f_1}, \tag{14}$$

$$\mathcal{P}f(y) := \int_Y f(z)p(y,dz), \tag{15}$$

$$\bar{f} := \int_Y f(z)\mu(dz). \tag{16}$$

1.3 The Black-Scholes Option Pricing Formula

The development of option pricing models in [6] and [7] is based on existence of a dynamic investment strategy involving the underlying asset and risk free bonds that exactly replicates payoff of the option. In case when stock price $S(t)$ follows log-normal diffusion process

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t), \tag{17}$$

where σ is the diffusion coefficient, μ - the drift coefficient and $W(t)$ - a standard Wiener process. It is assumed that trading is frictionless and continuous. Then the no-arbitrage condition yields the following differential equation on the call price $C(t)$

$$\frac{1}{2}\sigma^2 S^2(t)\frac{\partial^2 C(t)}{\partial S^2(t)} + \mu S(t)\frac{\partial C(t)}{\partial S(t)} + \frac{\partial C(t)}{\partial t} = \mu C(t), \tag{18}$$

where μ is the instantaneous risk-free rate of return. Given the two boundary conditions for the European call option

$$C(S(T), T) = \max(S(T) - K, 0), \tag{19}$$

$$C(0, t) = 0, \tag{20}$$

there exists a unique solution to the partial differential equation (18) and is called Black-Sholes formula

$$C_{BS}(S(t), t) = S(t)N(d_1) - K\exp(-\mu(T-t))N(d_2), \tag{21}$$

where

$$d_1 \equiv \frac{\log(S(t)/K) + (\mu + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}, \tag{22}$$

and

$$d_2 \equiv d_1 - \sigma\sqrt{T-t}, \tag{23}$$

where $N()$ is the standard normal cumulative distribution function.

The Black-Sholes formula (21) does not depend on drift μ, but may be an arbitrary function of $S(t)$ and other economical variables. This feature implies that the Black-Sholes formula is applicable to a different asset return processes and could reflect complex patterns of predictability and dependence on other observed and unobserved economic factors.

1.4 The Greeks

The Greeks are vital tools in risk management. Each Greek measures the sensitivity of the value of a portfolio to a small change in a given underlying parameter, so that component risks may be treated in isolation, and the portfolio rebalanced accordingly to achieve a desired exposure.

The Greeks in the BlackScholes model are relatively easy to calculate, a desirable property of financial models, and are very useful for derivatives traders, especially those who seek to

hedge their portfolios from adverse changes in market conditions. For this reason, those Greeks which are particularly useful for hedging delta, gamma and vega are well-defined for measuring changes in Price, Time and Volatility. Although μ is a primary input into the BlackScholes model, the overall impact on the value of an option corresponding to changes in the risk-free interest rate is generally insignificant and therefore higher-order derivatives involving the risk-free interest rate are not common.

The most common of the Greeks are the first order derivatives: Delta, Vega, Theta and Rho as well as Gamma, a second-order derivative of the value function. The above mentioned sensitivities are common enough that they have common names, and we will list explicit formulas as they have been derived for a European call $C(S(t), t)$ by Haug in [9].

Delta, Δ, measures the rate of change of option value with respect to changes in the underlying asset's price. Delta is the first derivative of the value C of the option with respect to the underlying instrument's price S.

$$\Delta(S(t), t) \equiv \frac{\partial C}{\partial S} = N(d_1) \tag{24}$$

Theta,Θ, measures the sensitivity of the value of the derivative to the passage of time t.

$$\Theta(S(t), t) \equiv \frac{\partial C}{\partial t} = -\frac{S(t)N(d_1)\sigma}{2\sqrt{T-t}} - \mu K \exp(-\mu(T-t))N(d_2) \tag{25}$$

Vega, ν measures sensitivity to volatility σ. Vega is the derivative of the option value with respect to the volatility of the underlying.

$$\nu(S(t), t) \equiv \frac{\partial C}{\partial \sigma} = S(t)N(d_1)\sqrt{T-t} \tag{26}$$

Rho, $\mathcal{R}$, measures sensitivity to the applicable interest rate. Rho is the derivative of the option value with respect to the risk free rate. Except under extreme circumstances, the value of an option is least sensitive to changes in the risk-free-interest rates. For this reason, rho is the least used of the first-order Greeks.

$$\mathcal{R}(S(t), t) \equiv \frac{\partial C}{\partial \mu} = K(T-t)\exp(-\mu(T-t))N(d_2) \tag{27}$$

Gamma, Γ, measures the rate of change in the delta with respect to changes in the underlying asset price. Gamma is the second derivative of the value function with respect to the underlying price. Gamma is important because it corrects for the convexity of value. When a trader seeks to establish an effective delta-hedge for a portfolio, the trader may also seek to neutralize the portfolio's gamma, as this will ensure that the hedge will be effective over a wider range of underlying price movements.

$$\gamma(S(t), t) \equiv \frac{\partial \Delta}{\partial S} = \frac{N(d_1)}{S(t)\sigma\sqrt{T-t}} \tag{28}$$

2 Derivation of a Formula for Serially Correlated Stock Return Process

The simplest mathematical model describing development of stock's price S_t and involving assumption of serial autocorrelation in stock's returns under commonly used condition on risk neutrality of probabilistic measure $\mathbb{P}$ may be written in the following way

$$S_{t+1} = S_t(1 + \varepsilon^2\mu + \varepsilon\sigma y_{t+1}), \tag{29}$$

where y_t is a Gaussian random sequence with zero mean and unit variance. When it is considered that these random numbers are independent we may write that y_t follows AR(1):

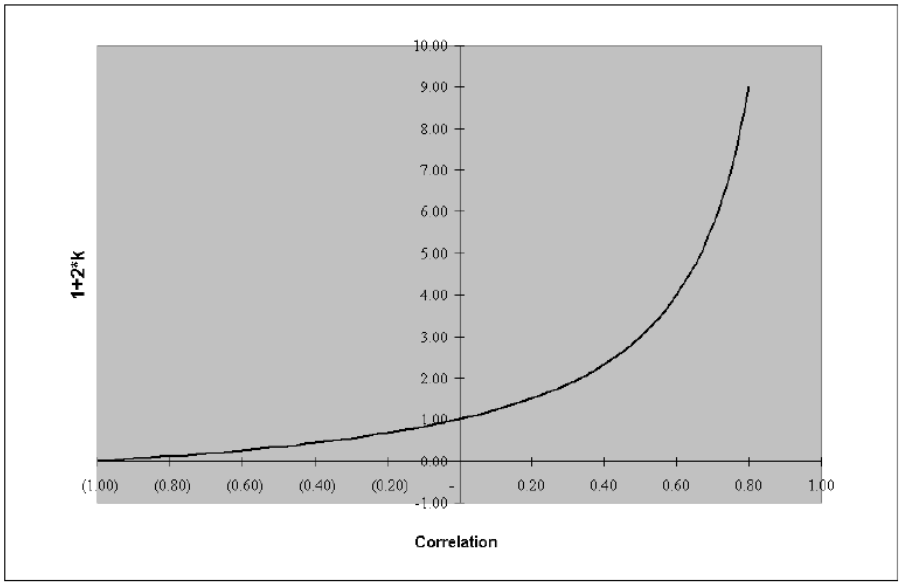


Figure 1: Ratio of $\frac{\sigma^2_{eff}}{\sigma^2}$ as a function of autocorrelation coefficient ρ

$$y_{t+1} = \rho y_t + \sqrt{1-\rho^2}\xi_{t+1}, \tag{30}$$

where $\{\xi_t\}$ is i.i.d. Gaussian sequence, $\mathbb{E}\xi_t = 0$, $\mathbb{E}\xi_t^2 = 1$.

To be able use formulas (12)-(15) derived by Carkovs in [8] we denote $x_t \equiv S_t$ and rewrite equation (29) in the following form

$$x_{t+1} = x_t + \varepsilon\sigma y_{t+1}x_t + \varepsilon^2\mu x_t \tag{31}$$

and now we can use (12) with

$$f_1(x_t, y_{t+1}) = \sigma y_{t+1}x_t \tag{32}$$

and

$$f_2(x_t, y_{t+1}) = \mu x_t. \tag{33}$$

After calculation of (13) and (14) we derive continuous time approximation of stochastic difference equation (29) in a form of diffusion process satisfying stochastic Ito differential equation

$$dS(t) = S(t)(\mu + \sigma^2 k)dt + S(t)\sqrt{1+2k}\sigma dw(t), \tag{34}$$

where

$$k := \sum_{m=1}^{\infty} Corr\{y_{t+m}, y_t\} = \frac{\rho}{1-\rho}. \tag{35}$$

Here we can introduce σ_{eff}^2 as an effective volatility of the diffusion process (34)

$$\sigma_{eff}^2 = \sigma^2(1+2k). \tag{36}$$

From Figure 1 we can observe that this effective volatility can be substantially greater then σ^2 if serial correlation is positive and converges to 0 as correlation approaches -1.

After substitution of (35) into (34) we get the final equation

$$dS(t) = S(t)(\mu + \sigma^2\frac{\rho}{1-\rho})dt + S(t)\sqrt{\frac{1+\rho}{1-\rho}}\sigma dw(t). \tag{37}$$

3 Option Pricing on Stocks with Autocorrelated Returns

Now let's derive European call option pricing formulas if underlying stock's price process $S(t)$ satisfies the stochastic differential equation (34). The boundary conditions for the European call option is given by (19) and (20). Using well known techniques we get the following results

$$C(S(t),t) = S(t)N(d_1) - K\exp(-(\mu+\sigma^2 k)(T-t))N(d_2), \tag{38}$$

where

$$d_1 = \frac{\log(S(t)/K) + (\mu + \sigma^2 k + \frac{1}{2}\sigma^2(1+2k))(T-t)}{\sigma\sqrt{(1+2k)(T-t)}}, \tag{39}$$

and

$$d_2 = d_1 - \sigma\sqrt{(1+2k)(T-t)}, \tag{40}$$

where $N()$ is the standard normal cumulative distribution function.

Now we are ready to derive formulas used to calculate sensitivities of call option price to changes in underlying parameters.

Delta, Δ, the first derivative of the value C of the option with respect to the underlying instrument's price S will have the same form as in (24):

$$\Delta(S(t),t) = \frac{\partial C}{\partial S} = N(d_1). \tag{41}$$

Theta,Θ, the sensitivity of the value of the derivative to the passage of time t now will have the following form

$$\Theta(S(t),t) = \frac{\partial C}{\partial t} = -\frac{S(t)N(d_1)\sigma\sqrt{1+2k}}{2\sqrt{T-t}} - (\mu+\sigma^2 k)K\exp(-(\mu+\sigma^2 k)(T-t))N(d_2) \tag{42}$$

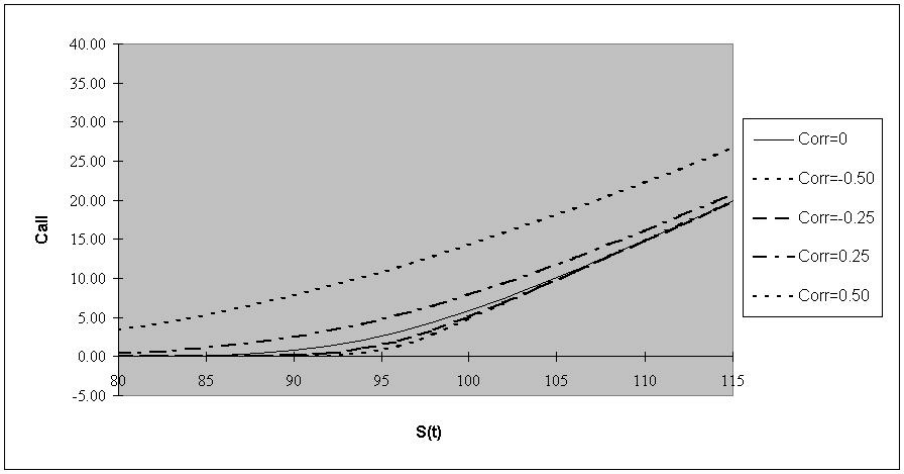


Figure 2: Value of Call Option for Different Correlation Coefficients ρ

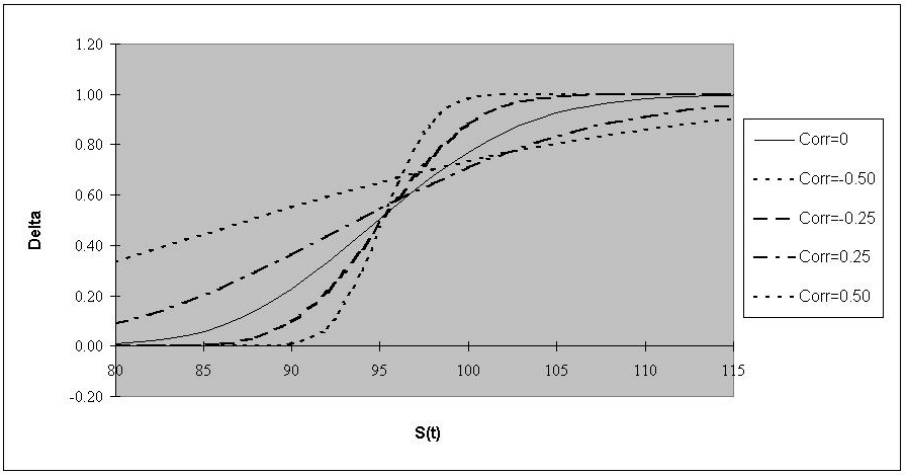


Figure 3: Value of Call Option's Delta for Different Correlation Coefficients ρ

Vega, ν, the sensitivity to volatility σ will be

$$\nu(S(t),t) = \frac{\partial C}{\partial \sigma} = S(t)N(d_1)\sqrt{(1+2k)(T-t)} \tag{43}$$

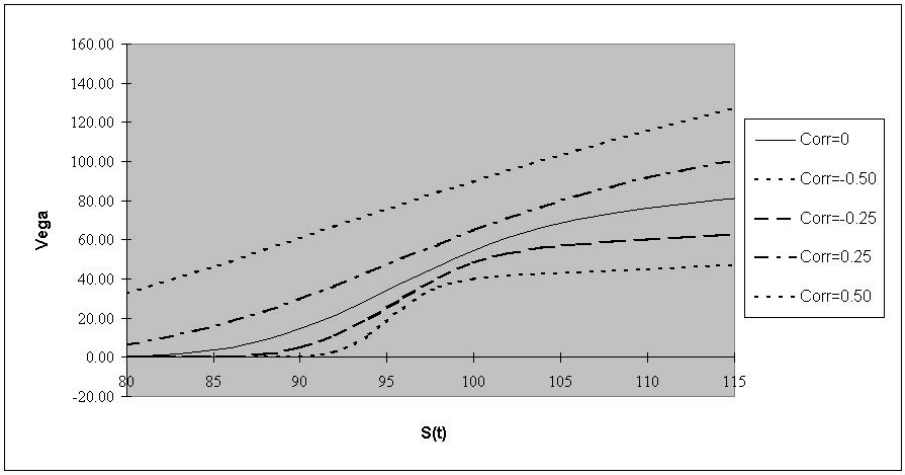


Figure 4: Value of Call Option's Vega for Different Correlation Coefficients ρ

Rho, $\mathcal{R}$, the sensitivity to the applicable interest rate

$$\mathcal{R}(S(t),t) \equiv \frac{\partial C}{\partial \mu} = K(T-t)\exp(-(\mu+\sigma^2 k)(T-t))N(d_2) \tag{44}$$

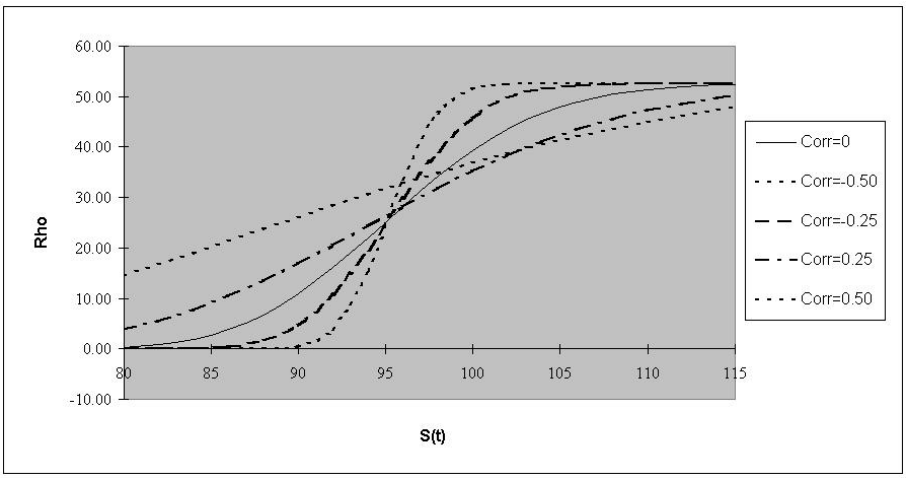


Figure 5: Value of Call Option's Rho for Different Correlation Coefficients ρ

Gamma, Γ, that measures the rate of change in the delta with respect to changes in the underlying asset price will be

$$\gamma(S(t),t) = \frac{\partial \Delta}{\partial S} = \frac{N(d_1)}{S(t)\sigma\sqrt{(1+2k)(T-t)}} \tag{45}$$

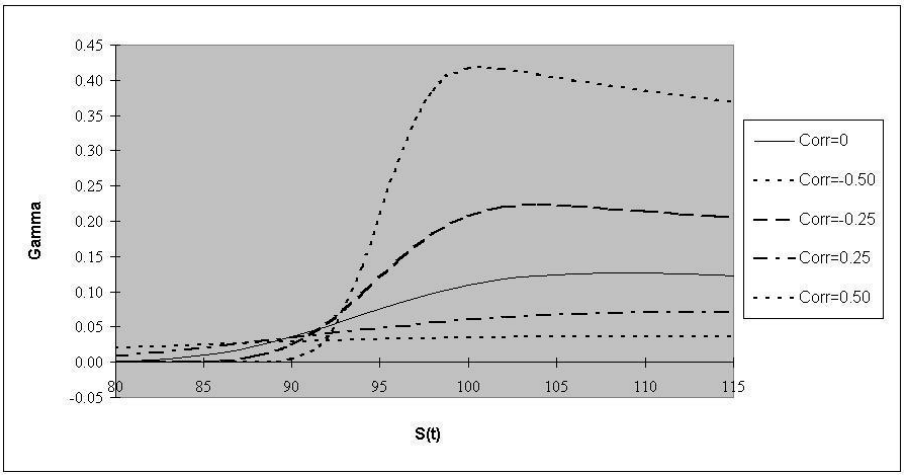


Figure 6: Value of Call Option's Gamma for Different Correlation Coefficients ρ

4 Conclusions and Further Research

In our paper we have derived an analytical formulas for calculation of the value of European call option and its sensitivities to underlying parameters. The formulas demonstrate an important relationship between the value of the option, its risk parameters and underlying asset return autocorrelation coefficient. We have been able to show that autocorrelation can have substantial impact on obtained results and it should be considered when someone tries to estimate correct historical volatility of the underlying return process. In a case of no autocorrelation, our result reduces to a well known Black-Sholes option pricing formula.

The approach used in our paper can be further applied to discrete time stochastic difference equation systems where volatility is stochastic or it is modeled by generalized autoregressive conditional heteroscedasticity process.

Another topic which deserves further analysis is how the convergence of discrete time stochastic difference equation to its continuous time approximation depends on autocorrelation coefficient.

References

[1] LO, A., MACKINLEY, A. C.: *An Econometric Analysis of Non-Synchronous Trading.* Journal of Econometrics, No. 45, pp. 181-211, 1990.

[2] JOKIVUOLLE, E.: *Measuring True Stock Index Value in the Presence of Infrequent Trading.* Journal of Financial and Quantitative Analysis, September, 1995.

[3] STOLL, H. R., WHALEY, R. E.: *The Dynamics of Stock Index and Stock Index Futures Returns.* Journal of Financial and Quantitative Analysis, No. 25, pp. 441-468, 1990.

[4] COUNRAD, J., KAUL, G.: *Time-variation in Expected Returns.* Journal of Business, No. 61, pp. 409-425, 1988.

[5] FAMA, E. F., FRENCH, K. R.: *Permanent and Temporary Components of Stock Market Prices*. Journal of Political Economy, No. 96, pp. 246-273, 1988.

[6] BLACK, F., SCHOLES, M.: *The pricing of options and corporate liabilities*. Journal of Political Economy, No. 81, pp. 637-659, 1973.

[7] MERTON, R.: *Theory of Rational Option Pricing* . Bell Journal of Economics, No. 4, pp. 141-183, 1973.

[8] CARKOVS, J.: *On Diffusion Approximation of Discrete Markov Dynamical Systems*. In Computational Geometry, Proceedings of World Academy of Science, Engineering and Technology, Vol. 30, July, pp. 1-6, 2008.

[9] HAUG, E. D.: *The Complete Guide to Option Pricing Formulas*. McGraw-Hill Professional, 2007.

[10] NELSON, D. B.: *ARCH models as diffusion approximations*. Journal of Econometrics, Vol. 45, Issues 1-2, Pages 7-38, July-August 1990

Current address

Aigars Egle, PhD st.
Riga Technical University, Kaļķu 1, LV-1050, Riga, LATVIA, phone: +371 67010847,
e-mail: aigars.egle@citadele.lv

Journal of Applied Mathematics

volume IV (2011), number III

LATVIAN GDP: THE OPTIMAL TIME SERIES FORECASTING ALGORITHM

BEZRUCKO Aleksandrs, (LV)

Abstract: In this work an algorithm is developed for finding optimal time series model for GDP forecasting. Latvian GDP data with quarterly observation frequency is taken as time series. ARMA Analysis of Latvian GDP time series is performed. The set of model has been constructed. In order to check the accuracy of models, different residual tests are performed: autocorrelation, heteroscedasticity and normality of residual distribution. Models are compared in their forecast quality.

Keywords: *time series, Gross Domestic Product, ARMA (Autoregressive Moving Average) Analysis, Residual tests, Serial Correlation, Heteroskedasticity*

1 Introduction

The analysis and forecast of GDP for any time and any country is important task for economists, policy makers and entrepreneurs. These researches are consisting of many objective and subjective factors. In econometrics forecast, not only statistical methods are used, but a lot of economical and political events must be taken into account.

Working on the paper, different methods of econometrical modelling have been analyzed. For example, analysis methods for German GDP forecast that are described by Lutkepohl in "Applied Time Series Analysis" [1]. Lutkepohl described different ways of ARMA and Residual analysis of time series. In this paper author uses familiar methods of statistical analysis of time series for forecasting Latvian GDP. Computer software enabled the author to perform the search for the best models for certain time series. Based on the analysis of these models, a search algorithm of optimal model is created.

In order to find an optimal model of forecasting Latvian Gross Domestic Product, two different cases of Latvian GDP series with quarterly observation frequency are taken. The first case is Latvian quarterly GDP series in levels (Latvian lats) and second case is the same data in percentage growth. The GDP series are given in Figure 1. The time series length is T = 57. The time series is taken from the first quarter of year 1996 till the first quarter of year 2009. All searches and forecasts are made using econometrical software EViews 6.0.

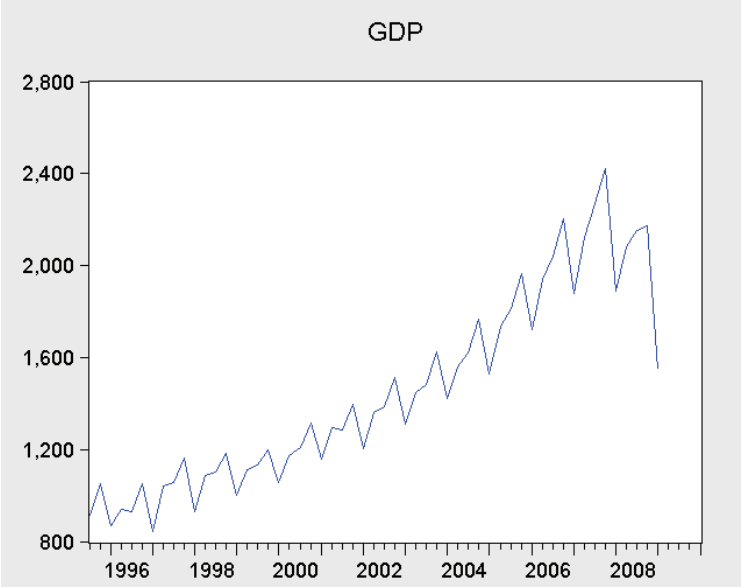


Figure 1 Latvian GDP (Lats) 1995Q1-2009Q1

2 Analysis description

2.1. The analysis of criteria

At the first stage of choosing the best model, 3 criteria are analyzed: Akaike info, Schwarz and Hannan-Quinn. The best model has minimal values. R-squared statistic is also present. At this stage models with best criteria are taken. ARMA Analysis is done in EViews program language and statistical criteria represent the result of the program (Fig.2).

Table: CRITS Workfile: GDP_AUTOREGRESSION_1::Untitled\

View | Proc | Object | Print | Name | Edit+/- | CellFmt | Grid+/- | Title | Comments+/-

	A	B	C	D	E
1	AR / MA	0.000000	1.000000	2.000000	3.000000
2	0.000000	0.347646	-0.239621	-1.415521	-1.446306
3	1.000000	-1.378436	-1.841098	-2.093701	-2.105267
4	2.000000	-1.662691	-2.019881	-2.076773	-2.204613
5					

Figure 1 ARMA Analysis in EViews 6.0

The R-squared (R^2) statistic measures the success of the regression in predicting the values of the dependent variable within the sample. In standard settings, R^2 may be interpreted as the fraction of the variance of the dependent variable explained by the independent variables. The statistic will equal one if the regression fits perfectly, and zero if it fits no better than the simple mean of the dependent variable. It can be negative for a number of reasons. For example, if the regression does not have an intercept or constant, if the regression contains coefficient restrictions, or if the estimation method is two-stage least squares or ARCH.
The Akaike Information Criterion (AIC) is computed as: $AIC = -2l/T + 2k/T$ where l is the log likelihood. The AIC is often used in model selection for non-nested alternatives-smaller values of

the AIC are preferred. For example, you can choose the length of a lag distribution by choosing the specification with the lowest value of the AIC.
The Schwarz Criterion (SC) is an alternative to the AIC that imposes a larger penalty for additional coefficients: $SC = -2l/T + (k\log T)/T$

2.2. Residual tests

The second stage is represented by Residual tests: Serial Correlation LM test, Histogram – Normality test, Hetereskedasticity ARCH test and Correlogram Square Residual test. Models have passed the test if P Value is higher than 0,1.
Serial Correlation LM test is an alternative to the Q-statistics for testing serial correlation. The test belongs to the class of asymptotic (large sample) tests known as Lagrange multiplier (LM) tests Serial Correlation LM test has the higher importance because on this step we are concerning with the possibility that our errors exhibit autocorrelation. LM test check for higher order ARMA errors and is applicable whether or not there are lagged dependent variables.
The null hypothesis of the LM test is that there is no serial correlation up to lag order p, where p is a pre-specified integer. The local alternative is ARMA(r, q) errors, where the number of lag terms p=max(r, q). Note that this alternative includes both AR(p) and MA(p) error processes, so that the test may have power against a variety of alternative autocorrelation structures.
The test tatistic is computed by an auxiliary regression as follows. First, suppose you have estimated the regression;
$y_t = X_t\beta + e_t$
where b are the estimated coefficients and e are the errors. The test statistic for lag order p is based on the auxiliary regression for the residuals $e = y - X\hat{\beta}$:

$$e_t = X_t\gamma + \left(\sum_{s=1}^{p} \alpha_s e_{t-s}\right) + v_t .$$

Histogram and normality tests are displays a histogram and descriptive statistics of the residuals, including the Jarque-Bera statistic for testing normality. If the residuals are normally distributed, the histogram should be bell-shaped and the Jarque-Bera statistic should not be significant. The Jarque-Bera statistic has a χ^2 distribution with two degrees of freedom under the null hypothesis of normally distributed errors. [2]
The ARCH test is a Lagrange multiplier (LM) test for autoregressive conditional heteroskedasticity (ARCH) in the residuals. This particular heteroskedasticity specification was motivated by the observation that in many financial time series, the magnitude of residuals appeared to be related to the magnitude of recent residuals. ARCH in itself does not invalidate standard LS inference. However, ignoring ARCH effects may result in loss of efficiency.
The ARCH LM test statistic is computed from an auxiliary test regression. To test the null hypothesis that there is no ARCH up to order q in the residuals, we run the regression:

$$e_t^2 = \beta_0 + \left(\sum_{s=1}^{q} \beta_s e_{t-s}^2\right) + v_t ,$$

where e is the residual. This is a regression of the squared residuals on a constant and lagged squared residuals up to order q. The F-statistic is an omitted variable test for the joint significance

of all lagged squared residuals. The Obs*R-squared statistic is Engle's LM test statistic, computed as the number of observations times the R^2 from the test regression. The exact finite sample distribution of the F-statistic under H_0 is not known, but the LM test statistic is asymptotically distributed as a $\chi^2(q)$ under quite general conditions.
Correlogram of squared residuals test displays the autocorrelations and partial autocorrelations of the squared residuals up to any specified number of lags and computes the Ljung-Box Q-statistics for the corresponding lags. The correlograms of the squared residuals can be used to check autoregressive conditional heteroskedasticity (ARCH) in the residuals.
If there is no ARCH in the residuals, the autocorrelations and partial autocorrelations should be zero at all lags and the Q-statistics should not be significant inclusion of ARMA terms. [2]

2.3. Out-Of-Samle Forecasting

The final evaluation test is “Out-Of-Sample Forecasting”. At this stage forecasts are compared to real data that we have for the period of the last 3 quarters of 2009.

3. Latvian GDP in levels

The analyzed series consist of seasonally unadjusted Latvian quarterly GDP in levels for the period of 1995Q1 – 2009Q1. It is represented in Figure 1. Constructing a model for the logs is more advantageous because the changes in the log series display a more stable variance than the changes in the original series. Time series in logs are shown in Figure 3.

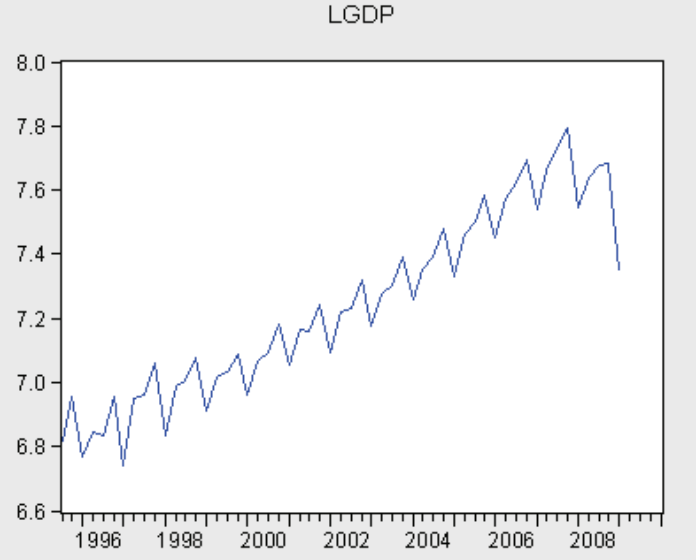


Figure 2 Latvian GDP in logs

Table 1

The analysis of criteria (Levels)

Nr.	LGDP	Akaike	Schwarz	Han-Quinn
1	Ar(1)	-1.326483	-1.290316	-1.312461
2	Ar(2)	-1.487976	-1.451479	-1.473862
3	Ma(1) C	-0.182363	-0.110677	-0.154503
4	Ma(2) C	-0.143123	-0.071437	-0.115263
5	AR(1) MA(1)	-1.878947	-1.806613	-1.850903
6	AR(1) AR(2) MA(1) MA(2)	-2.087592	-1.941604	-2.031137
7	AR(1) AR(2) SAR(4) MA(1)	-3.746621	-3.595105	-3.688722
8	AR(1) AR(2) SAR(4) MA(4)	-4.18207	-4.030555	-4.124172
9	Trend C	-1.914989	-1.843303	-1.887129
10	AR(1) Trend	-1.294331	-1.221997	-1.266287
11	AR(2) SAR(4)	-3.027406	-2.951648	-2.998456
12	AR(1) SAR(4) MA(4)	-3.646746	-3.534175	-3.603589
13	AR(1) AR(2) SAR(4) MA(4) SEAS(1)	-4.224953	-4.035559	-4.15258
14	AR(1) SAR(4) MA(4) D1997Q2	-4.047836	-3.89774	-3.990293

Best models: Nr. 5,7,8,12,13,14. Other models are excluded from further evaluation process. The residual test results are given in Table 2. Models Nr.13 and Nr.14 have undergone all tests. Residuals graph of Model Nr. 13 is given in Figure 4.

Table 2

Residual Test (Levels)

Nr.	Serial Correlation	Histogram	Heteroskedasticity
5	0.69290	0.00000	0.61360
7	0.69350	0.00000	0.79830
8	0.08070	0.40000	0.53920
12	0.61610	0.00003	0.84840
13	0.09250	0.83015	0.87060
14	0.10670	0.40223	0.85250

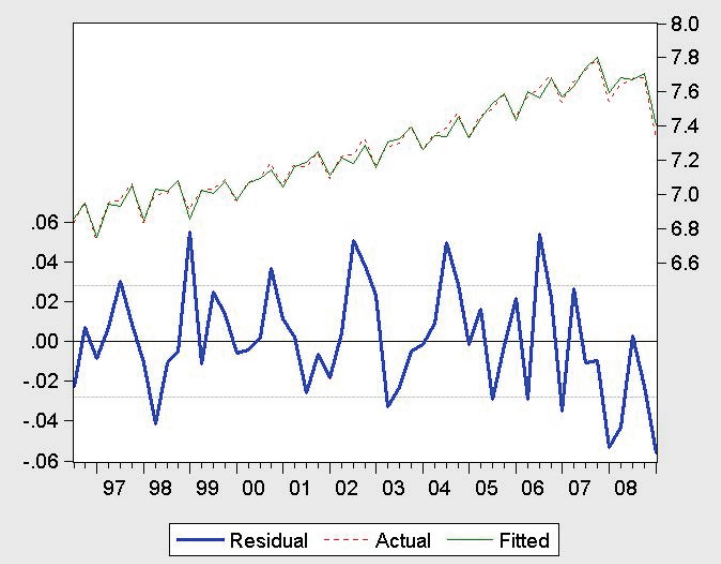


Figure 3 Residuals graph for Model Nr.13

The worst model residuals are given in Figure 5 for comparison.

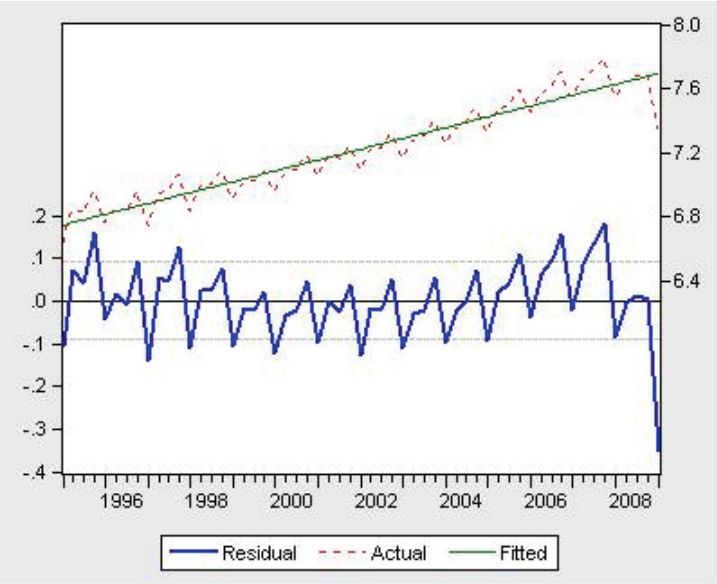


Figure 4 Residuals graph for Model Nr.9

"Out-Of-Sample Forecasting" test (Table 3) shows that the most real forecasts are gained from models Nr.12: AR(1) SAR(4) MA(4). Absolute difference (0.037) is minimal in this case. The best model has passed all Residual tests except Normality. The second and the third results are shown by models Nr. 7 and Nr.13., which have passed almost all residual tests except Nr.7., which also did not pass the Normality test.

Table 3

Out-Of-Sample Forecasting (Levels)

	Forecast - 1 Step - 09q2			Forecast - 2 Step -09q3			Forecast - 3 Step -09q4			
Nr.	LGDP real	LGDPf	Diff	LGDP real	LGDPf	Diff	LGDP real	LGDPf	Diff	| Sum % |
5	0.088	0.435	**-0.347**	0.031	0.022	**0.008**	0.034	0.018	**0.016**	0.371
7	0.088	0.082	**0.007**	0.031	0.017	**0.014**	0.034	-0.012	**0.045**	0.066
8	0.088	0.044	**0.044**	0.031	0.023	**0.007**	0.034	-0.020	**0.054**	0.106
12	0.088	0.103	**-0.014**	0.031	0.042	**-0.012**	0.034	0.023	**0.011**	0.037
13	0.088	0.057	**0.031**	0.031	0.017	**0.014**	0.034	-0.021	**0.055**	0.099
14	0.088	0.050	**0.039**	0.031	0.013	**0.018**	0.034	-0.019	**0.053**	0.109

LATVIAN GDP IN PERCENTAGE GROWTH

All evaluations are made with Latvian GDP. The differences and the log difference of time series are shown in Figure 6. The log difference displays a more stable variance than the changes in the original series.

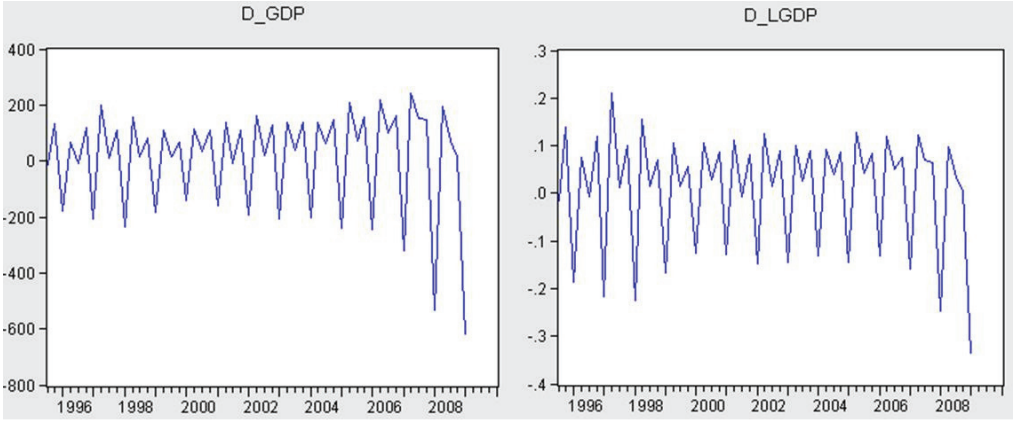


Figure 5 Time series in differences and in log differences

Table 4

The analysis of criteria (Difference)

Nr.	d (LGDP)	Akaike	Schwarz	Han-Quinn
1	Ar(1)	-1.659793	-1.623296	-1.645679
2	Ar(2)	-1.379102	-1.342269	-1.364897
3	Ma(1) C	-1.876758	-1.804424	-1.848714
4	Ma(2) C	-1.309161	-1.236827	-1.281117
5	AR(1) MA(1)	-1.712348	-1.639354	-1.684121
6	AR(1) AR(2) MA(1) MA(2)	-2.033619	-1.886286	-1.976798
7	AR(1) AR(2) SAR(4) MA(1)	-3.728444	-3.575482	-3.670195
8	AR(1) AR(2) SAR(4) MA(4)	-4.197451	-4.044489	-4.139202
9	Trend C	-1.306934	-1.2346	-1.278891
10	AR(1) Trend	-1.636801	-1.563807	-1.608574
11	AR(2) SAR(4)	-3.803588	-3.727107	-3.774463
12	AR(1) SAR(4) MA(4)	-4.191773	-4.078136	-4.148349
13	AR(1) AR(2) SAR(4) MA(4) SEAS(1)	-4.307962	-4.116759	-4.235151
14	AR(1) SAR(4) MA(4) D1997Q2	-4.273152	-4.121636	-4.215253

Models with the best criteria: Nr. 7, 8, 11, 12, 13, 14. Other models are excluded from further evaluation process.

Table 5

Residual Test (Difference)

Nr.	Serial Correlation	Histogram	Heteroskedasticity: ARCH
7	0.6938	0	0.825
8	0.6895	0.393373	0.3453
11	0.8852	0	0.017538
12	0.0503	0.421976	0.6305
13	0.64	0.560093	0.2823
14	0.5554	0.997972	0.7499

Models Nr. 8, 13, 14 have completed all tests. Model Nr. 12 also has good statistic. "Out-Of-Sample" forecasting test (Table 6) shows that the most real forecasts are gained from models Nr.11: AR(2) SAR(4). Absolute difference (0.037) is minimal in this case. This model did not complete the histogram test, but passed all other residual tests. The second result belongs to model Nr.7, which has the same problem with Normality test. Third result is shown by model Nr.14 – this model passed all residual tests.

Table 6

Out-Of-Sample Forecasting (Difference)

	Forecast - 1 Step - 09q2			Forecast - 2 Step -09q3			Forecast - 3 Step -09q4			
Nr.	dLGDP real	dLGDPf	Diff	dLGDP real	dLGDPf	Diff	LGDP real	LGDPf	Diff	| Sum % |
7	0.088	0.084	**0.004**	0.031	0.022	**0.009**	0.034	0.004	**0.030**	0.043
8	0.088	0.034	**0.054**	0.031	0.019	**0.012**	0.034	-0.020	**0.054**	0.120
11	0.088	0.091	**-0.003**	0.031	0.024	**0.007**	0.034	0.007	**0.027**	0.037
12	0.088	0.045	**0.043**	0.031	0.024	**0.007**	0.034	-0.020	**0.054**	0.104
13	0.088	0.022	**0.067**	0.031	0.009	**0.022**	0.034	-0.021	**0.054**	0.143
14	0.088	0.070	**0.018**	0.031	0.011	**0.020**	0.034	-0.020	**0.054**	0.091

4 The search alorithm

The search algorithm is shown in Figure 7. Step-by-step description looks as follows:
Input data: GDP Time series
1. Construction of ARMA models in levels and in differences separately. Starting from this point the model is divided in two branches and the subsequent steps are carried out in parallel for levels and for differences.
2. ARMA Analysis.
3. Performing Residual tests.
4. If during residual tests probability value is less than 10%, the model is excluded from further evaluation. This is not a reliable model.
5. If P Value is higher than 10%, the forecast for specified periods of time is performed.
6. Comparing forecast data with real data. Making an evaluation.
7. The analysis of the results. Two branches of models come back to one point.
Output data: Best model for the GDP Forecasts.

Conclusion

In the given paper the author described search algorithm of optimal time series. With a help of statistical modelling the econometric analysis of Latvian GDP is done. Different cases of constructing model are made in Latvian lats (in levels) and in percentage growth (in difference).

Comparison of 1 and 2 cases shows that case in levels and in differences gave approximately the same result – 3.7% deviation from real data in absolute value for forecasts for 3 steps in future. It seems that it is does not matter which way to use. But this is surely not the right path. It is very important to analyze all the results and understand how they are calculated and evaluated. If a model gives the best forecast for one, two or three steps separately – it is does not mean that this model will be best in other cases. Figure 7 shows the algorithm of searching for optimal forecasting model.

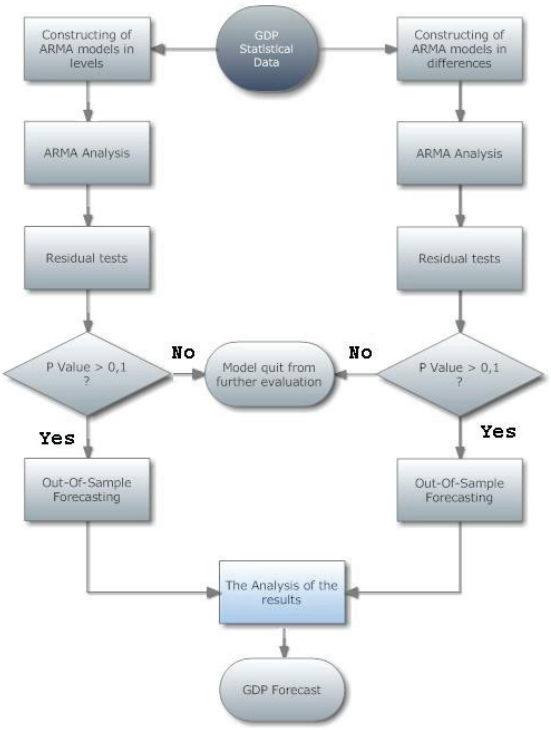


Figure 6 The search algorithm

References

[1] LUTKEPOHL H., KRATZIG M. Applied Time Series Econometrics. – Cambridge: Cambridge University press, 2004, 323 pp.
[2] EViews software documentation: Quantitative Micro Software– http://www.eviews.com – Resource is described 2010, 30 of august.
[3] NOSKO V., Econometrics: introduction to regression analysis of time series. – Moscow: Moscow Institute of Physics and Technology , 2002, 273 pp.
[4] MOLCANOV I., ARZENOVSKI S. Statistical methods of forecasting. – Rostov-na-Donu: Rostov State University, 2001, 74 pp.

Current address

Bezrucko Aleksandrs, Mgr.
Chair of the theory of probability and the mathematical statistics
Faculty of Computer Science and Information Technology
Riga Technical University, 1/4 Meza Street, LV-1048 Riga, Latvia
e-mail: bezrucko@gmail.com

Journal of Applied Mathematics

volume IV (2011), number III

NELSON-SIEGEL MODEL FOR THE ESTIMATION OF YIELD CURVE DERIVED FROM THE CZECH COUPON BOND MARKET

HLADÍKOVÁ Hana, (CZ)

Abstract. The zero coupon yield curve is one of the most fundamental tools in finance and is essential in the pricing of various fixed-income securities. Zero coupon rates are not observable in the market for a range of maturities. Therefore, an estimation methodology is required to derive the zero coupon yield curves from observable data. If we deal with approximations of empirical data to create yield curves it is necessary to choose suitable mathematical functions. We use parametric model of Nelson and Siegel. The current mathematical apparatus employed for this kind of approximation is outlined. In order to find parameters of the model we employ the least squares minimization of computed and observed prices. This theoretical background is applied to an estimation of the zero-coupon yield curve derived from the Czech coupon bond market. On an initial test data sample we have not faced any problems, reported elsewhere, of not having found the global optimum or having found multiple local minima.

Keywords. Yield curve estimation, Nelson-Siegel model, nonlinear least squares methods.

Mathematics Subject Classification: Primary 91G60, 91G30; Secondary 97M30.

1 Introduction

The term structure of interest rates is defined as the relationship between the yields of default-free pure discount (zero-coupon) bonds and their time to maturity. It provides a basis for pricing fixed-income securities and interest rate derivatives, as well as other capital assets.

Yield curve estimation plays a central role in pricing fixed-income derivatives, risk management and for national central banks. Since the yield curve cannot be directly observed, and there are not enough zero coupon bonds existing, it has to be derived from observed market prices of coupon bearing bonds.

The first class are parametric models. This class of function-based models includes the model proposed by Nelson and Siegel (1987) and its extension by Svensson (1994).

The second class of term-structure estimation employs a B-spline basis for the space of cubic splines to fit observed coupon-bond prices. As a consequence, we call these spline-based models. This approach includes a penalty in the generalized least-squares objective function.

Bolder and Gusba (2002), Marciniak(2006), Lin (2002) provide an extensive review and comparison of a number of estimation algorithms.

As to the Czech coupon bond market, the function-based construction of yield curve has not yet been satisfactorily explored. Construction of yield curves by the Svensson method is dealt with in Slavík (2001), Radova, Málek and Štěrba (2007) and Kladivko(2009).

The model of Nelson and Siegel (1987) and its extension by Svensson (1994) are used by central banks and other market participants as a model for the term structure of interest rates (Table 1).

Central bank	Model
Belgium	Svensson/Nelson–Siegel
Canada	Exponential spline
Finland	Nelson–Siegel
France	Svensson/Nelson–Siegel
Germany	Svensson
Italy	Nelson–Siegel
Japan	Smoothing splines
Norway	Svensson
Spain	Svensson
Sweden	Smoothing splines/Svensson
Switzerland	Svensson
UK	VRP
USA	Smoothing splines

Table 1. Estimation of the term structure of interest rates by different central banks.
Source: BIS (2005)

In Section 2 we repeat three equivalent descriptions of the term structure of interest rates, namely, the discount function d, the spot yield curve z and forward yield curve f. In Section 3 we define the Nelson-Siegel model and propose an iterative method to solve arising nonlinear least squares problem. The minimization problem is stated in terms of observed and computed prices rather than in observed and computed yields to maturity (YTM's). In Section 4 the data sample from the Czech coupon bond market is described. In Section 5 numerical experiments on these data are performed. Problem of finding global optimum is explored.

2 Term structure

The spot interest rate *z(t,T)* of a given maturity *T* is defined as the yield on a pure discount bond of that maturity. The spot rates are the discount rates determining the present value of a unit payment at a given time in the future. Spot rates considered as a function of maturity are referred to as the term structure of interest rates.

Each coupon bond can be considered as a package of discount bonds, namely one for each of the coupon payments and one for the principal payment. The price of such component discount bonds is equal to the amount of the payment discounted by the spot rate of the maturity corresponding to this payment. The price of the coupon bond is then the sum of the prices (discount function) *d(t,T)* of these component discount bonds.

The implications of the current spot rates for future rates can be described in terms of the forward rates *f(t,T)*. The forward rates are one-period future reinvestment rates, implied by the current term structure of spot rates.

There are three equivalent descriptions of the term structure of interest rates the discount function *d*, the spot yield curve *z* and forward yield curve *f*. We use $m = T - t$ to denote the time to maturity.

$$\begin{aligned} d(m) &= e^{-(m)z(m)}, \qquad z(m) = \frac{-\ln(d(m))}{m}\ , \\ f(m) &= \frac{\partial}{\partial m}\ln(-d(m)) = z(m) + (m)z'(m), \qquad z(m) = \frac{\int_0^m f(u)du}{m}., \end{aligned} \tag{1}$$

3 The Nelson –Siegel model

We do not actually observe zero-coupon rates, forward rates, or the discount function. We observe the set of coupon bond prices that are traded in the bond market at a given point in time. We minimize the weighted sum of the squared deviations of the fitted prices from the quoted prices.

Nelson-Siegel (1985) suggested forward curve to be estimated as:

$$f(m) = \beta_0 + \beta_1 e^{-\frac{m}{\tau}} + \frac{m}{t}\beta_2 e^{-\frac{m}{\tau}}\ . \tag{2}$$

The model has interesting economic interpretation of parameters and good asymptotical characteristics (Seppälä and Viertiö, 1996).

- $\lim_{m\to\infty} f(m) = \beta_0$, $\lim_{m\to 0} f(m) = \beta_0 + \beta_1$,
- The value of parameter $\beta_0 > 0,$ represents the asymptote of zero coupon yield curve function,

- The asymptote of forward curve as remained maturity approaches to infinity and can be interpreted as long term interest rate,
- The sum of parameters $\beta_0 + \beta_1$ represent initial value of forward curve $f(0) = \beta_0 + \beta_1$, which can be interpreted as instantaneous spot interest rate, thus we require $\beta_0 + \beta_1 > 0$.
- The value of parameter β_1 represents the deviation of the function values from the asymptote and can intuitively be explained as the curvature of the function or as the difference between long term and short term forward interest rates,

Using (1) we obtain from Equation (2) the zero coupon rate *z* and the discount function *d* as follows:

$$z(m) = \frac{1}{m}\int_0^m \beta_0 + \beta_1 e^{-\frac{u}{\tau}} + \frac{u}{\tau}\beta_2 e^{-\frac{u}{\tau}} du = \beta_0 + (\beta_1 + \beta_2)\left[\frac{1 - e^{-\frac{m}{\tau}}}{\frac{m}{\tau}}\right] - \beta_2 e^{-\frac{m}{\tau}} \tag{3}$$

$$d(m) = e^{-m.z(m)} = e^{-m\left(\beta_0 + (\beta_1+\beta_2)\left[\frac{1-e^{-\frac{m}{\tau}}}{\frac{m}{\tau}}\right] - \beta_2 e^{-\frac{m}{\tau}}\right)} \tag{4}$$

We define:

P_i - theoretical price of *i*-th bond, $\overline{P_i}$ - observed price of *i*-th bond,

l_i number of the payments of the *i*-th bond

t_{ij} - the time when the *j*-th payment of the *i*-th bond occurs; $m_{ij} = T_i - t_{ij}$

c_{ij} - the *j*-th payment of the *i*-th bond.

Let $\theta = (\beta_0, \beta_1, \beta_2, \tau)^{\mathrm{T}}$. The theoretical price P_i of bond number *i* is given by the sum of the discounted values of its cash flows, which using (4) is

$$P_i(\theta) = \sum_{j=1}^{l_i} c_{ij} d(m_{ij}, \theta) = \sum_{j=1}^{l_i} c_{ij} e^{-m_{ij} z(m_{ij}, \theta)} \tag{5}$$

The final step is to actually estimate the parameters of the Nelson-Siegel model. A natural requirement is to find these parameters such that the theoretical prices P_i are as close as possible to the observed prices $\overline{P_i}$. Thus, in the sense of the least squares method we want to find a set of parameters $\beta_0, \beta_1, \beta_2, \tau$ that minimizes the function *H(P)* given as,

$$H(P) := \sum_{i=0}^{N} w_i (P_i - \overline{P_i})^2, \quad \text{where } w_i \text{ is weight of the } i\text{-th bond.} \tag{6}$$

Our choice for the weights was the reciprocal of the modified duration.
We need to estimate four parameters: $\beta_0, \beta_1, \beta_2, \tau$. For *N* observed prices with different maturities $T_1, \cdots T_N$, we have *N* equations.
There is a natural strategy to obtain parameters for this model: fix parameter τ, and then estimate the $\beta_0, \beta_1, \beta_2$ values with least squares method. The model's parameters can change over time. We define $\theta_\tau = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$

$$\widetilde{P}_i(\theta_\tau) = \sum_{j=1}^{m_i} c_{ij} d(m_{ij}, \theta_\tau) = \sum_{j=1}^{l_i} c_{ij} e^{-m_{ij}\left(\beta_0 + (\beta_1+\beta_2)\left[\frac{1-e^{-\frac{m_{ij}}{\tau}}}{\frac{m_{ij}}{\tau}}\right] - \beta_2 e^{-\frac{m_{ij}}{\tau}}\right)}. \tag{7}$$

We let $\widetilde{P}(\theta_\tau) = [\widetilde{P}_1(\theta_\tau), \cdots \widetilde{P}_N(\theta_\tau)]^T$ be a vector of theoretical prices for the set of N bond observations. Our objective, therefore, is to solve the minimization problem,

$$\min_{\theta_\tau}((P - \widetilde{P}(\theta_\tau))^T W (P - \widetilde{P}(\theta_\tau)), \text{ where } W \text{ is an } N \times N \text{ weighting matrix.} \tag{8}$$

Equation (8) is a nonlinear least-squares problem. We apply the following nonlinear optimization algorithm (see e.g. Fischer, Nychka a Zervos, 1994):

1. Employ the linear first-order Taylor series approximation:
$$\widetilde{P}(\theta_\tau) \approx \widetilde{P}(\theta_\tau^0) - (\theta_\tau - \theta_\tau^0) X(\theta_\tau^0), \text{ where } X(\theta_\tau) = \frac{\partial \widetilde{P}(\theta_\tau)}{\partial \theta_\tau{}^T} \tag{9}$$
2. Define:
$$Y(\theta_\tau^0) = P - \widetilde{P}(\theta_\tau^0) + \theta^0 X(\theta^0), \tag{10}$$
3. Solve the linear least-squares approximation to the original problem given as:
$$\begin{aligned} &\min_{\theta_\tau}(Y(\theta_\tau^0) - \theta_\tau X(\theta_\tau^0))^T W (Y(\theta_\tau^0) - \theta_\tau X(\theta_\tau^0)), \text{ which is solved by,} \\ &\theta_\tau^1 = (X(\theta_\tau^0)^T W X(\theta_\tau^0))^{-1} (X(\theta_\tau^0)^T W Y(\theta_\tau^0)) \end{aligned} \tag{11}$$
4. Return to Step 1 with $\theta^0 := \theta^1$ until convergence is not achieved.

Note that the above algorithm defined by Equations (9) to (11) is well suited for finding a local minimum of problem (8). The question whether this local minimum is also a global minimum will be addressed in Section 4 (cf. Gauthier and Simonato, 2009). We also did not impose any constraints on β's ($\beta_0 > 0,\ \beta_0 + \beta_1 > 0$). It seems that if the problem is well posed then these constraints are automatically satisfied for 'reasonable' values of τ.

Alternatively, in place of using observed and theoretical prices in Equation (6) we can minimize the error of observed and theoretical yields to maturity (YTM's) to find the Nelson-Siegel model parameters.

4 Data from the Czech coupon bond market

The Czech market is small and not as liquid as other developed markets. The original life of the Czech government bond is from 3 to 50 years. The government issued bonds with annual coupon payments. We consider here data for a selected day as given in Table 2.

	Coupon	Maturity	Duration	Price+AUV	Years to maturity
CZ0001000731	6,4	14.4.10	-	106,3589	0,139726
CZ0001001242	2,55	18.10.10	0,64	101,8496	0,652055
CZ0001002158	4,1	11.4.11	1,08	106,7261	1,131507
CZ0001000764	6,55	5.10.11	1,53	110,7972	1,616438
CZ0001001887	3,55	18.10.12	2,49	104,5524	2,654795
CZ0001000814	3,7	16.6.13	3,03	105,9092	3,315068
CZ0001001143	3,8	11.4.15	4,47	105,6644	5,134247
CZ0001000749	6,95	26.1.16	4,95	119,4099	5,928767
CZ0001001903	4	11.4.17	5,91	103,8389	7,136986
CZ0001000822	4,6	18.8.18	6,8	105,7394	8,490411
CZ0001002471	5	11.4.19	7	109,8111	9,136986
CZ0001001317	3,75	12.9.20	8,31	94,89792	10,56164
CZ0001001945	4,7	12.9.22	9,13	101,5281	12,56164
CZ0001001796	4,2	4.12.36	14,95	87,945	26,8
CZ0001002059	4,85	26.11.57	17,64	93,69903	47,79178

Table 2. Government coupon bonds (22.2.2010).
Source: www.patria.cz, personal computing

We exclude two bonds with less than three months to maturity, since the yields on these securities often seem to behave oddly and one bond with more than forty-seven years to maturity, since price of bond will evidently include also another risk premium.

5 Numerical experiments

With the set of data described in Section 4 we performed a couple of numerical experiments. We used our own code written in FORTRAN.

The following measures of goodness of fit are used:

- $L2 = \sqrt{\sum_{i=1}^{N}\left(\overline{P}_i - P_i\right)^2}$, $L2W = \sqrt{\sum_{i=1}^{N}\left(\overline{P}_i - P_i\right)^2 w_i}$ l_2-norm of error and weighted l_2-norm of error (12)

- $RMSE = \sqrt{\sum_{i=1}^{N}\frac{\left(\overline{P}_i - P_i\right)^2}{N}}$ squared error (RMSE),

- $MAE = \sum_{i=1}^{N}\frac{\left|\overline{P}_i - P_i\right|}{N}$ absolute error, (12)

- $HR = \frac{card\left(\overline{P}_i,\ P_i^O \le \overline{P}_i \le P_i^B\right)}{N}$ hit ratio (HR).

P_i^O and P_i^B offer and bid price of the i-th bond.

In RMSE more weight is assigned to extraordinarily high error values. Large differences between RMSE and MAE indicate a large number of large errors of fit. HR is the number of theoretical bond prices, as a proportion of the overall number of bonds in the daily sample.

Two measures of maximum smoothness of a curve $y = g(x)$ between a and b are used.:

- $s = \int_a^b \sqrt{1 + \left[g'(x)\right]^2}\,dx$ "minimum length"
- $z = \int_a^b g''(x)^2\,dx$ "smoothness" (13)

The smoothest possible function has the minimum z value. Since a second derivative of a straight line equals zero, z is zero in that case and a straight line is perfectly smooth.

In this case, the function $g(x)$ stands for the discount, spot or forward rate curve. The two measures are once again very consistent. We have critiqued the results from some yield curve smoothing techniques because of the lack of smoothness in either discount function, spot rates or forward rates. In order to evaluate z over the full maturity spectrum of the rates curve, the rate segments must be at least twice differentiable at each point.

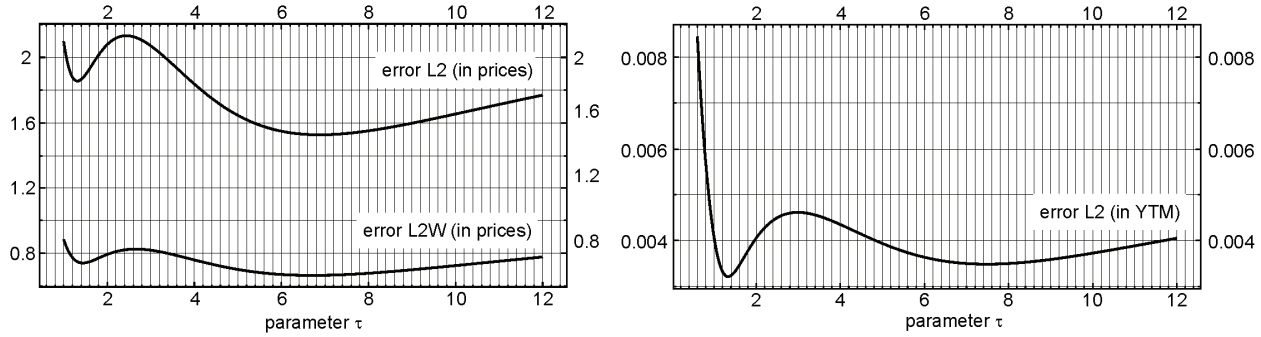


Figure 1. The L2-error, L2W-error (in prices) and L2-error (in yields, YTM = Yield To Maturity) for different values of parameter τ

Our initial tests revealed that values of τ could be restricted to $0 < \tau < 12$ (cf. Gilli at al., 2010). For a fixed τ we repeatedly solved minimization problem (8) to obtain $\beta_0, \beta_1, \beta_2$ applying algorithm defined by Equations (9) to (11). For these solutions we compared the L2W-errors of observed and estimated prices (see Figure 1). The least L2W-error was obtained for value $\tau = 6.7$. For this solution we computed the discount, forward and spot yield curves (Figure 2). In order to check the quality of our solution we compared the results with a time consuming global optimization strategy. This global strategy used coarse-fine bracketing of the four parameters requiring over one million attempts. In terms of the L2W-error the global strategy did not find a better solution for our test data.

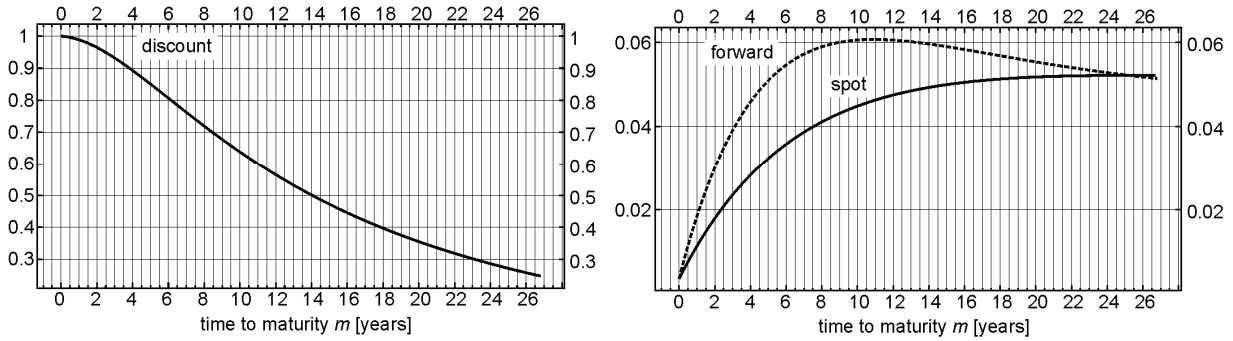


Figure 2. Computed discount function, spot and forward rates vs. time for the best solution in L2W-error

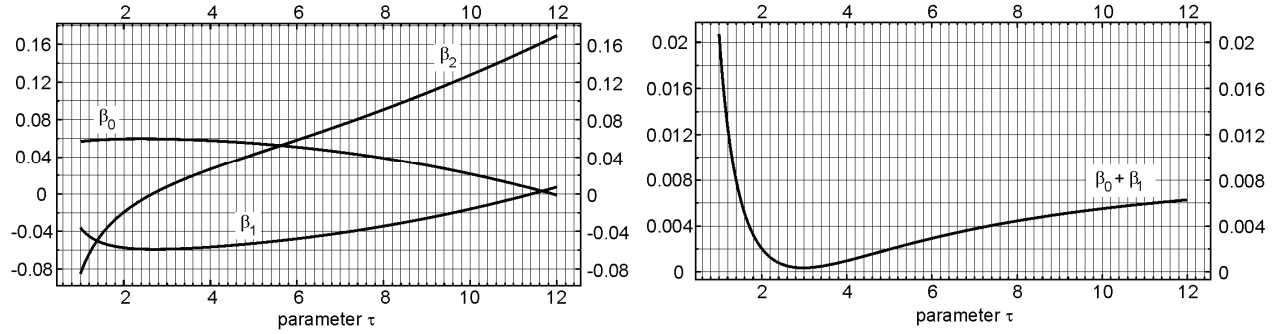


Figure 3. Parameters $\beta_0, \beta_1, \beta_2$ and $\beta_0 + \beta_1$ computed for different values of τ by algorithm (9)-(11).

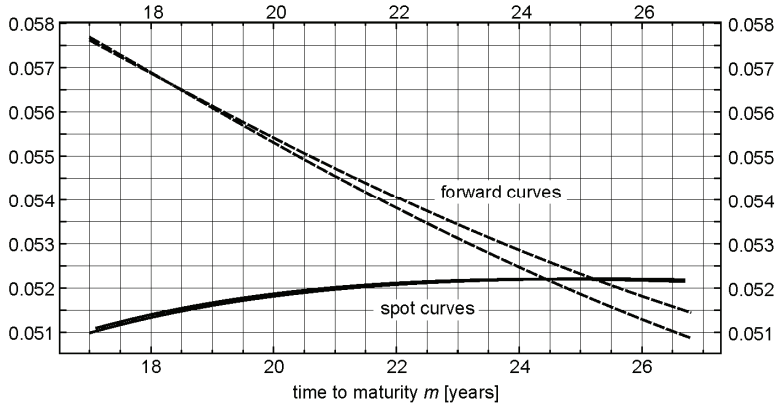


Figure 4. Comparison of spot and forward curves for solutions with and without weights.

The error in prices does not show erratic behavior in dependence on parameter τ (Figure 1). The minimization algorithm (9) - (11) found always the global minimum. Its convergence was fast and robust. The results were compared with a global strategy where we used values of $\beta_0, \beta_1, \beta_2, \tau$ from given intervals. The initial coarse estimates were: $0 < \beta_0 < 0.15$, $-0.15 < \beta_1 < 0.3$, $-0.3 < \beta_2 < 0.3$, $0 < \tau < 30$. For given $\beta_0, \beta_1, \beta_2, \tau$'s we recorded not only the L2W-error (objective function in Equation (6)) but also the other measures of error, namely RMSE[2], MAE, L2. Moreover, the MAE- and L2-errors were used to measure the error of observed and computed YTM's. Computed solutions of the minimization problem (6) with and without weights are given in Table 3. It is apparent that the obtained coefficients $\beta_0, \beta_1, \beta_2, \tau$ do not differ much. The use of the reciprocal of the modified duration as a weight w_i in Equation (6) does not show much influence on the obtained solution. This is also demonstrated in Figure 4 where we can see the differences in solutions with and without weights only on the long end.

τ	β_0	β_1	β_2	$\beta_0 + \beta_1$	Estimated Repo (%)	Weight	Measure of error	Value
6,70	0,04717	-0,04365	0,06925	0,00352	0,35236	1/MDur	L2W	0,66485
7,00	0,04568	-0,04175	0,07327	0,00393	0,39352	1	L2W=L2	1,52796

Table 3. Characterization of the obtained solutions.

The length and curvature of the two methods is compared in Table 4. The different error values are compared in Table 5. The obtained values are again very similar.

	Length			**Smoothness**		
Solution	**Discount**	**Spot**	**Forward**	**Discount**	**Spot**	**Forward**
L2=L2W	26.81211359	26.70010485	26.80026663	0.44603790	0.00536658	0.03474004
L2W	26.81211632	26.70010765	26.80027282	0.45881991	0.00577223	0.03734770

Table 4. Evaluation of the obtained solutions according to the length of the curves and smoothness.

Price				YTM (%)	
Solution	**MAE**	**RMSE[2]**	**HR**	**MAE**	**L2**
L2=L2W	0.32041911	0.17958841	53.85	0.07316269	0.34828361
L2W	0.32453623	0.17993799	53.85	0.07639381	0.35238783

Table 5. Evaluation of the obtained solutions according to accuracy of price and YTM estimations.

From our solutions we obtain estimates of the instantaneous forward rate curve from Czech government coupon bonds. We can understand this estimate as an approximation of market expectations regarding future short-term interest rates. We can see that the starting value of estimated forward rate (column $\beta_0 + \beta_1$ in Table 3) does not fit quite well the actual repo rate (the repo rate set by ČNB was 1 % on 22.2.2010). The estimated value is between 0.35% and 0.48% - it is below the actual repo rate. The estimates do provide picture of evolution of the forward curve. This is low level of expected repo rate in near future. The figures also show the gradually increasing forward curve. This corresponds to expectation of gradual increase of repo rate, which was consistent with market expectation as measured by CNB. Despite these expectations the repo rate dropped to value of 0,75% on 7.5.2010.

6 Conclusions

Results presented in this paper were based on interest rate estimates from the Czech coupon bond market, which is characterized by a relatively low number of bonds, by moderate liquidity and periodically reduced efficiency. We explored Nelson-Siegel model to create yield curves. This approach produced a reasonably looking zero-coupon yield curve. After some numerical experimentations we found the Nelson-Siegel model to be a stable and potentially useful for our data sample. We have not faced any problems, reported elsewhere, of not having found the global optimum or having found multiple local minima. This will be clarified in our subsequent work when compared to other methods and on larger set of data than just one day.

References

[1.] BOLDER,D.J. ,GUSBA,S: *Exponentials, Polynomials and Fourier Series: More Yield Curve Modelling at the Bank of Canada*, Bank of Canada Working Paper, no.2002-29,(2002).

[2.] BING-HUEI L: *B-splines: the case of Taiwanese Government bonds*, Applied Financial Economics,(2002, 12), pp. 57-75

[3.] EILERS, P. H.,, B. D. MARX : *Flexible Smoothing with B-splines and Penalties*," Statistical Science, 11, (1996) pp.89-102.

[4.] FISHER, M.,. NYCHKA, D,ZERVOS, D.,: *Fitting the Term Structure of Interest Rates with Smoothing Splines*, U.S. Federal Reserve Board Working Paper,(1994).

[5.] LI, B., E. DeWETERING, G. LUCAS, R. BRENNER, and A. SHAPIRO.. *Merrill Lynch Exponential Spline Model*., Merrill Lynch Working Paper,(2001)

[6.] McCULLOCH, J. H.: *Measuring the Term Structure of Interest Rates*, Journal of Business, 44,19{31, (1971).

[7.] MÁLEK,J., *Dynamika úvěrových měr a úrokové deriváty*. Praha:Ekopress(2005).
[8.] MÁLEK,J.,RADOVÁ,J.-ŠTERBA F., **Konstrukce výnosové křivky pomocí vládních dluhopisů v České republice**,Politická ekonomie 6(2007),pp792-827.
[9.] SLAVÍK, M., *Odhad časové struktury úrokových sazeb z cen domácích dluhopisů*, Finance a úvěr 11(2001), pp. 591-606.
[10.] MARCINIAK, M., *Yield Curve Estimation at the National Bank of Poland* ,Bank i kredit (2006).
[11.] SVENSSON, L.E.. *Estimating and Interpreting Forward Interest Rates: Sweden 1992-1994.* Centre for Economic Policy Research, Discussion Paper 1051.(1994)
[12.] WAGGONER, D. F. (1997): *Spline Methods for Extracting Interest Rate Curves from Coupon Bond Prices*, Federal Reserve Bank of Atlanta: Working Paper 97-10 (1997).
[13.] GAUTHIER, G. and SIMONATO, J. G. *Linearized Nelson-Siegel and Svensson models for the estimation of spot interest rates*. Available at *http://ssrn.com* (2009).
[14.] GILLI, M., GROSSE, S., SCHUMANN, E. *Calibrating the Nelson-Siegel-Svensson model*. Computational Optimization Methods in Statistics, Econometrics and Finance (COMISEF), Working Papers Series, WPS-031 30/03/2010, pp 1-23 (2010).
Electronic funds: www.cnb.cz; www.patria.cz, www.bankofcanada.com, www.twiter.com

Current address

RNDr. Hana Hladíková
University of Economics,W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
e-mail: Hana.Hladikova@vse.cz

Journal of Applied Mathematics

volume IV (2011), number III

INTEREST RATES ON RETAIL HOUSE PURCHASE LOANS: IS SLOVAKIA AN EXEPTION IN THE EUROZONE?

KLACSO Ján, (SK)

Abstract. After the European Central Bank lowered its base rate and adopted several non-standard measures under the pressure of the financial crisis, interbank interest rates and consequently client interest rates in the Eurozone dropped to their historical minimum. Interest rates on retail house purchase loans in Slovakia, however, remained relatively high compared to other member states. Based on cointegration techniques and error-correction equations, it is shown in the paper that it is the development of the government bond yields, which can be related to the development of the interest rates on retail hose purchase loans rather than the interbank market interest rates and also the liquidity margin affects partially the level of these interest rates in Slovakia. It means that the higher value of the interest rates on client loans in Slovakia is a result of historical differences in setting these interest rates between Slovakia and other member states rather than of differences in the reaction to the developments in the first half of 2009.

Key words. Interest rate pass-through, cointegration, error-correction

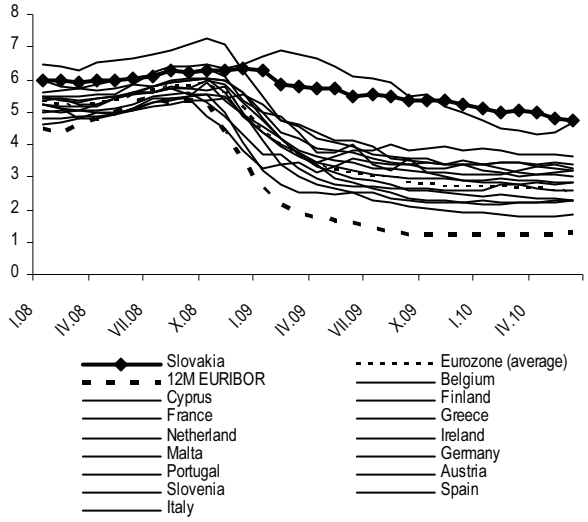
Mathematics Subject Classification: **91G70**

1 Introduction

Under the pressure of the turmoil on the financial markets and the global economic downturn, the European Central Bank decided to gradually decrease its base rate in 2008 and 2009 from more than 3 % to 1 % and to adopt several non-standard measures to lower the impact of the crisis on the banking sector and the real economy. As a consequence, the European interbank market interest rates dropped to their historical minimum. Interest rates on client loans and deposits in the member states of the Eurozone gradually followed the movements of the interbank interest rates. This development is also observable in case of Slovakia. However, for some type of loans, especially for retail house purchase loans, interest rates in Slovakia remained among the highest within the Eurozone, as their decrease was not as significant as in other member states. It is therefore a question, if the movements of these interest rates in Slovakia can be explained by the movements of the interbank interest rates and what other determinants lay behind the development of the interest

rates on retail house purchase loans. As house purchase loans with interest rate fixation of up to 1 year has the highest share in the total volume of retail house purchase loans, the interest rates on these loans are investigated in the paper.

Chart 1 Development of interest rates on newly granted retail house purchase loans with interest rate fixation of up to 1 year in the Eurozone member states



- Data are in percentages
- Source: www.ecb.int

2 Finding for long-term relationship

Based on economic theory, banks transmit the price of the funds they use to finance the loans to the interest rates on these loans. As it is interbank interest rates that determine primarily the price of funds, it is expected that mainly the value of these interest rates is reflected in the value of client interest rates. If banks really derive interest rates on retail house purchase loans from the level of interbank interest rates, and if there are no significant changes in the way of determining these interest rates, respectively market conditions are not subject to any significant changes, the existence of long-term relationship between the client interest rates and interbank interest rates can be expected. As all interbank interest rates and interest rates on newly granted retail house purchase loans are regarded as non-stationary time series of order 1 (so called I(1) processes) based on unit root tests (Table 1), the existence of long-term relationship was investigated using tests of cointegration. As ECB reduced its key interest rate and adopted the non-standard measures at the end of 2008 and during 2009, the existence of long-term relationship was tested first for the period 2005-2008.

After verifying the existence of long-term relationship between the client interest rates and the interbank interest rates (

Table 2), an EC (error correction) equation was estimated for client interest rates in the following form:

$$\Delta IR_t = \alpha\left(IR_{t-1} + \beta_0 + \beta_1 IBR_{t-1}\right) + \gamma_0 + \sum_{i=1}^{p}\left(\gamma_i \Delta IR_{t-i} + \delta_i \Delta IBR_{t-i}\right) + \varepsilon_t \qquad (2.1)$$

where IR_t is the average interest rate on newly granted retail house purchase loans, IBR_t is the respective interbank interest rate, α is the speed of adjustment in case of a deviation from the long-term relationship, β_0 is the estimated long-term spread between the client interest rate and the interbank interest rate, β_1 determines up to what extend are changes in the interbank interest rate transmitted into the client interest rate in the long term and ε_t are residuals. For all the above described coefficients a negative value is expected.

As the existence of cointegrating relationship cannot be rejected for any of the interbank interest rates and the best fit is for the error-correction equations in case of interbank rates of up to 6, 9 and 12 months (Table 3), only these interbank rates were used for further estimates.

As the existence of long-term relationship cannot be rejected for the period 2005-2008, the next question is whether this long-term relationship exists also for the prolonged period from 2005 to the end of the first half of 2010. As Slovak interbank interest rates (BRIBOR) existed only up to the end of 2008 (as Slovakia joined the Eurozone as of 1st January 2009), the time series of BRIBOR interest rates were prolonged using EURIBOR interbank interest rates. Tests of cointegration rejected the hypothesis of the existence of long-term relationship between the prolonged time series (

Table 2) and also the estimated equation (2.1) had poor performance for this prolonged period (

Table 4). It means that since January 2009 the relationship between the rates and/or the market conditions have substantially changed. Potential reasons for the elimination of the long-term relationship include:

- Euro interbank rates extending the time series from January 2009 do not reflect the credit risk, which was specific for Slovakia and was reflected in the BRIBOR interest rates (i.e. by how much, compared to banks of other member states, do Slovak banks pay more on the money market if they wish to borrow funds);
- Differences between short-term and long-term interbank rates have been substantially changed since the beginning of 2009. That could have become evident through an increase in the liquidity margin included in the client interest rate (expressing the risk associated with the fact that despite shorter fixation of the interest rate, the maturity of house purchase loans is usually relatively long and the bank has to use long-term funds to cover the loans and/or create a buffer to secure the availability of funds in the future, as their price can be higher than the present price). The liquidity margin was not reflected in the estimated EC equations, however, up to the end of 2008 spreads between short-term and long-term rates were relatively negligible compared to values from 2009.
- After euro introduction banks lost certain sources of income (EUR/SKK foreign exchange operations, sterilizing operations with the NBS, etc.), which could be replaced by increasing margins on products with relatively low competitive pressures (e.g. retail house purchase loans) compared to other products (e.g. municipal loans or corporate loans). Moreover, from 2009, household credit risk might have been perceived more sensitively owing to the adverse developments in the global (and consequently local) economy, which became evident in increased uncertainty regarding future developments on the labour market.

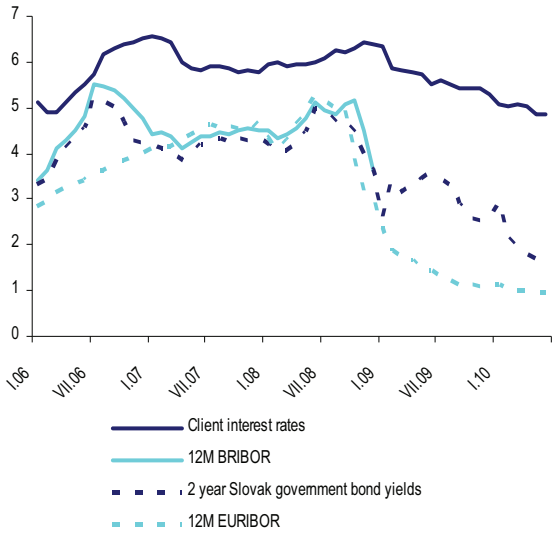
3 Inclusion of government bond yields and liquidity margin

As the available interbank interest rates from the beginning of 2009 do not reflect the credit risk specific for Slovakia (so called sovereign credit risk), time series of interbank interest rates were approximated by time series of the yields on 2 year Slovak government bonds. Tests of cointegration didn't reject the existence of long-term relationship for the prolonged period (
Table 2), therefore an EC equation was estimated for the client interest rates of the form:

$$\Delta IR_t = \alpha\left(IR_{t-1} + \beta_0 + \beta_1 GBY_{t-1}\right) + \gamma_0 + \sum_{i=1}^{p}\left(\gamma_i \Delta IR_{t-i} + \delta_i \Delta GBY_{t-i}\right) + \varepsilon_t \qquad (3.1)$$

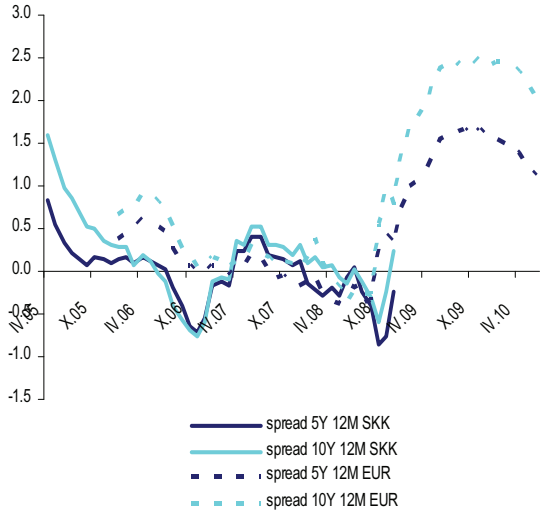
where GBY_t is the yield on 2 year government bonds. The explanation of the coefficients is as described above. Based on the estimation of equation (3.1) and the tests of cointegration it can be concluded that the yield on 2 year Slovak government bonds is a good approximation of the BRIBOR interbank interest rates for the period 2005-2008 and as long-term relationship cannot be rejected for the prolonged period it seems that Slovak banks reflect this yields in client interest rates rather than interbank interest rates (
Table 5).

Chart 2 Development of client interest rates, interbank interest rates and government bond yields



- Client interest rates are interest rates on newly granted retail house purchase loans with fixation of up to 1 year

Chart 3 Development of spread between short and long-term interest rates



- The spread is calculated as the difference between the short and long-term discount rate

Values are in percentage points

To test whether the increased steepness of the yield curve affected the value of the client interest rates the liquidity margin approximated by the difference between 10 years and 5 years interbank interest rate was included into the EC equation after tests didn't reject the existence of long term relationship between the client interest rates, the government bond yields and the liquidity margin (
Table 2). The estimated EC equation has the form:

$$\Delta IR_t = \alpha\left(IR_{t-1} + \beta_0 + \beta_1 GBY_{t-1} + \beta_2 LM_{t-1}\right) + \gamma_0 + \sum_{i=1}^{p}\left(\gamma_i \Delta IR_{t-i} + \delta_i \Delta GBY_{t-i} + \rho_i \Delta LM_{t-i}\right) + \varepsilon_t \quad (3.2)$$

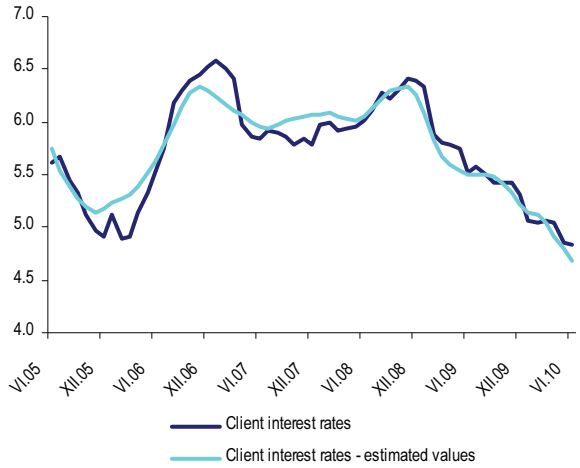
where LM_t is the liquidity margin. A negative value is expected for the coefficient β_2. Estimation of equation (3.2) showed that the spread between the long-term and short-term interbank interest rates is reflected in the client interest rates, however to a much lover extent than the country specific credit risk included in the government bond yields.

4 Testing for structural breaks

The last question is if there was a structural change in setting up the client interest rates by the banks after joining the Eurozone and after the effects of the financial crisis and the global economic recession became more pronounced in 2009. To test the stability of the estimated coefficients of the equation (3.2) Chow breakpoint test was used for each month of the period December 2008 – May 2010. Based on the breakpoint test, the null hypothesis of no structural brakes in the first months of 2009 cannot be rejected (

Table 7). It means that the changing market conditions in 2009 didn't cause any structural changes in the way banks set up their client interest rates.

Chart 4 Interest rates on newly granted retail house purchase loans with fixation of up to 1 year – real values and estimation



5 Conclusion

Until the end of 2008 it was hard to decide whether it is the yield on Slovak government bonds or the interbank market interest rates that is reflected in the client interest rates, as their development was strongly correlated. However, based on cointegration tests and the estimated EC equations it can be concluded, that at least from the beginning of 2009 it is the 2 year government bond yield which determinates the value of the interest rates on house purchase loans. It means that banks reflect also the country specific risk related to Slovakia when deciding the value of the loan

interest rate. Estimations showed that the spread between the long-term and short-term interbank rates is also included in the value of the client interest rates in the form of liquidity margin. As there is no evidence of any structural change in the way Slovak banks determine the client interest rates during the first half of 2010, it can be concluded that the difference in the client interest rates between Slovakia and other member states can be related to historical differences in setting these interest rates rather than to different reactions to changes at the end of 2008 and during the first half of 2009.

References

[1.] SVERIGES RISKBANK: *The transmission mechanism*, Available on internet: < http://www.riksbank.com/templates/Page.aspx?id=10547 >

[2.] BANK FOR INTERNATIONAL SETTLEMENTS: *The transmission of monetary policy in emerging market economies*, BIS policy papers No. 3. January 1998

[3.] HEFFERNAN, Shelagh, FUERTES, Ana – Maria: *Bank Heterogeneities in the Interest Rate Transmission Mechanism*, Cass Business School Research Paper, July 2006, Available on internet: < http://papers.ssrn.com/sol3/papers.cfm?abstract_id=903348 >

[4.] LEON, Costas, CHIONIS, Dionysos: *Modeling Interest Rate Transmission Dynamics in Greece – Is There Any Structural Brake After EMU?*, Working Paper series, May 2005, Available on internet: < http://papers.ssrn.com/sol3/papers.cfm?abstract_id=815584 >

[5.] BURGSTALLER, Johann: *Interest Rate Transmission to Commercial Credit Rates in Austria*, Working Paper No. 0306, May 2003, Available on internet: < http://ideas.repec.org/p/jku/econwp/2003_06.html >

[6.] ENDERS, Walter: *Applied econometric time series*, 1995, John Wiley and Sons, Inc.

Current address

Ján Klacso, Mgr.

Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics
Mlynsk8 dolina
842 48 Bratislava
tel.: +421 2 5787 2899
E-mail.:klacso@gmail.com

Appendix

Table 1 Unit root tests

	ADF test		Philips – Perron test	
	Level	1st differences	Level	1st differences
Client rates	0.608	0.000	0.094	0.000
Interbank rates				
1M	0.815	0.001	0.842	0.000
2M	0.643	0.020	0.834	0.015
3M	0.598	0.026	0.818	0.026
6M	0.478	0.048	0.788	0.033
9M	0.455	0.048	0.773	0.034
12M	0.455	0.048	0.773	0.034
2Y	0.793	0.000	0.827	0.000
3Y	0.751	0.000	0.826	0.000
4Y	0.738	0.000	0.813	0.000
5Y	0.729	0.000	0.831	0.000
6Y	0.672	0.000	0.792	0.000
7Y	0.659	0.000	0.766	0.000
8Y	0.644	0.000	0.748	0.000
9Y	0.631	0.000	0.734	0.000
10Y	0.743	0.000	0.833	0.000
2Y GBY	0.897	0.000	0.818	0.000
Liquidity margin	0.766	0.000	0.624	0.000

- All unit root tests are calculated including an intercept

Table 2 Tests of cointegration

Cointegrating relationship with	Period 2005 - 2008				Period 2005 - 2010H1			
	Trace test		ME test		Trace test		ME test	
	0 CE	1CE	0 CE	1 CE	0 CE	1CE	0 CE	1 CE
Interbank rates								
1M	0.043	0.323	0.046	0.323	-	-	-	-
2M	0.011	0.123	0.025	0.123	-	-	-	-
3M	0.002	0.041	0.012	0.041	-	-	-	-
6M	0.000	0.062	0.000	0.062	0.167	0.731	0.099	0.731
9M	0.000	0.088	0.000	0.088	0.285	0.666	0.217	0.666
12M	0.000	0.125	0.000	0.125	0.440	0.641	0.391	0.641
2Y	0.001	0.161	0.001	0.161	-	-	-	-
3Y	0.003	0.218	0.003	0.218	-	-	-	-
4Y	0.004	0.220	0.004	0.220	-	-	-	-
5Y	0.008	0.259	0.008	0.259	-	-	-	-
6Y	0.008	0.258	0.009	0.258	-	-	-	-
7Y	0.009	0.252	0.011	0.252	-	-	-	-
8Y	0.015	0.273	0.017	0.273	-	-	-	-
9Y	0.030	0.311	0.033	0.311	-	-	-	-
10Y	0.058	0.344	0.061	0.344	-	-	-	-
2Y GBY	0.001	0.130	0.004	0.130	0.000	0.557	0.000	0.557
2Y GBY and LM	-	-	-	-	0.001	0.319	0.000	0.294

- All tests of cointegration are calculated including an intercept in CE and VAR

Table 3 Estimated coefficients of equation (2.1) for period 2005- 2008

Interbank rates:	β_0	β_1	α	# lags	aR^2
1M	-3.238	-0.646	-0.271	1	30.75%
2M	-3.177	-0.649	-0.272	1	34.38%
3M	-2.795	-0.735	-0.262	1	41.34%
6M	-1.982	-0.910	-0.243	1	55.49%
9M	-1.938	-0.911	-0.245	1	57.44%
12M	-2.007	-0.888	-0.244	1	57.99%
2Y	-2.307	-0.827	-0.223	1	47.47%
3Y	-1.583	-1.000	-0.192	1	46.68%
4Y	-1.147	-1.102	-0.185	1	46.35%
5Y	-0.674	-1.209	-0.176	1	45.03%
6Y	-0.269	-1.297	-0.175	1	44.83%
7Y	0.054	-1.365	-0.176	1	44.56%
8Y	0.480	-1.451	-0.173	1	43.50%
9Y	0.921	-1.538	-0.168	1	41.40%
10Y	1.373	-1.627	-0.159	1	39.24%

Table 4 Estimated coefficients of equation (2.1) for period 2005- 2010H1

Interbank rates:	β_0	β_1	α	# lags	aR^2
6M	-4.730	-0.298	-0.133	1	27.05%
9M	-4.602	-0.331	-0.149	1	28.48%
12M	-4.514	-0.353	-0.148	1	30.51%

Table 5 Estimated coefficients of equation (3.1)

Period:	β_0	β_1	α	# lags	aR^2
2005 – 2008	-1.784	-1.034	-0.172	1	46.43%
2005 – 2010H1	-3.080	-0.738	-0.221	1	46.42%

Table 6 Estimated coefficients of equation (3.2)

Period:	β_0	β_1	β_2	α	# lags	aR^2
2005 – 2010H1	-3.139	-0.706	-0.003	-0.243	1	49.99%

Table 7 Chow breakpoint test

	2008M12	2009M01	2009M02	2009M03	2009M04	2009M05
F statistics (p-value)	0.8266	0.8413	0.5737	0.5572	0.6266	0.7608
Log likelihood ratio (p-v.)	0.7749	0.793	0.4858	0.4683	0.5431	0.6959
Wald statistics (p v.)	0.8291	0.8439	0.5691	0.5519	0.6241	0.7623

Journal of Applied Mathematics

volume IV (2011), number III

COMPLEX PROJECTS' MANAGEMENT USING EVA – A CASE STUDY

PEDRO Maria I. (P), PEREIRA João (P), FILIPE José António (P), FERREIRA Manuel Alberto M. (P)

Abstract: Earned Value Analysis (EVA) is a method of measuring the project performance. Although the concept exists since the nineteen's century and it has been in use since the 1960s, only now it is gaining considerable popularity. Those in favour will base their arguments in the cost savings from the project and the improved communication, analysis and control that come from its implementation. Those who have a different opinion will cite the limited benefit from its use and the effort to make it work. There is no doubt that these different opinions come from different experiences. Nevertheless, everybody agrees that EVA is a powerful tool if applied correctly. The aim of this work is to implement this tool to a project in concrete and to evaluate the use of EVA as a complementary tool of the system currently used – SAPE, in determining and controlling costs associated with the project, at every moment of the project. This allows to evaluate the possible deviations and to enable timely corrections.

Keywords: Integrated Management Systems, Project Control, ERP, SAPE, Information Systems, Earned Value Analysis.

1 Introduction

Nowadays, consumers are highly exigent. Therefore, companies make great efforts to meet these requirements trying to offer a consistently high level of service, regardless of the area where they operate. To accomplish this objective companies must have the internal capabilities to reach not only the purposes that their clients require but also an efficient level of operation, based on reduced costs.

To achieve the goals mentioned above, the consulting company IA[1] is going to implement EVA (Earned Value Analysis) tool as a complement to the SAPE[2] Software, already in use in the company. Such integration would increase the strengths of both tools in the company. This enables

[1] The name of the company is not the real one, for confidentiality reasons.

[2] "**Systems, Applications and Products** in Data Processing"

IA to get a much more rigid control of its projects. In addition, the project manager will have better information to make his decisions. Consequently, as he will be better informed about the project development, failures can be detected earlier.

Considering this, this paper aims:

1) to get an overall view of the SAPE software, highlighting its main deficiencies,
2) to reach a close analysis to the EVA tool,
3) to propose the application of this model to a case study and
4) to make an analysis of the results and draw the inherent conclusions.

2 SAPE Software

Considering the competitive markets and global economy, companies feel the necessity to develop some vital characteristics in order to survive. For instance, the ability to obtain crucial information and the way to manage it are essential factors for the company. Enterprise Resource Planning (ERP) appeared in order to contribute to that.

SAPE main objective is to permit to companies to plan and to control their necessities through this kind of information systems. While in its early stages, ERP was used as a tool to manage operations, planning and controlling resources necessities. ERP was used to calculate the quantity of resources needed and the correct time to do so. It was used to evaluate the implications of the companies' future demand on the financial areas, as well as to calculate the resources` necessities. Nowadays it is used as a corporative system that supports and assists every business area [1]. ERP is used in a larger way to cover all functional sectors of the company. Therefore, it is now considered a "global plan" and became a software that allows the existence of a unified information system and that connects all business areas. It has become a very agile application that improves the whole Business Process [2].

SAPE or SAP ERP is the main product of SAP AG, a German company and the leader of the corporative software market. Being a type of ERP, SAPE can be now defined as an integrated management transactional system that connects all sectors of a company [3]. SAPE is a very useful tool mostly because it allows a very high integration in the company. Several software can be removed, once SAPE causes a reduction of data inconsistencies. However, this integration can be seen as a disadvantage because it only considers one user to have an effect on all departments. Another disadvantage of this system is that it can be seen as an account tool and, therefore, the user of SAPE must have some knowledge about taxes and tributary legislation, otherwise he will not be prone to use this tool correctly. Besides, SAPE is seen as a way to reduce costs and this might be seen as a way to sack people. Finally, it is important to understand that most of SAPE disadvantages are psychological, meaning that a strong motivation campaign to the advantages of this tool might be necessary to motivate workers to believe in the system [4].

3 Earned Value Analysis

EVA is a tool that controls and evaluates a project`s performance based on its costs, deadlines and progress. Its methodology consists in comparing what has been done or obtained (earned value) with what really has been done and what should have been done. EVA is a control tool that permits

to evaluate simultaneously and quantitatively the costs and delays in a specific time. It also makes a prediction on how much money will be spent at the end of the project.

EVA is a strong methodology due to three main aspects: it is a uniform unit of measure, it is consistent and it is a basis for cost performance analysis. Being a uniform unit of measure, it permits to combine and to compare the progress of completely different tasks. As far as being a consistent method, EVA allows that everybody inputs the information on how they are doing and realise if they are on schedule and the percentage of work done. Finally, EVA is a basis for cost performance analysis because it measures the quantity of work done in a consistent way, and compare unit costs, unit. In other words, it allows the comparison between physical progress of a project and its costs by using the same unit of measure [5].

After the presentation of the conceptual fundaments that support EVA, it is necessary to define three main variables in order to successively implement this tool. The first is the *Actual Cost of Work Performed* (ACWP). This is the money that has been spent to complete a task or, if the task is not finished, the money that has been spent so far. The second variable is the *Budgeted Cost of Work Performed (*BCWP). For completed tasks, this is the budget in the original task plan, regardless of the money actually spent on completing it. For unfinished tasks, it is the task budget multiplied by the percentage of completion so far. This might be seen as the achieved progress because this is the earning value for the cost that a task was expected to incur. Finally, the third and last primarily variable is the *Budgeted Cost of Work Scheduled* (BCWS). This is the money that is expected to be spent on the work that is expected to be accomplished by now. This means that it doesn't matter if the task is actually finished or on schedule. To do this, it is enough to look to the project plan and see what was planned to happen by now. For the project as a whole, these variables are the sum of all the ACWPs, BCWPs and BCWSs of all tasks. It is important to understand that these variables are functions of time. So, each time, EVA is used and these variables must be recalculated [6].

Considered the main variables for each task and for the project as a whole, earned value scores can be calculated. The first score is the *Schedule Variance* (SV):

$$SV = (BCWP - BCWS) / BCWS \tag{1}$$

The SV compares the achieved progress with the planned progress and divides it by the scheduled progress. This provides the percentage of deviation from what was planned.

The next secondary variable is the *Cost Variance* (CV):

$$CV = (BCWP - ACWP) / BCWP \tag{2}$$

The CV compares the actual cost to the planned cost for the work actually performed and divides it by the planned cost to provide us with the percentage of deviation from plan.
The *Schedule Performance Index* (SPI) is:

$$SPI = BCWP / BCWS \tag{3}$$

It is interesting to understand this variable meaning by using an example. If SPI = 0.9 this means that 90% of the time predicted in the schedule was actually converted into work. Consequently, there is a 10% loss of time and the work is behind the schedule. A SPI greater than 1 means that works are better than scheduled.

In the same line, the *Cost Performance Index* (CPI) is:

$$CPI = BCWP / ACWP \quad (4)$$

Presenting an example once again, if CPI = 0.9 it means that for every 1€ consumed, only 0,9€ are actually being converted into final product. Therefore, there is an overspending. A CPI greater than 1 means a spending of less money than what was predicted.
Now, for calculating the variables that are more meaningful and that permit more explicit results, let's begin to consider the variable *Estimate at Completion* (EAC):

$$EAC = ((BAC – BCWP) / CPI) + ACWP \quad (5)$$

The BAC, or *Budget at Completion*, is what is planned to spend at the end of the project. So, the EAC provides the work that has not been finished (BAC – BCWP), dividing it by the CPI and adding the ACWP, which is considered a sunk cost.
Finally the *Variation at Completion* (VAC):

$$VAC = BAC – EAC \quad (6)$$

It is easy to understand that this variable will show if one is spending more or less money than expected and will quantify it [7].

Although EVA is a powerful tool, its usage is not consensual. A survey was made to 400 professionals who worked in 180 projects and the results were that EVA was used by 41% of people but its value and popularity were very low. Trying to justify the low value proved by researchers, Thamhain [8] states that the little applicability, found as a result in the studies made, can be attributed to different barriers, either being internal or external.

To Wideman [9], a project of great importance requires a unit of planning and control with professionals capable of collecting the information and making the analysis of added value, making its applicability justifiable.

To Sparrow [10], the earned value analysis enables a supplementary value to the project because it offers a premature visibility of its results. So, it is possible to determine a tendency of costs and deadlines, while it is still possible to implement corrective actions.
On the contrary, West and McElroy [11] agree that EVA is an adequate tool for the generation of reports of work done, and not a managerial tool, since the control in real time of the project, using all parameters of analysis becomes unviable.

As can be seen, although it is consensual that EVA is a very powerful tool when applied correctly, even in the areas where it has been applied, it is still not very well defined [12].
According to Lukas [13], the top ten reasons why EVA does not work are:

a) no documented requirements,
b) incomplete requirements,
c) WBS not used or not accepted,
d) WBS incomplete,
e) plan not integrated,
f) schedule and/or budget incorrect,
g) change management not used or ineffective,
h) cost collection system inadequate,
i) incorrect progress and
j) management influence and or control.

Lukas [13] states that EVA is the most effective technique for providing information on project performance. It communicates scope, schedule and cost status information to project stakeholders. Properly used, earned value is a flexible process that provides timely information on the project "health". Effective use of EVA concepts can provide a competitive advantage in successfully delivering projects. Lukas [13] also mentions that if you have prepared your project plan properly, earned value analysis takes no additional effort to implement. The key is having complete requirements and a good project plan.

4 Case Study

Now, let's see how EVA works in a real life project. This project is called MSRCPF ("Monitoring System Radiation Coils Pyrolysis Furnace") and the objective is to implement an online monitoring system to the pyrolysis furnace. This will allow the detection of malfunctions and reduction of their consequences. This project also seeks to develop non destructive control techniques that will permit to determine the degradation state of these equipments. Such techniques will be applied during the programmed stops to avoid service failures that normally lead to production loss.

The main results that this project intends to achieve are:

a) increasing from four to six years the life of the equipment,
b) getting a reduction by 25% of the opportunity and maintenance costs that come from un-programmed stops,
c) getting a reduction by 29% of the opportunity costs that come from programmed stops,
d) getting a reduction of the systematic maintenance costs in 14%,
e) improving the availability of the equipment due to a reduction of the un-programmed stops and
f) getting a minimal energy consumption by doing a reduction in some operations.

As it is, the programmed stops have a cost of 813.600€ per year. This corresponds to a profit reduction of 520.000€ per day (opportunity cost). Looking at the equipment immobilization periods, this means that there is a loss of 3.094.000€ per year. Concerning the un-programmed stops, they have an opportunity cost of 2.080.000€ per year.

After a brief description of the project and after exposing the reasons for doing it, EVA can be put into action. Let`s assume a check point three months after the project begun. According to what was initially planned, the status of the project should be as shown in table1.

Tasks in Progress	Predicted Completion
1.A - Furnace Hardware Implementation Studies	100%
1.B - Acquiring Systems, Storage and Data Transfer Studies	25%
1.C - Coquefication Model Studies	50%

Table 1: Percentage of completion after 3 months of work - Tasks and predicted percentage

Knowing this and knowing some internal information about the project, the amount that should have been spent on each task and consequently the amount in the project as a whole can be determined three months after it has started.
The results are shown in table 2.

1.A	Technical Personnel	Worked Hours	Cost per Hour Worked	Total
	Fernando Afonso	45	34,74 €	1.563,30 €
	Carlos Jorge	30	25,13 €	753,90 €
	José António	120	26,31 €	3.157,20 €
	Rui João	55	26,31 €	1.447,05 €
				6.921,45 €
	Sub-Contracted Technicians	Worked Hours	Cost per Hour Worked	Total
	José Augusto	25	70,00 €	1.750,00 €

Table 2: Task 1. A. Predicted Costs

Task 1.A. will cost 8.671.45€. This task is finished so this is the amount of money that has to be spent on it after three months. Now let's see what happens with task 1.B., see table 3.

1.B	Technical Personnel	Worked Hours	Cost per Hour Worked	Total
	Fernando Afonso	285	34,74 €	9.900,90 €
	José António	212	26,31 €	5.577,72 €
	Mário Gonçalves	212	19,71 €	4.178,52 €
				19.657,14 €
	Sub-Contracted Technicians	Worked Hours	Cost per Hour Worked	Total
	José Augusto	40	70,00 €	2.800,00 €

Table 3: Task 1.B. Predicted Costs

In this case, although task 1.B. will cost 22.457,14€, only 25% is supposed to be completed. Therefore, it is supposed to have paid 5.614,29€ for this task so far. Finally, the status of the task 1.C. is shown in table 4.

1.C	Technical Personnel	Worked Hours	Cost per Hour Worked	Total
	José António	145	26,31€	3.814,95€
	Rui João	355	26,31€	9.340,05€
	Pedro Sousa	35	32,93€	1.152,55€
	Luís Sousa	140	21,99€	3.078,60€
	Celso Araújo	75	26,81€	2.010,75€
	Nuno Batista	130	16,96€	2.204,80€
	Manuel Soares	285	28,58€	6.435,30€
	Sandra dos Santos	305	15,58€	4.751,90€
				32.788,90€
	Technical Personnel	Worked Hours	Cost per Hour Worked	Total
	José Augusto	20	70,00 €	1.400,00€

Table 4: Task 1.C. Predicted Costs

Once again, when task 1.C. is completed, it is supposed to cost 34.188,90€ but now, three months after the beginning of the project, it has a 50% completion. So, the cost that now matters is 17.094,45€. How much each task is supposed to cost? and which tasks are supposed to be "on-going"? It is possible to determine the amount that should be spent: 31.380,19€.

This amount is now exactly known. Unfortunately, things almost never go the way desired. Let`s see what happens if the real scenario is the following:

- Task 1.A. was harder than it was initially thought to be and, to finish it on time, everybody had to work 5% more than expected.
- Due to technical problems, task 1.B. had a one month delay and so, in this third month checkpoint, it still was not started.
- Due to lack of equipment, task 1.C. was delayed by one week.

Now let`s see how EVA takes into account this new information. Firstly, task 1.A. is on time but it costs more than the expected. Now it is necessary to spend 9.105,02€ to do the same job that it was thought to cost 8.671,45€. Secondly, task 1.B. has not begun. So although it was not spent the expected 5.614,29€, this is not good because the task is delayed and it will be probably necessary to spend more money later. Finally, Task 1.C. is 46% completed and it was spent 15.669,91€ on it so far. As can be seen, after three months of work, 24.774,94€ were spent while it was expected to have spent 31.380,19€. To some people this might look like good news. But let's see what EVA configures about it.
The first thing to do is to determine the main variables for each task that has already begun or should have begun (see table 5).

1.A.	ACWP	9.105,02 €
	BCWP	8.671,45 €
	BCWS	8.671,45 €

Table 5: Task 1.A. main variables

Almost instinctively, ACWP (real cost) is determined and so is BCWS (predicted cost). The only tricky variable might be the BCWP. In this case, 1.A. is finished. So it does not matter if we have spent more or less than predicted or if we took more time or not to do it. When a task is completed, the BCWP is always equal to the BCWS. This shows that a good planning is absolutely imperative for the success of EVA. When a task is completed, it is possible to earn the value that it was initially established for that task. However, when a task is not finished this variable might cause some doubts. Basically, it is necessary to multiply the real percentage of completion of each task by the predicted costs of each one. In other words, it is necessary to determine how much the achieved work should have cost according to the first predictions. Now let`s analyze the development of the task 1.B. (table 6).

1.B.	ACWP	0,00 €
	BCWP	0,00 €
	BCWS	5.614,29 €

Table 6: Task 1.B. main variables

Once again, this comes almost naturally. ACWP and BCWP are obviously zero because the task has not started so there is no "work performed". Additionally, BCWS is what we have seen above, the

money that we were supposed to have spent by now. So far so good. This seems naturally understood, it is easy to implement it and the theory comes almost naturally. Finally, let's analyze the development of task 1.C. (table 7).

1.C.	ACWP	15.669,91 €
	BCWP	15.669,91 €
	BCWS	17.094,45 €

Table 7: Task 1.C. main variables

In this case BCWP is equal to ACWP because it is spent exactly what it should be according to the first predictions. On the other hand, BCWS is greater than BCWP showing that this task is delayed. But let´s now determine the secondary variables to put this information in a quantitative way.
For the task 1.A., and using the equations (1) and (3), it is realized that SV = 0 and SPI = 1. This means that this task is on time. On the other hand, when equations (2) and (4) are applied, it can be seen that CV has a negative value (showing that this task is overspending) and CPI = 0,95. This means that for every 1€ spent, only 0,95€ are actually being converted into work.
When looking at 1.B., it seems that it does not make much sense to calculate anything because the task is yet to start. But notice what happens when equation (1) is used. The result shows SV = -1, meaning that not only this task is delayed but also that it should have started but it did not.
Finally, when looking at 1.C., and applying equations (1) and (3), it is possible to see that SV = -0,083 (once again showing that this task is delayed) and SPI = 0,92 (meaning that only 92% of the time predicted was actually converted in work). Therefore, there was an 8% available time loss. On the other hand, equations (2) and (4) show a CV = 0 and CPI = 1, meaning that the amount spent is exactly what was predicted.
Now by knowing what is going on with each one of the tasks that should have begun by now, let`s look at the project as a whole and let's see if should exist any concerns there. As told before, to see the overall project, it is necessary to sum all the main variables. In this case, the results are in table 8.

Project - Global Overview	ACWP	24.774,94 €
	BCWP	24.341,36 €
	BCWS	31.380,19 €

Table 8: Project Main Variables

Just by looking at these numbers it is possible to understand that things are not doing great. Let`s quantify it. When (1) is calculated, SV = -0,22. Accordingly, (3) presents a value for SPI of 0,78. So this project is definitely behind the schedule. On the other hand, when (2) is applied, it shows a CV = -0,02. Again, (4) shows us an CPI = 0,98 which means that this project is overspending. These are really bad news. If things continue as they are, this project will not only take more time to finish but also will cost more money. Let's now quantify how much more. From the first predictions it is possible to know that this project has a budgeted cost (BAC) of 1.082.348,88€. When using (5), it can be realized that the EAC is 1.101.627,86€. Using these two variables, (6) can be calculated and it gives a VAC of -19.278,98€. This means that if the project keeps this tendency, when it is over, it will have cost more 19.278,98€ than expected. To do a time analysis a Gant graph should be built.

EVA does not take into account the critical path nor the slacks of each activity so it is impossible to know if a delay now will affect the final date for completion. Nevertheless, this deficiency is easily beaten and EVA`s advantages in cost control might be the difference between success and failure.

5 Conclusion and Limitations of the Study

First of all, it is important to say that the scenarios built for the case study were made up to provide the results wanted because there is no information about the real status of the project.
In this case study, for task 1.A., it was said that everybody worked 5% more. In real life these situations do not happen. One worker might work 10 hours more than expected, other one 7 hours less but it is always hard to realize the money that it is actually being spent. This is definitely a key success factor for EVA. A good informatics system and the habit of reporting progress and costs are the support that EVA requires.
Another thing to take to keep in mind is that, to simplify the math, we did not took into account the overheads nor equipment purchases. Basically, the results shown are just from tasks and their costs. Normally, in a project there are other factors that might influence the performances.
Another key success factor is the quality of the predictions and planning. As it was seen, the earned value at the end of a task was always what was predicted. Time allows that a project plan can be thought carefully.
It was also seen that EVA follows a standard or a tendency. So, spending an extra 1.000€ now, it does not mean that, at the end of the project, it will cost just 1000€ more. Probably, if you are overspending now you will keep on overspending and at the end of the project you will pay not 1.000€ but 20.000€ more than expected.
EVA may represent also a great control tool and it can be used many times to this aim. If something is wrong, EVA will definitely warn you on time so you can take corrective actions.
Finally, every time you use EVA you have to update your schedule. This might not look much important but it is and for that it is enough to see the following example. Imagine that a project lasts 12 months, you use the EVA model after the first one and you realize that the project is delayed. You take the actions needed but you forget to update your schedule. When you do another check point control, let`s say, at the end of the second month, even if there were not any more delays, the effects of that first delay will still be visible and you do not understand why. In other words, after that first delay, task X would finish two weeks later than first planned and if you do not update your schedule, every time you use EVA you will think that X is delayed another two weeks, when in fact it is already a consequence from the first delay.
To sum up, Earned Value Analysis is a very powerful project control/management tool. It might be expensive and it might be difficult to put it into work at the beginning. But if you have a good informatics system that collects your costs, if you have experienced personnel in planning and if you create your projects knowing that the EVA model will be applied, meaning that your project is formed by small and manageable tasks, than Earned Value Analysis may give you an edge and an advantage that can make all the difference.

References

[1] ESTEVES, J. M., SANTOS, A. A. e CARVALHO, J. Á., (2000), “O Ciclo de Vida dos Custos dos Sistemas ERP”, VII Congresso Brasileiro de Custos.

[2] GUSKUMA, E. A., (1999), "ERP: A Solução Final?", Administração de Informática.
[3] JESUS, R. G. e OLIVEIRA, M. O. F., (2006), "Implantação de Sistemas ERP: tecnologia e pessoas na implementação do SAP R/3", Revista de Gestão da Tecnologia e Sistemas de Informação.
[4] COLANGELO, L. F., (2001), "Implantação de Sistemas E.R.P. – Um enfoque de longo prazo", São Paulo, Atlas.
[5] WILKENS, T. T., (1999), "Earned Value, Clear and Simple", Los Angeles County Metropolitan Transportation Authority.
[6] CESARONE, J., (2007), "Project Management by the numbers – How Earned Value Analysis can keep you on track".
[7] GIACOMETTI, R. A., SILVA, C. E. S., SOUZA, H. J. C., MARINS, F. A. S. and SILVA E. R. S., (2007), "Aplicação do Earned Value em projectos complexos – um estudo de caso na EMBRAER".
[8] THAMHAIN, H. J., (1998), "Integrating Project Management Tools with the Project Team", Long Beach: 29th Annual Project Management Institute Seminars and Symposium.
[9] WIDEMAN, R. M., (1999), "Cost Control of Capital Projects and the Project Cost Management Systems Requirements", 2nd ed., Vancouver: AEW Services e Bi Tech Publishers.
[10] SPARROW, H., (2000), "EVM = Earned Value Management Results in Early Visibility and Management Opportunities", Houston: 31st Annual Project Management Institute Seminars and Symposium.
[11] WEST, S. M. e McELROY, S., (2001), "EVMS: A Managerial Tool vs. a Reporting Tool", Nashville: 32th Annual Project Management Institute Seminars and Symposium.
[12] VARGAS, R. V., (2003), "Earned Value Analysis in the Control of Projects: Success or Failure", AACE International Transactions.
[13] LUKAS, J. A., (2008), "Earned Value Analysis – Why it doesn´t work", AACE International Transactions.

Current addresses

Maria Isabel C. Pedro, Professor Auxiliar
CEGIST/IST - Portugal
Av. Rovisco Pais,LISBOA, Portugal
Phone: +(351) 214233267
e-mail: ipedro@ist.utl.pt

João Pereira, Masters Student
Av. Rovisco Pais, LISBOA, Portugal
Phone: +(351) 214233267
e-mail: jbleitaopereira@gmail.com

José António Filipe, Professor Auxiliar
ISCTE – IUL
UNIDE – IUL, Av. Forças Armadas 1649-026 Lisboa, Portugal
Tel.+351 217 903 000
e-mail: jose.filipe@iscte.pt

Manuel Alberto M. Ferreira, Professor Catedrático
ISCTE – IUL
UNIDE – IUL, Av. Forças Armadas 1649-026 Lisboa, Portugal
Tel. +351 217 903 000
e-mail: manuel.ferreira@iscte.pt

Journal of Applied Mathematics

volume IV (2011), number III

LAW OF IMPORTATION FOR GENERATED FUZZY IMPLICATORS

BIBA Vladislav, (CZ)

Abstract. In the classical logic, term $A \Rightarrow (B \Rightarrow C)$ is equivalent with $(A \wedge B) \Rightarrow C$. However, in the fuzzy logic it is not true for all implicators and t-norms. Mentioned equivalence is know as the Law of importation. Recently, weaker form of this rule were proposed by Massanet and Torrens in [8], where more general function is used instead of t-norm.

Key words and phrases. Fuzzy implicator, t-norm, law of importation, generator function

Mathematics Subject Classification. Primary 60A05, 08A72; Secondary 28E10.

1 Preliminaries

We briefly recall definitions and properties of the most important connectives of fuzzy logic.

Definition 1.1 *A unary operator* $N : [0,1] \to [0,1]$ *is called a fuzzy negation if, for any* $x, y \in [0,1]$,

- $x < y \Rightarrow N(y) \leq N(x)$,
- $N(0) = 1, N(1) = 0$.

The negator N is called a *strict negator* if and only if the mapping N is continuous and strictly decreasing. A strict negator is called *strong* if it is an involution, i.e. $N(N(x) = x \; \forall x \in [0,1]$. A *dual negator* based on N is given by $N^d(x) = 1 - N(1-x)$.

Definition 1.2 *A non-decreasing mapping $C : [0,1]^2 \to [0,1]$ is called a conjunctor if, for any $x, y \in [0,1]$, it holds*

- $C(x,y) = 0 \quad$ *whenever* $\quad x = 0 \quad$ *or* $\quad y = 0$,
- $C(1,1) = 1$.

Commonly used conjunctors in fuzzy logic are the triangular norms.

Definition 1.3 *A triangular norm (t-norm for short) is a binary operation on the unit interval $[0,1]$, i.e., a function $T : [0,1]^2 \to [0,1]$ such that for all $x, y, z \in [0,1]$, the following four axioms are satisfied:*
(T1) Commutativity $T(x,y) = T(y,x)$,
(T2) Associativity $T(x,T(y,z)) = T(T(x,y),z)$,
(T3) Monotonicity $T(x,y) \leq T(x,z)$ *whenever* $y \leq z$,
(T4) Boundary Condition $T(x,1) = x$.

We recall some important properties of t-norms which we will use:

Definition 1.4 *[6] A t-norm T is continuous if for all convergent sequences $\{x_n\}_{n\in N}$, $\{y_n\}_{n\in N}$ we have*

$$T(\lim_{n\to\infty} x_n, \lim_{n\to\infty} y_n) = \lim_{n\to\infty} T(x_n, y_n).$$

A t-norm T is left-continuous if for each $y \in [0,1]$ and for all non-decreasing sequences $\{x_n\}_{n\in N}$ we have

$$\lim_{n\to\infty} T(x_n, y) = T(\lim_{n\to\infty} x_n, y).$$

Proposition 1.5 *A t-norm T is left-continuous if and only if it is left-continuous in its first component, i.e., if for each $y \in [0,1]$ and for each non - decreasing sequence $(x_n)_{n\in N} \in [0,1]^N$ we have*

$$\sup_{n\in N} T(x_n, y) = T(\sup_{n\in N} x_n, y).$$

Now, we will turn our attention to the Archimedean property. We recall another definition of Archimedean t-norms, which is equivalent with the classical one.

Proposition 1.6 *A t-norm T is Archimedean if and only if for each $x \in]0,1[$ we have*

$$\lim_{n\to\infty} \underbrace{T(x, ..., x)}_{n\text{-times}} = 0.$$

Following important algebraic property is *strict monotonicity.*

Proposition 1.7 *A t-norm T is strictly monotone if and only if the cancelation law holds, i.e., if $T(x,y) = T(x,z)$ and $x > 0$ imply $y = z$.*

Remark 1.8 *Note that the dual operator to the conjunctor C, defined by $D(x,y) = 1 - C(1-x, 1-y)$ is called the disjunctor. Equivalently, a disjunctor can be defined as a non-decreasing mapping $D : [0,1]^2 \to [0,1]$ such that $D(x,y) = 1$ whenever $x = 1$ or $y = 1$ and $D(0,0) = 0$. Commonly used disjunctors in fuzzy logic are the triangular conorms. A triangular conorm (also called a t-conorm) is a binary operation S on the unit interval $[0,1]$ which, for all $x, y, z \in [0,1]$, satisfies $(T1) - (T3)$ and $(S4)$ $S(x,0) = x$. The original definition of t-conorms given in [10] is completely equivalent to the previous axiomatic definition, where the t-conorm is based on a given t-norm T by formula*

$$S(x,y) = 1 - T(1-x, 1-y).$$

For more information, see [7].

In the literature, we can find several different definitions of fuzzy implications. In this paper we will use the following one, which is equivalent to the definition introduced by Fodor and Roubens in [4]. The readers can obtain additional informations by reading [1] and [9].

Definition 1.9 *A function $I : [0,1]^2 \to [0,1]$ is called a fuzzy implicator if it satisfies the following conditions:*

(I1) I is decreasing in its first variable,

(I2) I is increasing in its second variable,

(I3) $I(1,0) = 0$, $I(0,0) = I(1,1) = 1$.

We recall definitions of some important properties of implicators which we will investigate.

Definition 1.10 *A fuzzy implicator $I : [0,1]^2 \to [0,1]$ satisfies:*

(NP) the left neutrality property, or is called left neutral, if

$$I(1,y) = y; \qquad y \in [0,1],$$

(EP) the exchange principle if

$$I(x, I(y,z)) = I(y, I(x,z)) \text{ for all } x, y, z \in [0,1],$$

(IP) the identity principle if

$$I(x,x) = 1; \quad x \in [0,1],$$

(OP) the ordering property if

$$x \leq y \iff I(x,y) = 1; \quad x, y \in [0,1],$$

(CP) the contrapositive symmetry with respect to a given negator N if

$$I(x,y) = I(N(y), N(x)); \quad x, y \in [0,1],$$

(LI) the law of importation with a t-norm T if

$$I(T(x,y),z) = I(x,I(y,z)); \quad x,y \in [0,1],$$

(WLI) the weak law of importation with a given function F if

$$I(F(x,y),z) = I(x,I(y,z)); \quad x,y,z \in [0,1].$$

Definition 1.11 *Let $I : [0,1]^2 \to [0,1]$ be a fuzzy implicator. The function N_I defined by $N_I(x) = I(x,0)$ for all $x \in [0,1]$, is called the natural negator of I.*

Our constructions of implicators will make use extensions of the classical inverse of function. One way of extending is described in next definitions.

Definition 1.12 *Let $\varphi : [0,1] \to [0,\infty]$ be a non-decreasing function. The function $\varphi^{(-1)}$ which is defined by*

$$\varphi^{(-1)}(x) = \sup\{z \in [0,1]; \varphi(z) < x\},$$

is called the pseudo-inverse of the function φ, with the convention $\sup \emptyset = 0$.

Definition 1.13 *Let $f : [0,1] \to [0,\infty]$ be a non-increasing function. The function $f^{(-1)}$ which is defined by*

$$f^{(-1)}(x) = \sup\{z \in [0,1]; f(z) > x\},$$

is called the pseudo-inverse of the function f, with the convention $\sup \emptyset = 0$.

Lemma 1.14 *[3] Let $N : [0,1] \to [0,1]$ be a negator. Then $N^{(-1)}$ is a negator if and only if*

$$N(x) = 0 \quad \Leftrightarrow \quad x = 1. \tag{1}$$

2 Generated implicators

It is well-known that it is possible to generate t-norms from one variable functions. It means it is enough to consider one variable function instead of two variable function. In our works [3, 5] and [11] we defined several types of these so-called generated implicators. The first possibility use a strictly decreasing function f.

Theorem 2.1 *[5] Let $f : [0,1] \to [0,\infty]$ be a strictly decreasing function such that $f(1) = 0$. Then the function $I_f(x,y) : [0,1]^2 \to [0,1]$ which is given by*

$$I_f(x,y) = \begin{cases} 1 & \text{if } x \le y, \\ f^{(-1)}(f(y^+) - f(x)) & \text{otherwise,} \end{cases} \tag{2}$$

where $f(y^+) = \lim\limits_{y \to y^+} f(y)$ and $f(1^+) = f(1)$ is a fuzzy implicator.

On the other hand for strictly increasing functions g, [11] gives operator I^g.

Theorem 2.2 *[11] Let $g : [0,1] \to [0,\infty]$ be a strictly increasing function such that $g(0) = 0$. Then the function $I^g(x,y) : [0,1]^2 \to [0,1]$ which is given by*

$$I^g(x,y) = g^{(-1)}(g(1-x) + g(y)), \tag{3}$$

is an fuzzy implicator.

The implicator I^g can be generalized. This generalization is based on replacing standard negator by arbitrary one.

Theorem 2.3 *[11] Let $g : [0,1] \to [0,\infty]$ be a strictly increasing function such that $g(0) = 0$ and N be a fuzzy negator. Then the function I_N^g:*

$$I_N^g(x,y) = g^{(-1)}(g(N(x)) + g(y)), \tag{4}$$

is an implicator.

If we compose a strictly decreasing function f with a negator N then $g(x) = f(N(x))$ is again an increasing function (though not necessarily strictly increasing). We can apply such a function g to formula (4) and have another possibility how to generate implicators.

Theorem 2.4 *[3] Let $f : [0,1] \to [0,\infty]$ be a strictly decreasing function with $f(1) = 0$, and $N : [0,1] \to [0,1]$ be a negator such that formula (1) is fulfilled for N. Then the function $I_f^{(N,N^{(-1)})} : [0,1]^2 \to [0,1]$ defined by*

$$I_f^{(N,N^{(-1)})}(x,y) = N^{(-1)}\left(f^{(-1)}\left(f(x) + f(N(y))\right)\right) \tag{5}$$

is an implicator.

If we consider formula (5), we can see that $N^{(-1)}$ is just another negator. Really, it might be replaced by an arbitrary negator. However, we would like to keep the procedure of generating of implicators as simple as possible. Still, there are at least two negators (in general different from $N^{(-1)}$) which are related to N. Namely, N itself and N^d. Hence we have the following two possibilities of generating of implicators.

Theorem 2.5 *[3] Let $f : [0,1] \to [0,\infty]$ be a strictly decreasing function with $f(1) = 0$ and $N : [0,1] \to [0,1]$ be a negator. Then function $I_f^N : [0,1]^2 \to [0,1]$ defined by*

$$I_f^N(x,y) = N\left(f^{(-1)}\left(f(x) + f(N(y))\right)\right), \tag{6}$$

and function $I_f^{(N,N^d)} : [0,1]^2 \to [0,1]$ defined by

$$I_f^{(N,N^d)}(x,y) = N^d\left(f^{(-1)}\left(f(x) + f(N(y))\right)\right), \tag{7}$$

are implicators.

3 The (weak) law of importation

In the last section we investigate the *law of importation* and the *weak law of importation* for classes of defined generated implicators. In the first example, we deal with implicator I_f given by continuous bounded function:

Example 3.1 *Let $f_1(x) = 1 - x$, for implicator I_{f_1} we get*

$$I_{f_1}(x,y) = \begin{cases} 1 & \text{if } x \le y, \\ 1 - x + y & \text{otherwise.} \end{cases}$$

It is easy to show that operators $I_{f_1}(x, I_{f_1}(y,z))$ and $I_{f_1}(T(x,y),z)$ are:

$$I_{f_1}(x, I_{f_1}(y,z)) = \begin{cases} 1 & \text{if } y \le z \vee x \le 1 - y + z, \\ 2 - x - y + z & \text{otherwise,} \end{cases}$$

$$I_{f_1}(T(x,y),z) = \begin{cases} 1 & \text{if } T(x,y) \le z, \\ 1 - T(x,y) + z & \text{otherwise.} \end{cases}$$

Now we can see that I_f satisfies LI with t-norm $T(x,y) = x + y - 1$.

The t-norm in the previous example is actually an archimedean t-norm with generator f. In the following example we deal with different situation, where f is not bounded function:

Example 3.2 *Let $f_2(x) = -\ln x$, implicator I_{f_2} is given by*

$$I_{f_2}(x,y) = \begin{cases} 1 & \text{if } x \le y, \\ \frac{y}{x} & \text{otherwise.} \end{cases}$$

For operators $I_{f_2}(x, I_{f_2}(y,z))$ and $I_{f_2}(T(x,y),z)$ we get:

$$I_{f_2}(x, I_{f_2}(y,z)) = \begin{cases} 1 & \text{if } y \le z \vee x \le \frac{z}{y}, \\ \frac{z}{x \cdot y} & \text{otherwise,} \end{cases}$$

$$I_{f_2}(T(x,y),z) = \begin{cases} 1 & \text{if } T(x,y) \le z, \\ \frac{z}{T(x,y)} & \text{otherwise.} \end{cases}$$

It is obvious that I_{f_2} satisfies LI with t-norm $T(x,y) = x \cdot y$.
Another implicator satisfying LI with the same t-norm is I^g implicator generated by function $g(x) = -\ln(1-x)$. Implicator I^g is given by

$$I^g(x,y) = 1 - x + x \cdot y.$$

The following proposition is generalization of these examples.

Proposition 3.3 *Let $f : [0,1] \to [0,\infty]$ be a continuous strictly decreasing function such that $f(1) = 0$ and $f(0) = 1$. Then the implicator I_f satisfies the law of importation with t-norm*

$$T(x,y) = f^{(-1)}(f(x) + f(y)).$$

Proof. Let f be function as described. By definition we have

$$I_f(y,z) = \begin{cases} 1 & y \le z, \\ f^{(-1)}\left(f(z^+) - f(y)\right) & \text{otherwise.} \end{cases}$$

Fact, that f is continuous (and strictly decreasing) allows a simplification of previous statement:

$$I_f(y,z) = \begin{cases} 1 & y \le z, \\ f^{-1}\left(f(z) - f(y)\right) & \text{otherwise.} \end{cases} \tag{8}$$

Using this fact, we can write $I_f(x, I_f(y,z))$ as

$$I_f(x, I_f(y,z)) = \begin{cases} 1 & x \le I_f(y,z), \\ f^{-1}(f(z) - f(y) - f(x)) & x > I_f(y,z). \end{cases} \tag{9}$$

As f is continuous and strictly decreasing, we have $\forall t \in [0,\infty];\ f^{(-1)}(t) = f^{-1}\left(\min(t, f(0))\right)$. For any $x, y \in [0,1]^2$, both $f(x)$ and $f(y)$ are non-negative. It means that T can be expressed as

$$T(x,y) = f^{-1}\left(\min(f(x) + f(y), f(0))\right). \tag{10}$$

Using simplified formula (8) we get

$$I_f(T(x,y),z) = \begin{cases} 1 & T(x,y) \le z, \\ f^{-1}(f(z) - f(T(x,y))) & \text{otherwise.} \end{cases}$$

- The condition $T(x,y) \le z$ means that either $T(x,y) = 0$ or $f^{-1}(f(x)+f(y)) \le z$. In both cases $f(x) + f(y) \ge f(z)$, i.e. $f(x) \ge f(z) - f(y)$. Therefore we get $x \le I_f(y,z)$ (whether $f(z) > f(y)$ or not) and subsequently $I_f(x, I_f(y,z)) = 1$.
- In case that $T(x,y) > z$, we get $f(x) + f(y) < f(z)$. It means that $y > z$ and $x > f^{-1}(f(z) - f(y))$ as well. Subsequently $I_f(x, I_f(y,z)) = f^{-1}(f(z) - f(y) - f(x))$.

In the previous proposition we demand f to be continuous function. However, there are also not continuous functions which generates I_f implicator that satisfies law of importation:

Example 3.4 *Let $f : [0,1] \to [0,1]$ be a function given by*

$$f(x) = \begin{cases} 1 - \frac{2}{3}x & \text{if } x < 1, \\ 0 & \text{if } x = 1. \end{cases}$$

Implicator I_f and corresponding mapping $I_f(x, I_f(y, z))$ are given by

$$I_f(x,y) = \begin{cases} 1 & \text{if } x \leq y, \\ y & \text{if } x = 1, \\ \min(\frac{3}{2} - x + y, 1) & \text{otherwise,} \end{cases}$$

$$I_f(x, I_f(y,z)) = \begin{cases} 1 & \text{if } y \leq z \text{ or } x \leq I_f(y,z), \\ z & \text{if } x = y = 1, \\ \min(\frac{3}{2} - x + z) & \text{if } x < 1 \wedge y = 1, \\ \min(\frac{3}{2} - y + z) & \text{if } x = 1 \wedge y < 1. \end{cases}$$

Implicator I_f satisfies the law of importation with t-norm T defined as:

$$T(x,y) = \begin{cases} \min(x,y) & \text{if } \max(x,y) = 1, \\ \max(x + y - \frac{3}{2}, 0) & \text{otherwise.} \end{cases}$$

Proof is done examining each possibility separately.

- *Implicator I_f satisfies the law of importation with all t-norms T^* such that $T^* \leq T$.*
- *Moreover, I_f also satisfies the weak law of importation with two-variable functions F with two properties: $F \leq T$ and $F(x,1) = F(1,x) = x$ for all $x \in [0,1]$.*

Propositions similar to proposition 3.3, considering only continuous g and strict negators holds also for I^g, I^g_N. Proofs to these facts are similar to previous proof, Therefore we will focus to more general case, where negator is not necessary strict:

Proposition 3.5 *Let $g : [0,1] \to [0,\infty]$ be a continuous strictly increasing function such that $g(0) = 0$ and let $N : [0,1] \to [0,1]$ be a continuous negator satisfying equation 1. Then the implicator I^g_N holds weak law of importation with function*

$$F(x,y) = N^{(-1)}\left(g^{(-1)}(g(N(x)) + g(N(y)))\right).$$

Proof. Let g and N be a function and negator as mentioned in the proposition. Since $I^g_N(y,z) = g^{(-1)}\left(g(N(y)) + g(z)\right)$, it is easy to show, that

$$I^g_N(x, I^g_N(y,z)) = g^{(-1)}\left(g(N(x)) + g(N(y)) + g(z)\right). \tag{11}$$

If $I^g_N(y,z) < 1$, it is trivial. If $I^g_N(y,z) = 1$, then $I^g_N(x, I^g_N(y,z)) = I^g_N(x,1) = 1$ as well, i.e. previous equality is correct.
Let F be the function as was defined in the proposition. Continuous negator N implies that $N(N^{(-1)}(t)) = t$ for all $t \in [0,1]$ and continuous and strictly increasing g means that $g^{(-1)}(g(t)) = \min(t,1)$ for all $t \geq 0$. Using this two facts, we get

$$g(N(F(x,y))) = \min\left(g(N(x)) + g(N(y)), g(1)\right), \tag{12}$$

$$I_N^g(F(x,y),z) = g^{(-1)}\left(\min(g(N(x)) + g(N(y)), g(1)) + g(z)\right).$$

We can remove minimum from the upper term, since any $t \geq g(1)$ leads to $g^{(-1)}(t) = 1$. Therefore we get

$$I_N^g(F(x,y),z) = g^{(-1)}\left(g(N(x)) + g(N(y)) + g(z)\right),$$

which is the same term as in equation 11.

Remark 3.6 *Function from previous proposition possess some additional properties:*

- *F is a commutative conjunctor, which is obvious. Moreover, it is associative as well: Using equality 12, we get*

  $$F(F(x,y),z) = N^{(-1)}\left(g^{(-1)}(g(N(x)) + g(N(y)) + g(N(z))\right)$$

  and the same is true for $F(x,F(y,z))$.

- *In case of strict negator, F is a t-norm since $F(x,1) = N^{(-1)}(N(x)) = x$.*

- *If N is strict, then I_N^g satisfies the law of importation only with this t-norm: Take $z = 0$, then*

  $$I_N^g(x, I_N^g(y,z)) = g^{(-1)}(g(N(x)) + g(N(y)),$$

  $$I_N^g(F(x,y),z) = N(F(x,y)).$$

  Let I_N^g holds the law of importation with F, then

  $$F(x,y) = N^{-1}(g^{(-1)}(g(N(x)) + g(N(y)))).$$

There are also not continuous negators N that generates I_N^g implicator which still satisfies *LI.* The last example shows one such implicator and corresponding t-norm:

Example 3.7 *Let $g(x) = x$ and $N(x) = 1 - \frac{2}{3}x$ if $x < 1, N(0) = 1$, then implicator I_N^g is given by*

$$I_N^g(x,y) = \begin{cases} \min(1 - \frac{2}{3}x + y, 1) & \text{if } x < 1, \\ y & \text{if } x = 1. \end{cases}$$

This implicator satisfies the law of importation with t-norm

$$T(x,y) = \begin{cases} \max(x + y - \frac{3}{2}, 0) & \text{if } x, y < 1, \\ \min(x,y) & \text{if } \max(x,y) = 1. \end{cases}$$

Note, that both implicator and t-norm are not continuous.

References

[1] BACZYŇSKI, M., BALASUBRAMANIAM, J.: *Fuzzy implications* (Studies in Fuzziness and Soft Computing, Vol. 231), Springer, Berlin (2008)

[2] BIBA, V., HLINĚNÁ, D., KALINA, M., KRÁĽ, P.: *Generated fuzzy implications and known classes of implications* (to appear)

[3] BIBA, V., HLINĚNÁ, D., KALINA, M., KRÁ Ľ, P.: *I-fuzzy equivalences and I-partitions* 17th East-West Fuzzy Colloquium Conference Proceedings, Zittau 2010

[4] FODOR, J. C., ROUBENS, M.: *Fuzzy Preference Modeling and Multicriteria Decision Support* Kluwer Academic Publishers, Dordrecht 1994.

[5] HLINĚNÁ, D., BIBA, V.: *Generated fuzzy implications and known classes of implications* In: Acta Universitatis Matthiae Belii ser. Mathematics. 1. Banská Bystrica: Matej Bel University, 2010, 25–34.

[6] KLEMENT, E. P., MESIAR, R.: *Logical, algebraic, analytic and probabilistic aspects of triangular norms*, Elsevier, Amsterdam 2005, ISBN 0-444-51814-2

[7] KLEMENT, E. P., MESIAR, R., PAP, E.: *Triangular Norms* Kluwer, Dordrecht 2000.

[8] MASSANET, S., TORRENS, J.: *Fuzzy implications and Weak law of importation* IFSA-EUSFLAT, 2009

[9] MAS, M., MONSERRAT, M., TORRENS, J., TRILLAS, E.: *A survey on fuzzy implication functions* IEEE Transactions on Fuzzy Systems 15 (6) 1107–1121, (2007).

[10] SCHWEIZER, B., SKLAR, A.: *Probabilistic Metric Spaces* North Holland, New York 1983

[11] SMUTNÁ, D.:*On many valued conjunctions and implications* Journal of Electrical Engineering, 10/s, vol. 50, 1999, 8–10

Current address

Biba Vladislav, Mgr.
Department of Mathematics,
Faculty of Electrical Engineering and Communication,
Brno University of Technology,
Technická 8, Brno Czech Republic
e-mail: xbibav00@stud.feec.vutbr.cz

Journal of Applied Mathematics

volume IV (2011), number III

INVOLVING FUZZY ORDER IN THE DEFINITION OF MONOTONICITY FOR AGGREGATION FUNCTION

GRIGORENKO Olga, (LV)

Abstract. In this paper we introduce a fuzzy order relation in the definition of monotonicity for aggregation function. We use the fuzzy order relation to define the degree of monotonicity, which takes values from the unit interval and is equal to 1 for a monotone function with respect to a crisp order relation. Further we illustrate this definition by examples and study the properties of aggregation functions which have a degree of monotonicity equal to 1. We also introduce α levels in the definition of the degree of monotonicity and thereby obtain a more general and consistent theory.

Key words and phrases. aggregation function, fuzzy order relation, monotonicity.

Mathematics Subject Classification. Primary 03E72, 06A06; Secondary 62C86.

1 Introduction

The aim of this work is to introduce a fuzzy order relation in aggregation process, namely, to use a fuzzy order relation instead of the crisp order relation in the definition of monotonicity. Recall that an aggregation function is a mapping satisfying boundary conditions and the condition of monotonicity. In our work we focus only on the condition of monotonicity.

The next two examples illustrate our inspiration which led to the present research:

Let us observe first the aggregation which is illustrated by the Table 1. According to the definition of aggregation function, the mapping defined above is an aggregation function (obviously, the monotonicity condition is fulfilled). But if we have a more attentive look at this example and consider the aggregation results of alternatives a_1 and a_2 we will find out that the results are intuitively wrong. We see that the first attribute of the alternative a_2 is less than the first attribute of the alternative a_1; the second attribute of a_2 is greater than the

Alternat.	First attrib.	Second attrib.	Aggreg. result
a_1	0.1	0.3	0.2
a_2	0.01	0.31	0.29
a_3	0.2	0.4	0.3

Table 1: Motivate example 1.

Alternat.	First attrib.	Second attrib.	Aggreg. result
a_1	0.1	0.3	0.2
a_2	0.2	0.399	0.301
a_3	0.2	0.4	0.3

Table 2: Motivate example 2.

second attribute of a_1 only a little (is equal to the second attribute of a_1 "in a fuzzy sense"), so, intuitively, we expect that if even the aggregation result of the alternative a_2 is greater than the aggregation result of the alternative a_1, then it should be greater only a little. But in our example aggregation result of the alternative a_1 is less than the aggregation result of the alternative a_2 and we have a big difference between the aggregation results. To avoid such situations we involve fuzzy order relation in order to define the degree of monotonicity.

Another possible situation where fuzzy order could help is the problem when we have small mistakes in aggregation, what is actually illustrated by the "Motivate example 2". The small variation of data could change the result drastically. For the question "Is this an aggregation function?" we could only answer "Yes" or "No", thus a very small mistake or destroy of data could change the answer from "Yes" to "No". Let us observe Example 2:
In this case it is not an aggregation function since the monotonicity condition for the pair (a_2, a_3) is not fulfilled. But the second row could be realized just as the damaged third one. Thus, in this case it would be useful to delete the second row or to involve the degree of monotonicity which not only says "it is an aggregation function" or "it is not an aggregation function" but gives us the degree to which a mapping is a monotone function.

Thus the aim of this work is to define the degree of monotonicity, to observe illustrating examples and to study the properties of the degree of monotonicity. The problem of the use of fuzzy order relation in the context of monotonicity of aggregation process was first considered in our paper [3]. In this work we continue research in this area.

2 Preliminaries

In the sequel we will use the basic notions and properties of t-norms. For this information we refer the reader to [4]. For the details on aggregation functions we refer the reader to [2].

We continue with an overview of basic definitions and results on fuzzy relations which will be important for our further research.

Definition 2.1 *(see e.g. [1]) A fuzzy binary relation E on a set X is called fuzzy equivalence relation with respect to a t-norm T, if and only if the following three axioms are fulfilled for all $x, y, z \in X$:*
1) $E(x, x) = 1$ reflexivity;
2) $E(x, y) = E(y, x)$ symmetry;
3) $T(E(x, y), E(y, z)) \leq E(x, z)$ T-transitivity.

The following result establishes principles of construction of fuzzy equivalence relations from pseudo-metrics.

Theorem 2.2 *(see e.g. [1]) Let T be a continuous Archimedean t-norm T with an additive generator t. For any pseudo-metric d, the mapping $E_d(x,y) = t^{(-1)}(\min(d(x,y), t(0)))$ is a T-equivalence.*

Example 2.3 *Let us consider the set of real numbers $X = \mathbb{R}$ and metric $d(x,y) = |x-y|$ on it. Taking into account that $t_L(x) = 1 - x$ is an additive generator of T_L (Łukasievicz t-norm) and that $t_P(x) = -ln(x)$ is an additive generator of T_P (product t-norm), we obtain two fuzzy equivalence relations:*

$$E_L(x,y) = \max(1 - |x-y|, 0);$$
$$E_P(x,y) = e^{-|x-y|}.$$

Definition 2.4 *[1] Let $\preceq$ be a crisp order on X and let E be a fuzzy equivalence relation on X. E is called compatible with $\preceq$ if and only if the following implication holds for all $x, y, z \in X$: $x \preceq y \preceq z \Rightarrow (E(x,z) \leq E(y,z)$ and $E(x,z) \leq E(x,y))$.*

Remark 2.5 *Let X be the set of real numbers and $\leq$ be a linear order on it. Then, for a fixed element x_0, $E(x,x_0)$ is non-decreasing in the interval $[-\infty, x_0]$ and non-increasing in the interval $[x_0, \infty]$, where E is a fuzzy equivalence relation which is compatible with $\leq$.*

Definition 2.6 *(see e.g. [1]) A T-transitive fuzzy relation $L : X^2 \to [0,1]$ is called fuzzy order relation with respect to a t-norm T and a T-equivalence E, if and only if it additionally fulfills the following two axioms for all $x, y \in X$:*

1. *$L(x,y) \geq E(x,y)$ E-reflexivity;*
2. *$T(L(x,y), L(y,x)) \leq E(x,y)$ T-E-antisymmetry.*

A fuzzy order L is called strongly linear if and only if $\forall x, y \in X : \ \max(L(x,y), L(y,x)) = 1$.

The following theorem states that strongly linear fuzzy order relations are uniquely characterized as fuzzifications of crisp linear orders.

Theorem 2.7 *[1] Let L be a binary fuzzy relation on X and let E be a T-equivalence on X. Then the following two statements are equivalent:*

1. *L is a strongly linear T-E-order on X.*
2. *There exists a linear order $\preceq$ the relation E is compatible with, such that L can be represented as follows:*

$$L(x,y) = \begin{cases} 1, & \text{if } x \preceq y \\ E(x,y), & \text{otherwise.} \end{cases}$$

This theorem shows that if we have a set X, a linear order $\preceq$ on it and a T-equivalence on X which is compatible with $\preceq$, then we can build a fuzzy linear order as it was shown above.

Further, in some theorems we consider fuzzy order relation on interval $[0,1]$ as a function

$$L(x,y) = \begin{cases} 1, & \text{if } x \leq y \\ g(|x-y|), & \text{otherwise,} \end{cases}$$

where g is a non-increasing function. We do not claim that in general L thus defined is a fuzzy order relation, because it depends on the choice of a function g. But the necessary condition for a relation $E(x,y) = g(|x-y|)$ to be compatible with $\leq$ is that function g is non-increasing. Therefore, if we prove a result for an arbitrary fuzzy relation L defined above, the theorem will also hold for a fuzzy order relation

$$R(x,y) = \begin{cases} 1, & \text{if } x \leq y \\ E(x,y) = g(|x-y|), & \text{otherwise,} \end{cases}$$

where E is a fuzzy equivalence relation compatible with $\leq$.

3 Degree of monotonicity

We start with definition of the degree of monotonicity.

Definition 3.1 *Let $f : [0,1]^n \to [0,1]$ be a function (aggregation function), $P : [0,1]^2 \to [0,1]$ be a fuzzy order relation and $\mapsto_T$ a residuum corresponding to the t-norm T. We define the degree of monotonicity for a function (aggregation function) f w.r.t fuzzy relation P and residuum $\mapsto_T$ in the following way:*

$$M_{P,\mapsto_T}(f) = \inf_{x,y}(\wedge_i P(x_i,y_i) \mapsto_T P(f(x),f(y))).$$

In the sequel, we will often write x to denote an element $x = (x_1,...,x_n)$ and for simplicity of notation we write $x \leq y$ where $x = (x_1,...,x_n)$, $y = (y_1,...,y_n)$ if $x_i \leq y_i$ for all $i \in \{1,...,n\}$.

Example 3.2 *Let us observe the examples which we have presented in the introduction and let us calculate the degree of monotonicity for these aggregations. Let us denote by A the aggregation function, namely, $A(a_k)$ denotes the aggregation result for the alternative a_k. We calculate the degree of monotonicity with respect to the fuzzy order relation:*

$$P(a_i,b_i) = \begin{cases} 1, & \textit{if } a_i \leq b_i \\ \max(1-|a_i-b_i|,0), & \textit{otherwise} \end{cases},$$

based on Lukasiewicz T-norm (see Example 2.3 and Theorem 2.7) and the residuum corresponding to the same t-norm: $a \mapsto_T b = \min(1-a+b,1)$ (Łukasiewicz residuum).

The preliminary results which we get calculating the value $\wedge_i P(a_{ki},a_{ni}) \mapsto_T P(A(a_k),A(a_n))$ (let us denote this value by $\omega(a_k,a_n)$) for every two alternatives a_k and a_n we summarize in Table 3 and Table 4:

For the first example the degree of monotonicity is equal to 0.92*. We note "deficiency" for the alternatives a_2 and a_1 - the result which we expected to get (see Introduction).*

For the example which we have presented in the introduction in the Table 2. the degree of monotonicity is equal to 0.999*.*

Further we study the properties of aggregation functions which have a degree of monotonicity equal to 1.

Table 3: Motivate example 1.

Alt. a_k and a_n	$\omega(a_k, a_n)$
a_1 and a_2	1
a_1 and a_3	1
a_2 and a_1	0.92
a_2 and a_3	1
a_3 and a_1	1
a_3 and a_2	1

Table 4: Motivate example 2.

Alt. a_k and a_n	$\omega(a_k, a_n)$
a_1 and a_2	1
a_1 and a_3	1
a_2 and a_1	0.999
a_2 and a_3	0.999
a_3 and a_1	1
a_3 and a_2	1

Proposition 3.3 *The degree of monotonicity for a function f with respect to a crisp linear order is equal to 1 if and only if f is a monotone function.*

Proof is obvious, see e.g. [3].

We continue with the proposition stating that the degree of monotonicity for the weighted mean with respect to certain fuzzy order relation and residuum, corresponding to the left-continuous t-norm is equal to 1.

Proposition 3.4 *Let f be the weighted mean: $f(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} w_i x_i$ where weight w_i are non negative and $\sum_{i=1}^{n} w_i = 1$, and let g be a non-increasing function. Then the degree of monotonicity for function f with respect to the fuzzy order relation*

$$P(x_i, y_i) = \begin{cases} 1, & \text{if } x_i \leq y_i \\ g(|x_i - y_i|), & \text{otherwise} \end{cases}$$

and the residuum $\mapsto_T$, where T is a left-continuous t-norm, is equal to 1.

Proof. *To find the value* $\inf_{x,y}(\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y)))$ *we consider two cases:*

1. *If $x_i \leq y_i$, $\forall i \in \{1, 2, ..., n\}$ then $P(x_1, y_1) = \cdots = P(x_n, y_n) = 1$ by the definition of fuzzy relation P.*
   Since obviously $\sum_{i=1}^{n} w_i x_i \leq \sum_{i=1}^{n} w_i y_i$ it follows that $P(\sum_{i=1}^{n} w_i x_i, \sum_{i=1}^{n} w_i y_i) = 1$.
   Hence $(\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y))) = (1 \mapsto_T 1) = 1$.

2. *Now we consider the case when there exists such a set $K \subseteq I$ $(I = \{1, 2, ..., n\})$:*
   $\forall k \in K$ $x_k > y_k$.
   Let $g(|x_l - y_l|) = \min_{k \in K} g(|x_k - y_k|) = \wedge_i P(x_i, y_i)$. Thus $\forall k \in K$ $g(|x_l - y_l|) \leq g(|x_k - y_k|)$ and therefore $|x_k - y_k| \leq |x_l - y_l|$ since the function f is non-increasing.
   If $f(x) \leq f(y)$ then $P(f(x), f(y)) = 1$ and $(\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y))) =$
   $= (P(x_l, y_l) \mapsto_T 1) = 1$.

We consider the case when $f(x) > f(y)$*:* $\sum_{i=1}^{n} w_i x_i > \sum_{i=1}^{n} w_i y_i$.
Further $|f(x) - f(y)| = |\sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i y_i| = \sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} w_i (x_i - y_i) \leq$
$\leq \sum_{k \in K} w_k (x_k - y_k) \leq (x_l - y_l) \cdot \sum_{k \in K} w_k \leq |x_l - y_l|$ *and then* $g(|x_l - y_l|) \leq g(|f(x) - f(y)|)$.
Finally, $(\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y))) = (P(x_l, y_l) \mapsto_T P(f(x), f(y))) = 1$.

We have shown that $(\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y))) = 1$ *for all* x *and* y*. Hence* $M_{P,\mapsto_T}(f) = 1$.

Theorem 3.5 *Let* $f : [0,1]^n \to [0,1]$ *be a monotone function,* T *be a continuous Archimedean t-norm with an additive generator* t *and let* d *be an arbitrary pseudo-metric in interval* $[0,1]$ *such that* $d(a,b) \leq t(0)$ *for all* a *and* b *from the unit interval. Then the degree of monotonicity for function* f *with respect to the residuum* $\mapsto_T$ *and the fuzzy order relation*

$$P(x_i, y_i) = \begin{cases} 1, & \text{if } x_i \leq y_i \\ E_d(x_i, y_i), & \text{otherwise} \end{cases}$$

is equal to 1 if and only if $\forall\ x = (x_1, ..., x_n),\ \ y = (y_1, ..., y_n)$

$$f(x) > f(y) \Rightarrow d(f(x), f(y)) \leq \max_{x_i > y_i}(d(x_i, y_i)).$$

Proof. *The sufficiency is proved in [3]. We continue by proving the necessity. Let us assume that there exist such elements* x *and* y *that* $f(x) > f(y)$ *but* $d(f(x), f(y)) > \max_{x_i > y_i}(d(x_i, y_i))$.
Then let us calculate the value $\wedge_i P(x_i, y_i) \to P(f(x), f(y))$*. Let* k *be an integer for which*
$\wedge_i P(x_i, y_i) = P(x_k, y_k) = E_d(x_k, y_k) = t^{(-1)}(\min(d(x_k, y_k), t(0)))$.
Since $f(x) > f(y)$ *we have* $P(f(x), f(y)) = t^{(-1)}(\min(d(f(x), f(y)), t(0)))$.
Then $\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y)) =$
$= t^{(-1)}(\max(t(t^{(-1)}(\min(d(f(x), f(y)), t(0)))) - t(t^{(-1)}(\min(d(x_k, y_k), t(0)))), 0)) =$
$= t^{(-1)}(\max(\min(d(f(x), f(y)), t(0)) - \min(d(x_k, y_k), t(0)), 0)) < t^{(-1)}(0)$*. The last inequality is true since* $\min(d(f(x), f(y)), t(0)) - \min(d(x_k, y_k), t(0)) > 0$ *and* $t^{(-1)}$ *is a strictly decreasing mapping. Thus* $\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y)) \neq 1$.

4 Involving α-levels in the definition of the degree of monotonicity

Not for every monotone function f $M_{P,\mapsto_T}(f)$ is equal to 1. Let illustrate this with the following example:

Example 4.1 *Let us evaluate the degree of monotonicity for weak t-norm*

$$T_W(x_1, x_2) = \begin{cases} min(x_1, x_2), & \text{if } x_1 \vee x_2 = 1 \\ 0, & \text{otherwise} \end{cases},$$

(which is obviously a monotone function), with respect to the fuzzy order relation

$$P(x_i, y_i) = \begin{cases} 1, & \text{if } x_i \leq y_i \\ g(|x_i - y_i|), & \text{otherwise} \end{cases},$$

where g is a continuous non-increasing mapping, and the residuum $\mapsto_T$ *corresponding to a left-continuous t-norm:*

$M_{P,\mapsto_T}(T_W) \leq \inf_{\substack{x=(1,1),\\ y=(y_0,y_0),\\ y_0\in[0,1)}} (\wedge_i P(x_i, y_i) \mapsto_T P(T_W(x), T_W(y))) =$

$= \inf_{y_0\in[0,1)} (P(1, y_0) \mapsto_T P(1, 0)) = \sup_{y_0\in[0,1)} P(1, y_0) \to_T P(1, 0) = 1 \to_T P(1, 0) = P(1, 0) = g(1).$

It is not naturally to define a mapping g in such way that $g(1) = 1$*, in this case g is equal to 1 for all arguments from the unit interval. So, for the mappings g such that* $g(1) \neq 1$ $M_{P,\mapsto_T}(T_W)$ *is not equal to 1.*

Calculating the degree of monotonicity of a monotone function f for every two elements $x, y : \ x < y$ we have to compute the value $\wedge_i P(y_i, x_i) \mapsto_T P(f(y), f(x))$ which is equal to $\wedge_i E(y_i, x_i) \mapsto_T E(f(y), f(x))$ in case when $P(x_i, y_i) = \begin{cases} 1, & \text{if } x_i \leq y_i \\ E(x_i, y_i), & \text{otherwise} \end{cases}$.

Then if f is a monotone function the necessary condition for $M_{P,\mapsto_T}(f) = 1$ is $\inf_{x<y}(\wedge_i E(y_i, x_i) \mapsto_T E(f(y), f(x))) = 1$.

Intuitively this is the degree of the statement: "if x and y are indistinguishable then $f(x)$ and $f(y)$ are indistinguishable". This is something more than just generalization of monotonicity. But we think that it could be a useful condition for the study of aggregation processes.

Actually, if we want to be closer to the classical (crisp) definition of monotonicity, we can calculate the value $\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y))$ only for those elements x, y which are in the relation $x \leq y$ in a certain fuzzy sense. By this we mean that the value $\wedge_i P(x_i, y_i)$ should be close to 1. One can choose a constant α from the interval $[0, 1]$ to define what "close to 1" does mean and calculate the degree of α-monotonicity:

Definition 4.2 *Let* $f : [0, 1]^n \to [0, 1]$ *be a function (aggregation function),* $P : [0, 1]^2 \to [0, 1]$ *be a fuzzy order relation and* $\mapsto_T$ *a residuum corresponding to the t-norm T. We define the degree of* α*-monotonicity for a function (aggregation function) f w.r.t fuzzy relation P and residuum* $\mapsto_T$ *in the following way:*

$$M^{\alpha}_{P,\mapsto_T}(f) = \inf_{\wedge_i P(x_i,y_i)\geq\alpha} (\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y))).$$

It is easy to see that if a fuzzy order P and a t-norm T are fixed and $\alpha_1 \leq \alpha_2$ then $S_{\alpha_1} \subseteq S_{\alpha_2}$, where $S_{\alpha_1} = \{f : M^{\alpha_1}_{P,\to_T}(f) = 1\}$ and $S_{\alpha_2} = \{f : M^{\alpha_2}_{P,\to_T}(f) = 1\}$.

5 Example

Let observe the following example, where f is an arithmetic mean destroyed at one point $x_1 = x_2 = 0.5$: $f(x_1, x_2) = \begin{cases} \frac{x_1+x_2}{2}, x_1, x_2 \neq 0.5 \\ 0.6, \ x_1 = x_2 = 0.5 \end{cases}$.

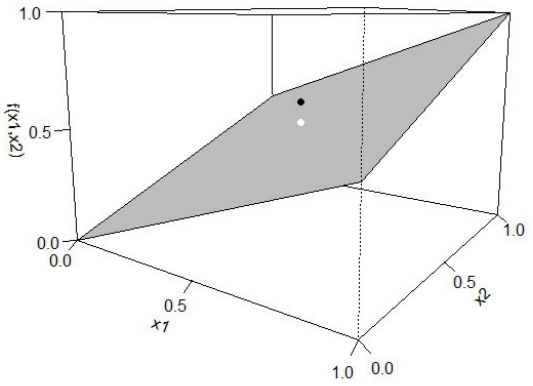


Figure 1: $f(x_1, x_2)$

We involve the fuzzy order relation:

$$P(x_i, y_i) = \begin{cases} 1, & \text{if } x_i \leq y_i \\ \max(1 - |x_i - y_i|, 0), & \text{otherwise} \end{cases},$$

based on Łukasiewicz t-norm (see Example 2.3 and Theorem 2.7). and calculate the degree of monotonicity with respect to three different residuums:
a) residuum corresponding to the Łukasiewicz t-norm: $a \mapsto_{T_L} b = \min(1 - a + b, 1)$;
b) residuum corresponding to the minimum t-norm: $a \mapsto_{T_M} b = \begin{cases} 1, & \text{if } a \leq b \\ b, & \text{otherwise} \end{cases}$;
c) residuum corresponding to the product t-norm: $a \mapsto_{T_P} b = \begin{cases} 1, & \text{if } a \leq b \\ \frac{b}{a}, & \text{otherwise} \end{cases}$.

The function f is monotone everywhere except of point $(0.5, 0.5)$, so we define the defect of monotonicity of function f as

$$def(f) = f(0.5, 0.5) - \lim_{(x_1,x_2)\to(0.5,0.5)} f(x_1, x_2) = 0.1.$$

We calculate the degree of monotonicity for function f with respect to the fuzzy order relation P and each residuum (we will use the abbreviation $\mapsto_T$ if we mean any of the residuums $\mapsto_{T_L}$,$\mapsto_{T_M}$ or $\mapsto_{T_P}$). According to Proposition 3.4, for every two elements x, y $((x_1, x_2) \neq (0.5, 0.5)$ and $(y_1, y_2) \neq (0.5, 0.5)) \wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y)) = 1$. Thus we must find the value $\wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y))$ for all x, y where $(x_1, x_2) = (0.5, 0.5)$ or $(y_1, y_2) = (0.5, 0.5)$. For the brevity we involve notation $\omega_{P,\mapsto_T}(x, y) = \wedge_i P(x_i, y_i) \mapsto_T P(f(x), f(y))$.

Then $M_{P,\mapsto_T}(f) = \inf_{x,y} \omega_{P,\mapsto_T}(x,y)$. We know that: $\inf_{x,y\neq(0.5,0.5)} \omega_{P,\mapsto_T}(x,y) = 1$, so we have to find $\inf_{x=(0.5,0.5)} \omega_{P,\mapsto_T}(x,y)$ and $\inf_{y=(0.5,0.5)} \omega_{P,\mapsto_T}(x,y)$.

- $(x_1, x_2) = (0.5, 0.5)$.
  The results are visualized by the following illustration (for more details see [3]):

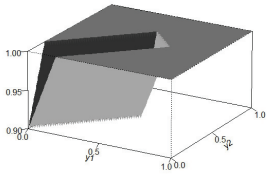


Figure 2: $\omega_{P,\mapsto_{T_L}}(x,y)$, where $x = (0.5, 0.5)$

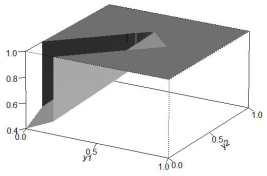


Figure 3: $\omega_{P,\mapsto_{T_M}}(x,y)$, where $x = (0.5, 0.5)$

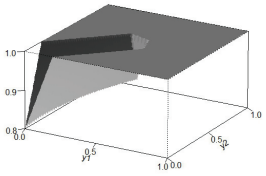


Figure 4: $\omega_{P,\mapsto_{T_P}}(x,y)$, where $x = (0.5, 0.5)$

- $(y_1, y_2) = (0.5, 0.5)$
  According to the Proposition 3.4,
  $P(x_1, 0.5) \wedge P(x_2, 0.5) \mapsto_T P(f(x), 0.5) = 1$. It is easy to verify (using the properties of the residuum $\mapsto_T$ and the properties of the fuzzy order relation P) that if $f(0.5, 0.5) \geq 0.5$ and $f(x_1, x_2) = \frac{x_1+x_2}{2}$ for all (x_1, x_2), where $(x_1, x_2) \neq (0.5, 0.5)$ then
  $P(x_1, 0.5) \wedge P(x_2, 0.5) \mapsto_T P(f(x), f(0.5, 0.5)) = 1$.

Therefore using the above investigations, we obtain the following results:
a) $M_{P,\mapsto_{T_L}}(f) = 1 - def(f) = 0.9$;
b) $M_{P,\mapsto_{T_M}}(f) = 0.4$;
c) $M_{P,\mapsto_{T_P}}(f) = 0.8$.

Thus in these examples the best result is when both the fuzzy order P and the residuum correspond to the same, Łukasiewicz t-norm.

Now let observe the first case from the previous example, when the degree of monotonicity for the destroyed arithmetic mean is calculated with respect to the Łukasiewicz residuum. It is interesting to see what result we obtain if we suppose that two points are indistinguishable if the distance between them is less or equal to 0.1. Thus we would like "to get round" the deficiency of 0.1. Further we calculate the degree of monotonicity for the function $f(x_1, x_2) = \begin{cases} \frac{x_1+x_2}{2}, x_1, x_2 \neq 0.5 \\ 0.6, \ x_1 = x_2 = 0.5 \end{cases}$ with respect to Łukasiewicz residuum and the following fuzzy order relation: $P_{Mod}(x_i, y_i) = \begin{cases} 1, \ if \ x_i \leq y_i + 0.1 \\ 1.1 - |x_i - y_i|, \ otherwise \end{cases}$.

As in the previous example $\omega_{P_{Mod},\mapsto_{T_L}}(x,y)$ is equal to 1 if $y=(0.5,0.5)$. The results when $x=(0.5,0.5)$ are visualized by the following illustration. To compare the results for $\omega_{P_{Mod},\mapsto_{T_L}}$ and $\omega_{P,\mapsto_{T_L}}(x,y)$ we fix in the graph two points: $(0.5,0.5,0.9)$ and $(0.5,0.5,1)$.

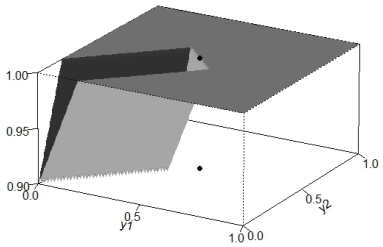


Figure 5: $\omega_{P_{Mod},\mapsto_{T_M}}(x,y)$,where $x=(0.5,0.5)$

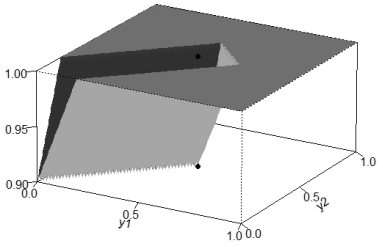


Figure 6: $\omega_{P,\mapsto_{T_M}}(x,y)$,where $x=(0.5,0.5)$

We see that using fuzzy order relation P_{Mod} where we tried "to get round" the deficiency of 0.1 we get the same result $M_{P_{Mod},\mapsto_{T_L}}(f)=1-def(f)=0.9$. To improve the situation we can calculate the degree of α-monotonicity for $\alpha=0.9$. In this case $M^{0.9}_{P,\mapsto_T}(f)=1$. So "to get round" the deficiency we should both modify the fuzzy order relation and use the definition of the degree of α-monotonicity.

6 Conclusion

The degree of monotonicity is calculated for a concrete mapping and depends on this mapping, fuzzy order relation and residuum which are chosen by an expert. The degree of monotonicity takes its values from the interval [0,1]. In case of a crisp order relation the property of having the degree of monotonicity equal to 1 is equivalent to the property of being monotone in the crisp sense. We consider the behavior of the degree of monotonicity calculated with respect to the given fuzzy order relation and residuum, illustrating it with examples. Besides, we study how a deficiency of monotonicity influences the degree of monotonicity. In the work we also study necessary properties when the degree of monotonicity is equal to 1. Note that a more complete theory is obtained by involving α levels in the definition of the degree of monotonicity.

Acknowledgement

The paper was partially supported by ESF project
Nr. 2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004.

References

[1] BODENHOFER, U.: *A similarity-based generalization of fuzzy orderings preserving the classical axioms*. In Internat. J. Uncertain. Fuzziness Knowledge-Based Systems, Vol. 8, No. 5, pp. 593-610, 2000.

[2] GRABISCH, M., MARICHAL, J.-L., MESIAR, R., PAP, E.: *Aggregation Functions (Encyclopedia of Mathematics and its Applications)*. Cambridge University Press, UK, 2009.

[3] GRIGORENKO, O.: *Degree of monotonicity in aggregation process*. In Proceedings of 2010 IEEE International Conference on Fuzzy Systems, pp. 1080-1087, 2010.

[4] KLEMENT, E.P., MESIAR, R., PAP, E.: *Triangular Norms*. Kluwer Academic Publishers, The Netherlands, 2002.

Current address

Olga Grigorenko, Mgr.
Department of Physics and Mathematics,
University of Latvia, Zellu street 8, Riga, LV-1002,
phone: +37129234989,
email: ol.grigorenko@gmail.com.

Journal of Applied Mathematics

volume IV (2011), number III

NON-CLAUSAL RESOLUTION AND FUZZY LOGIC

HABIBALLA Hashim (CZ)

Abstract. The paper presents experimental comparison of several resolution strategies for reasoning in Fuzzy Predicate Logic with evaluated syntax. Resolution-based reasoning is established on previous works concerning non-clausal resolution principle both theoretical and application-oriented (FPLGERDS inference engine).

Key words and phrases. Fuzzy inference, Fuzzy Description Logic, Resolution Strategies.

1 Introduction

Fuzzy Predicate Logic with Evaluated Syntax (FPL) [11] is a well-studied and wide-used logic capable of expressing vagueness. It has a lot of applications based on robust theoretical background. The knowledge representation itself doesn't lead to full applicable deductive system. It also requires an efficient formal proof theory. Since DL is semantically a subclass of FOL it may use various proved techniques like tableaux algorithm. However the most widely applied resolution principle [3] brings syntactically several obstacles mainly arising from normal form transformation. FPL is associating with even harder problems when trying to use the resolution principle. The solutions to these obstacles based on the non-clausal resolution [2] were already proposed in [7] and [6].

It leads to the refutational resolution theorem prover for FDL ($RRTP_{FDL}$). FDL has been presented in a simple form in [**?**] and in more general form in [10] with strong computability and time complexity results. The article refers to some definitions from cited refutational resolution theorem provers for DL ($RRTP_{DL}$) and FPL ($RRTP_{FPL}$).

Another issue addressed in the paper concerns to the efficiency of presented inference strategies developed originally for the proving system. We show their perspectives in combination with standard proof-search strategies. The main problem for the fuzzy logic theorem proving

lies in the large amount of possible proofs with different degrees and there is presented an algorithm (Detection of Consequent Formulas - DCF) solving this problem. The algorithm is based on detection of such redundant formulas (proofs) with different degrees.

2 General resolution and unification extensions for existentiality

For the purposes of ($RRTP_{FPL}$) we will use generalized principle of resolution, which is defined in the research report [1]. There is a propositional form of the rule defined at first and further it is lifted into first-order logic. We will introduce the propositional form of the general resolution.

General resolution - propositional version

$$\frac{F[G] \quad F'[G]}{F[G/\bot] \vee F'[G/\top]} \tag{1}$$

where the propositional logic formulas F and F' are the premises of inference and G is an occurrence of a subformula of both F and F'. The expression $F[G/\bot] \vee F'[G/\top]$ is the resolvent of the premises on G. Every occurrence of G is replaced by false in the first formula and by true in the second one. It is also called F the positive, F' the negative premise, and G the resolved subformula.

The proof of the soundness of the rule is similar to clausal resolution rule proof. Suppose the Interpretation I in which both premises are valid. In I, G is either true or false. If G ($\neg G$) is true in I, so is $F'[G/\top]$ ($F[G/\bot]$).

Revised version of the paper which forms the core of the handbook [2] is closely related with notion of selection functions and ordering constraints. By a selection functions we mean a mapping S that assigns to each clause C a (possibly empty) multiset $S(C)$ of negative literals in C. In other words, the function S selects (a possibly empty) negative subclause of C. We say that an atom A, or a literal $\neg A$, is selected by S if $\neg A$ occurs in $S(C)$. (There are no selected atoms or literals if $S(C)$ is empty.) As an usual ordering can be used lexicographic path ordering over a total precedence. But in this case the ordering is admissible if predicate symbols have higher precedence than logical symbols and the constants $\top$ and $\bot$ are smaller than the other logical symbols. It means the ordering is following $A \succ \equiv \succ \supset \succ \neg \succ \vee \succ \wedge \succ \top \succ \bot$. The handbook also addresses another key issues for automated theorem proving - the efficiency of the proof search. This efficiency is closely related with the notion of *redundancy*.

If we want to generalize the notion of resolution and lift it into first-order case we have to define first the notion of selection function for general clauses. General clauses are multisets of arbitrary quantifier-free formulas, denoting the disjunction of their elements. Note, that we can also work with a special case of such a general clauses with one element, which yields to a standard quantifier-free formula of first-order logic. A (general selection) function is a mapping S that assigns to each general clause C a (possibly empty) set C of non-empty sequences of (distinct) atoms in C such that either $S(C)$ is empty or else, for all interpretations I in which C is false, there exists a sequence $A_1, ..., A_k$ in $S(C)$, all atoms of which are true in I. A sequence $A_1, ..., A_k$ in $S(C)$ is said to be *selected* (by S).

We have to define the notion of polarity for these reasons according to the handbook [2]. It is based on the following assumption that a subformula F' in $E[F']$ is *positive* (resp. *negative*), if $E[F'/\top]$ (resp. $E[F'/\bot]$) is a tautology. Thus, if F' is *positive* (resp. *negative*) in E, F'

(resp. $\neg F'$) logically implies E. Even it should seem that determining of the polarity of any subformula is NP-complete (hard) problem, we can use syntactic criteria for this computation. In this case the complexity of the algorithm is linear (note that we base our theory on similar syntactic criteria below - structural notions definition).

When trying to refine the general resolution rule for fuzzy predicate logic , it is important to devise a sound and complete unification algorithm. Standard unification algorithms require variables to be treated only as universally quantified ones. We will present a more general unification algorithm, which can deal with existentially quantified variables without the need for those variables be eliminated by skolemization. It should be stated that the following unification process doesn't allow an occurrence of the equivalence connective. It is needed to remove equivalence by rewrite rule: $A \leftrightarrow B \Leftrightarrow [A \rightarrow B] \wedge [B \rightarrow A]$.

We assume that the language and semantics of FOL is standard. We use terms - individuals $(a, b, c, ...)$, functions (with n arguments) $(f, g, h, ...)$, variables $(X, Y, Z, ...)$, predicates(with n arguments) $(p, q, r, ...)$, logical connectives $(\wedge, \vee, \rightarrow, \neg)$, quantifiers $(\exists, \forall)$ and logical constants $(\bot, \top)$. We also work with standard notions of logical and special axioms (sets LAx, SAx), logical consequence, consistency etc. as they are used in mathematical logic.

Definition 1
Structural notions of a FOL formula
Let F be a formula of FOL then the structural mappings Sub (subformula), Sup (superformula), Pol (polarity) and Lev (level) are defined as follows:

$F = G \wedge H$ or $F = G \vee H$	$Sub(F) = \{G, H\}$, $Sup(G) = Sup(H) = F$ $Pol(G) = Pol(F) = Pol(H)$
$F = G \rightarrow H$	$Sub(F) = \{G, H\}$, $Sup(G) = Sup(H) = F$ $Pol(G) = -Pol(F)$, $Pol(H) = Pol(F)$
$F = \neg G$	$Sub(F) = \{G\}$, $Sup(G) = F$ $Pol(G) = -Pol(F)$
$F = \exists \alpha G$ or $F = \forall \alpha G$ (α is a variable)	$Sub(F) = \{G\}$, $Sup(G) = F$ $Pol(G) = Pol(F)$

$Sup(F) = \emptyset \Rightarrow Lev(F) = 0, Pol(F) = 1,$
$Sup(F) \neq \emptyset \Rightarrow Lev(F) = Lev(Sup(F)) + 1$
For mappings Sub and Sup reflexive and transitive closures Sub^* and Sup^* are defined recursively as follows:
1. $Sub^*(F) \supseteq \{F\}$, $Sup^*(F) \supseteq \{F\}$
2. $Sub^*(F) \supseteq \{H | G \in Sub^*(F) \wedge H \in Sub(G)\}$, $Sup^*(F) \supseteq \{H | G \in Sup^*(F) \wedge H \in Sup(G)\}$

Example: $A \rightarrow B$ - $Pol(A) = -1$, $Pol(B) = 1$, $Lev(A) = 1$

These structural mappings provide framework for assignment of quantifiers to variable occurrences. It is needed for the correct simulation of skolemization (the information about a variable quantification in the prenex form). Subformula and superformula mappings and its closures encapsulate essential hierarchical information of a formula structure. Level gives the ordering with respect to the scope of variables (which is also essential for skolemization simulation - unification is restricted for existential variables). Polarity enables to decide the global meaning of a variable (e.g. globally an existential variable is universal if its quantification subformula has negative polarity). Sound unification requires further definitions on variable

quantification. We will introduce notions of the corresponding quantifier for a variable occurrence, substitution mapping and significance mapping (we have to distinguish between original variables occurring in special axioms and newly introduced ones in the proof sequence).

Definition 2
Variable assignment, substitution and significance
Let F be a formula of FOL, $G = p(t_1, ..., t_n) \in Sub^*(F)$ atom in F and α a variable occurring in t_i. Variable mappings Qnt(quantifier assignment), Sbt (variable substitution) and Sig(significance) are defined as follows:

$Qnt(\alpha) = Q\alpha H, where Q = \exists \vee Q = \forall, H, I \in Sub^*(F), Q\alpha H \in Sup^*(G)$,
$\forall Q\alpha I \in Sup^*(G) \Rightarrow Lev(Q\alpha I) < Lev(Q\alpha H)$.
$F[\alpha/t']$ is a substitution of term t' into α in $F \Rightarrow Sbt(\alpha) = t'$.
A variable α occurring in $F \in LAx \cup SAx$ is significant w.r.t. existential substitution, $Sig(\alpha) = 1$ iff variable is significant, $Sig(\alpha) = 0$ otherwise.

Example: $\forall x(\forall x A(x) \rightarrow B(x))$ - $Qnt(x) = \forall x A(x)$, for x in $A(x)$ and $Qnt(x) = \forall x(\forall x A(x) \rightarrow B(x))$, for x in $B(x)$.

Note that with Qnt mapping (assignment of first name matching quantifier variable in a formula hierarchy from bottom) we are able to distinguish between variables of the same name and there is no need to rename any variable. Sbt mapping holds substituted terms in a quantifier and there is no need to rewrite all occurrences of a variable when working with this mapping within unification. It is also clear that if $Qnt(\alpha) = \emptyset$ then α is a free variable. These variables could be simply avoided by introducing new universal quantifiers to F. Significance mapping is important for differentiating between original formula universal variables and newly introduced ones during proof search (an existential variable can't be bounded with it).

Before we can introduce the standard unification algorithm, we should formulate the notion of global universal and global existential variable (it simulates conversion into prenex normal form).

Definition 3
Global quantification
Let F be a formula without free variables and α be a variable occurrence in a term of F.

1. α is a global universal variable ($\alpha \in Var_\forall(F)$) iff ($Qnt(\alpha) = \forall \alpha H$ $\wedge Pol(Qnt(\alpha)) = 1$) or ($Qnt(\alpha) = \exists \alpha H \wedge Pol(Qnt(\alpha)) = -1$)

2. α is a global existential variable ($\alpha \in Var_\exists(F)$) iff ($Qnt(\alpha) = \exists \alpha H$ $\wedge Pol(Qnt(\alpha)) = 1$) or ($Qnt(\alpha) = \forall \alpha H \wedge Pol(Qnt(\alpha)) = -1$)

$Var_\forall(F)$ and $Var_\exists(F)$ are sets of global universal and existential variables.

Example: $F = \forall y(\forall x A(x) \rightarrow B(y))$ - x is a global existential variable, y is a global universal variable.

It is clear w.r.t. skolemization technique that an existential variable can be substituted into an universal one only if all global universal variables over the scope of the existential one have been already substituted by a term. Skolem functors function in the same way. Now

we can define the most general unification algorithm based on recursive conditions (extended unification in contrast to standard MGU).

Definition 4
Most general unifier algorithm
Let $G = p(t_1, ..., t_n)$ and $G' = r(u_1, ..., u_n)$ be atoms. Most general unifier (substitution mapping) MGU(G, G') $= \sigma$ is obtained by following atom and term unification steps or the algorithm returns fail-state for unification. For the purposes of the algorithm we define the Variable Unification Restriction (VUR).

Variable Unification Restriction

Let F_1 be a formula and α be a variable occurring in F_1, F_2 be a formula, t be a term occurring in F_2 and β be a variable occurring in F_2. Variable Unification Restriction (VUR) for (α,t) holds if one of the conditions 1. and 2. holds:

1. α is a global universal variable and $t \neq \beta$, where β is a global existential variable and α not occurring in t (non-existential substitution)

2. α is a global universal variable and $t = \beta$, where β is a global existential variable and $\forall F \in Sup^*(Qnt(\beta))$, $F = Q\gamma G$, $Q \in \{\forall, \exists\}$, γ is a global universal variable, $Sig(\gamma) = 1$ $\Rightarrow (Sbt(\gamma) = r') \in \sigma$, r' is a term (existential substitution).

Atom unification

1. if $n = 0$ and $p = r$ then $\sigma = \emptyset$ and the unifier exists (success-state).

2. if $n > 0$ and $p = r$ then perform term unification for pairs $(t_1, u_1), \ldots, (t_n, u_n)$; If for every pair unifier exists then $MGU(G, G') = \sigma$ obtained during term unification (success state).

3. In any other case unifier doesn't exist (fail-state).

Term unification (t', u')

1. if $u' = \alpha$, $t' = \beta$ are variables and $Qnt(\alpha) = Qnt(\beta)$ then unifier exists for (t', u') (success-state) (occurrence of the same variable).

2. if $t' = \alpha$ is a variable and $(Sbt(\alpha) = v') \in \sigma$ then perform term unification for (v', u'); The unifier for (t', u') exists iff it exists for (v', u') (success-state for an already substituted variable).

3. if $u' = \alpha$ is a variable and $(Sbt(\alpha) = v') \in \sigma$ then perform term unification for (t', v'); The unifier for (t', u') exists iff it exists for (t', v') (success-state for an already substituted variable).

4. if $t' = a$, $u' = b$ are individual constants and a = b then for (t',u') unifier exists (success-state).

5. if $t' = f(t'_1, ..., t'_m)$, $u' = g(u'_1, ..., u'_n)$ are function symbols with arguments and $f = g$ then unifier for (t', u') exists iff unifier exists for every pair (t'_1, u'_1), ..., (t'_n, u'_n) (success-state).

6. if $t' = \alpha$ is a variable and VUR for (t', u') holds then unifier exists for (t', u') holds and $\sigma = \sigma \cup (Sbt(\alpha) = u')$ (success-state).

7. if $u' = \alpha$ is a variable and VUR for (u', t') holds then unifier exists for (t', u') holds and $\sigma = \sigma \cup (Sbt(\alpha) = t')$ (success-state).

8. In any other case unifier doesn't exist (fail-state).

$MGU(A) = \sigma$ for a set of atoms $A = \{G_1, \ldots, G_k\}$ is computed by the atom unification for $(G_1, G_i), \sigma_i = MGU(G_1, G_i), \forall i, \sigma_0 = \emptyset$, where before every atom unification (G_1, G_i), σ is set to σ_{i-1}.

With above defined notions it is simple to state the general resolution rule for FOL (without the equivalence connective). It conforms to the definition from [1].

Definition 5
General ordered resolution with selection for first-order logic (GR_{FOL})

$$\frac{F[G_1, , ..., G_k] \quad F'[G'_1, ..., G'_n]}{F\sigma[G/\bot] \vee F'\sigma[G/\top]} \tag{2}$$

where $\sigma = MGU(A)$ is the most general unifier (MGU) of the set of the atoms $A = \{G_1, \ldots, G_k, G'_1, \ldots, G'_n\}$, $G = G_1\sigma$. For every variable α in F or F', $(Sbt(\gamma) = \alpha) \cap \sigma = \emptyset$ $\Rightarrow Sig(\alpha) = 1$ in F or F' iff $Sig(\alpha) = 1$ in $F\sigma[G/\bot] \vee F'\sigma[G/\top]$. F is called positive and F' is called negative premise, G represents an occurrence of an atom. The expression $F\sigma[G/\bot] \vee F'\sigma[G/\top]$ is the resolvent of the premises on G.
and
(i) either G is selected by S in F', or else $S(F')$ is empty, G is maximal in F', (ii) atom G is maximal in F, and (iii) F does not contain a selected atom.

Note that with Qnt mapping we are able to distinguish variables not only by its name (which may not be unique), but also with this mapping (it is unique). Sig property enables to separate variables, which were not originally in the scope of an existential variable. When utilizing the rule it should be set the Sig mapping for every variable in axioms and negated goal to 1. We present a very simple example of existential variable unification before we introduce the refutational theorem prover for FOL.

3 Fuzzy Predicate Logic and refutational proof

The fuzzy predicate logic with evaluated syntax is a flexible and fully complete formalism, which will be used for the below presented extension [11]. In order to use an efficient form of the resolution principle we have to extend the standard notion of a proof (provability value and degree) with the notion of refutational proof (refutation degree). Propositonal version

of the fuzzy resolution principle has been already presented in [5]. We suppose that set of truth values is Łukasiewicz algebra. Therefore we assume standard notions of conjunction, disjunction etc. to be bound with Łukasiewicz operators. It is important that we assume that for every subformula $Sub, Sup, Pol, Lev, Qnt, Sbt, Sig$ and other derived properties defined in section 2 hold (where the classical FOL connective is presented the Łukasiewicz one has the same mapping value).

Definition 6
Evaluated proof, refutational proof and refutation degree
An evaluated formal proof of a formula A from the fuzzy set $X \underset{\sim}{\subseteq} F_J$ is a finite sequence of evaluated formulas $w := \ a_0/A_0, \ a_1/A_1, ..., \ a_n/A_n$ such that $A_n := A$ and for each $i \leq n$, either there exists an m-ary inference rule r such that
$a_i/A_i := \ r^{evl}(a_{i_1}, ..., a_{i_m})/r^{syn}(A_{i_1}, ..., A_{i_m})$,
$i_1, ..., i_m < n$ or $\ a_i/A_i := \ X(A_i)/A_i$.
We will denote the value of the evaluated proof by $Val(w) = a_n$.
An evaluated refutational formal proof of a formula A from X is w, where additionally $\ a_0/A_0 := 1/\neg A$ and $A_n := \bot$. $Val(w) = a_n$ is called refutation degree of A.

Definition 7
General ordered resolution with selection for fuzzy predicate logic (GR_{FPL})

$$r_{GR} : \frac{a/F[G_1, , ..., G_k], \ b/F'[G'_1, ..., G'_n]}{a \otimes b/F\sigma[G/\bot]\nabla F'\sigma[G/\top]} \tag{3}$$

where $\sigma = MGU(A)$ is the most general unifier (MGU) of the set of the atoms $A = \{G_1, \ldots, G_k, G'_1, \ldots, G'_n\}$, $G = G_1\sigma$. For every variable α in F or F', $(Sbt(\gamma) = \alpha) \cap \sigma = \emptyset$ $\Rightarrow Sig(\alpha) = 1$ in

F or F' iff $Sig(\alpha) = 1$ in $F\sigma[G/\bot] \vee F'\sigma[G/\top]$. F is called positive and F' is called negative premise, G represents an occurrence of an atom. The expression $F\sigma[G/\bot]\nabla F'\sigma[G/\top]$ is the resolvent of the premises on G.
and
(i) either G is selected by S in F', or else $S(F')$ is empty, G is maximal in F', (ii) atom G is maximal in F, and (iii) F does not contain a selected atom.

Lemma 1
Soundness of r_{GR}
The inference rule r_{GR} for FPL based on $\mathcal{L}_{\text{Ł}}$ is sound i.e. for every truth valuation $\mathcal{D}$,

$$\mathcal{D}(r^{syn}(A_1, ..., A_n)) \geq r^{evl}(\mathcal{D}(A_1), ..., \mathcal{D}(A_n)) \tag{4}$$

holds true.

Definition 8
Refutational resolution theorem prover
Refutational non-clausal resolution theorem prover for FPL $(RRTP_{FPL})$ is the inference system with the inference rule GR_{FPL} and sound simplification rules for $\bot$, $\top$ (standard equivalencies

for logical constants). A refutational proof by definition 6 represents a proof of a formula G (goal) from the set of special axioms N. It is assumed that $Sig(\alpha) = 1$ for $\forall\alpha$ in $F \in N \cup \neg G$ formula, every formula in a proof has no free variable and has no quantifier for a variable not occurring in the formula.

Definition 9
Simplification rules for $\nabla, \Rightarrow$

$$r_{s\nabla} : \frac{a/\perp \nabla A}{a/A} \quad and \quad r_{s\Rightarrow} : \frac{a/\top \Rightarrow A}{a/A}$$

Lemma 2
Provability and refutation degree for GR_{FPL}
$T \vdash_a A$
iff $a = \bigvee\{Val(w)|$ w is a refutational proof of A from LAx $\cup$ SAx$\}$

Theorem 1
Completeness for fuzzy logic with r_{GR}, $r_{s\nabla}$, $r_{s\Rightarrow}$ instead of r_{MP}
Formal fuzzy theory, where r_{MP} is replaced with r_{GR}, $r_{s\nabla}$, $r_{s\Rightarrow}$, is complete i.e. for every A from the set of formulas $T \vdash_a A$ iff $T \models_a A$.

4 Implementation and efficiency

The author also currently implements the non-clausal theorem prover into fuzzy logic and logic as an extension of previous prover for FOL (GEneralized Resolution Deductive System - GERDS) [4]. Experiments concerning prospective inference strategies can be performed with this extension. The prover called Fuzzy Predicate Logic GEneralized Resolution Deductive System - FPLGERDS provides standard interface for input (knowledge base and goals) and output (proof sequence and results of fuzzy inference, statistics).

There are already several efficient strategies proposed by author (mainly Detection of Consequent Formulas (DCF) adopted for the usage also in FPL). With these strategies the proving engine can be implemented in "real-life" applications since the complexity of theorem proving in FPL is dimensionally harder than in FOL (the need to search for all possible proofs - we try to find the best refutation degree). The DCF idea is to forbid the addition of a resolvent which is a logical consequence of any previously added resolvent. For refutational theorem proving it is a sound and complete strategy and it is emiprically very effective. Completeness of such a strategy is also straight-forward in FOL:

$$(R_{old} \vdash R_{new}) \wedge (U, R_{new} \vdash \perp) \Rightarrow (U, R_{old} \vdash \perp)$$

Example: $R_{new} = p(a)$, $R_{old} = \forall x(p(x))$, $R_{old} \vdash R_{new}$.

DCF could be implemented by the same procedures like General Resolution (we may utilize self-resolution). Self-resolution has the same positive and negative premise and needs to resolve all possible combinations of an atom. It uses the following scheme:

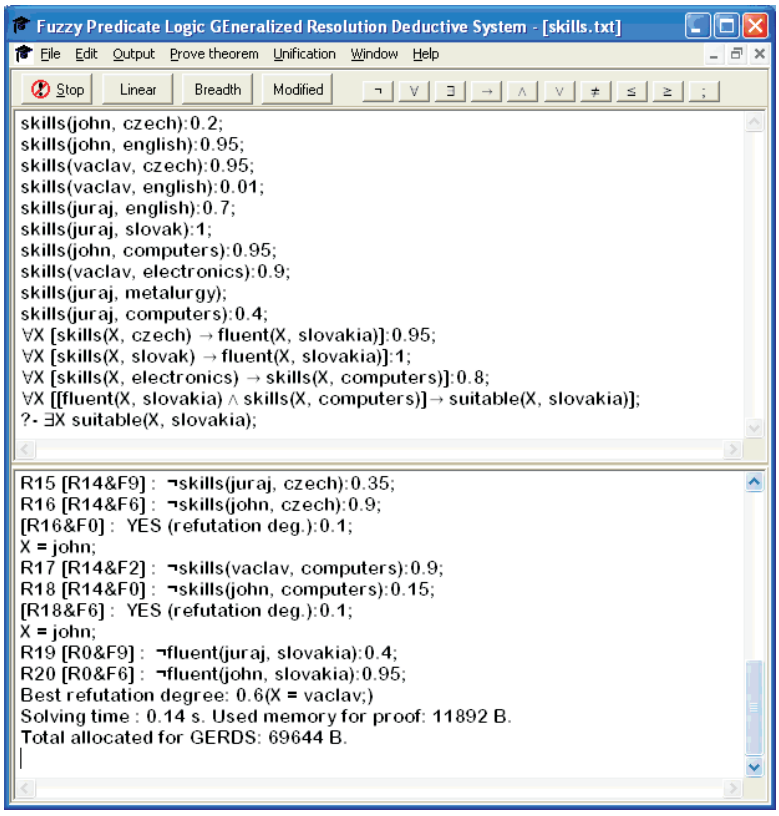


Figure 1: Fuzzy Predicate Logic GEneralized Resolution Deductive System

$$R_{old} \vdash R_{new} \Leftrightarrow \neg(R_{old} \rightarrow R_{new}) \vdash \bot$$

Even the usage of this teachnique is a semidecidable problem, we can use time or step limitation of the algorithm and it will not affect the completeness of the $RRTP_{FOL}$.
Example: $R_{new} = p(a)$, $R_{old} = \forall x(p(x))$, $\neg(\forall x(p(x)) \rightarrow p(a))$
MGU: $Sbt(x) = a$, $Res = \neg(\bot \rightarrow \bot) \vee \neg(\top \rightarrow \top) \Rightarrow \bot$
We have proved that R_{new} is a logical consequence of R_{old}.

In FPL we have to enrich the DCF procedure by the limitation on the provability degree. if $U \vdash_a R_{old} \wedge U \vdash_b R_{new} \wedge b \leq a$ then we can apply DCF. DCF Trivial check performs a symbolic comparison of R_{old} and R_{new} we use the same provability degree condition. In other cases we have to add R_{new} into the set of resolvents and we can apply "DCF Kill" procedure. DCF Kill searches for every R_{old} being a logical consequence of R_{new} and if $U \vdash_a R_{old} \wedge U \vdash_b R_{new} \wedge b \geq a$ then Kill R_{old} (resolvent is removed).

We will now show some efficiency results concerning many-valued logic both for Fuzzy Predicate Logic. We have used the above mentioned application FPLGERDS and originally developed DCF strategy for FPL. It is clear that inference in $RRTP_{FPL}$ and $RRTP_{FDL}$ on general knowledge bases is a problem solved in exponential time. Nevertheless as we would like to demonstrate the need to search for every possible proof (in contrast to the two-valued logic) will not necessarily in particular cases lead to the inefficient theory. We have devised

knowledge bases (KB) on the following typical problems related to the use of fuzzy logic.

We have performed experimental measurements concerning efficiency of the presented non-clausal resolution principle and also DCF technique. This measurements were done using the FPLGERDS application [8]. Special testing knowledge bases were prepared and several types of inference were tested on a PC with standard Intel Pentium 4 processor as described below.

Fuzzy Logic redundancy-based inefficient knowledge bases

As it was shown above in the theorem proving example the problem of proof search is quite different in FPL and FDL in comparison with the two-valued logic. We have to search for the best refutation degree using refutational theorem proving in order to make sensible conclusions from the inference process. It means we cannot accept the **"first successful"** proof, but we have to check **"all possible proofs"** or we have to be sure that every omitted proof is **"worse"** that some another one. The presented DCF and DCF Kill technique belong to the third sort of proof search strategies, i.e. they omit proofs that are really "worse" than some another (see the explication above). Proofs and formulas causing this could be called redundant proofs and redundant formulas. Fuzzy logic makes this redundancy dimensionally harder since we could produce not only equivalent formulas but also equivalent formulas of different evaluation degree.

We have compared efficiency of the standard **breadth-first search**, **linear search** and **modified linear search** (starting from every formula in knowledge base) and also combinations with DCF and DCF-kill technique [8]. We have prepared knowledge bases of the size 120, 240, 360, 480 and 600 formulas. It has been compared the time and space efficiency on the criterion of 2 redundancy levels. This level represents the number of redundant formulas to which the formula is equivalent (including the original formula). For example the level 5 means the knowledge base contain 5 equivalent redundant formulas for every formula (including the formula itself). The basic possible state space search techniques and DCF heuristics and their combinations are presented in the following tables.

Table 1: Proof search algorithms

Search method		**Description**
Breadth	B	Level order generation, start - special axioms + goal
Linear	L	Resolvent ⇒ premise, start - goal
Modified-Linear	M	Resolvent ⇒ premise, start - goal + special axioms

We use standard state space search algorithms in the FPLGERDS application - Breadth-first and Linear search. Breadth-first method searches for every possible resolvent from the formulas of the level 0 (goal and special axioms). These resolvents form formulas of the level 1 and we try to combine them with all formulas of the same and lower level and continue by the same procedure until no other non-redundant resolvent could be found. Linear search performs depth-first search procedure, where every produced resolvent is used as one of the premises in succeeding step of inference. The first produced resolvents arises from the goal formula. Modified linear search method posses the same procedure as linear one, but it starts from goal and also from all the special axioms.

Table 2: DCF heuristics

DCF Method		**Description**
Trivial	T	Exact symbolic comparison
DCF	DC	Potential resolvent is consequent (no addition)
DCF Kill	DK	DCF + remove all consequent resolvents

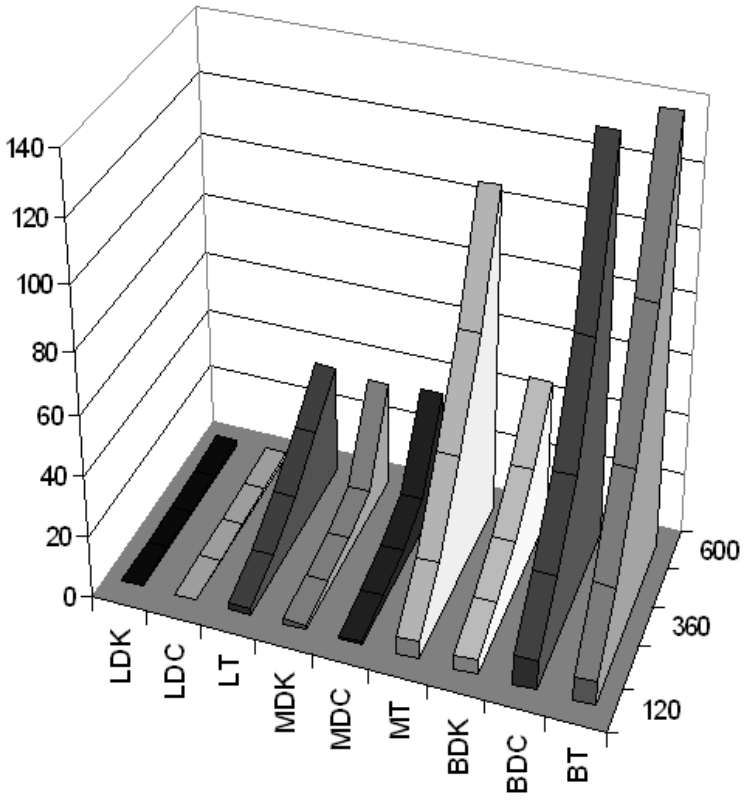
DCF methods for reduction of resolvent space are basically three. The simplest is trivial DCF method, which detects redundant resolvent only by its exact symbolic comparison, i.e. formulas are equivalent only if the are syntactically the same. Even it is a very rough method, it is computationally very simple and forms necessary essential restriction for possibly infinite inference process. The next method of DCF technique enables do detect the equivalency of a formula (potential new resolvent) by the means described above. DCF Kill technique additionally tries to remove every redundant resolvent from the set of resolvents.

Table 3: Inference strategies

Search	**DCF**	Code	**Description**
Breadth	Trivial	BT	Complete
Breadth	DCF	BDC	Complete
Breadth	DCF Kill	BDK	Complete
Mod. Linear	Trivial	MT	Incomplete (+)
Mod. Linear	DCF	MDC	Incomplete (+)
Mod. Linear	DCF Kill	MDK	Incomplete (+)
Linear	Trivial	LT	Incomplete
Linear	DCF	LDC	Incomplete
Linear	DCF Kill	LDK	Incomplete

We have built-up 9 combinations of inference strategies from the mentioned proof search and DCF heuristics. They have different computational strength, i.e. their completeness is different for various classes of formulas. Fully complete (as described above) for general formulas of FPL and FDL are only breadth-first search combinations. Linear search strategies are not complete even for two-valued logic and horn clauses. Modified linear search has generally bad completeness results when an infinite loop is present in proofs, but for guarded knowledge bases it can assure completeness preserving better space efficiency than breadth-first search.

We have tested presented inference strategies on the sample knowledge bases with redundancy level 5 with 20, 40, 60, 80 and 100 groups of mutually redundant formulas (the total number of formulas in knowledge base is 120, 240, 360, 480 and 600). At first we have tested their time efficiency for inference process. As it could be observed from figure 2, the best results have **LDK and LDC** strategies. For simple guarded knowledge bases (not leading to an infinite loop in proof search and where the goal itself assures the best refutation degree) these two methods are **very efficient**. DCF strategies significantly reduces the proof search even in comparison with LT strategy (standard), therefore the usage of any non-trivial DCF heuristics is significant. Next important result concludes from the comparison of BDK and MDK, MDC strategies. We can conclude that MDK and MDC strategies are relatively comparable to BDK and moreover BDK preserves completeness for general knowledge bases.



	120	240	360	480	600
BT	7,44	25,09	50,55	90,91	138,67
BDC	9,52	23,73	43,76	76,63	129,94
BDK	4,68	10,1	17,64	29,54	47,31
MT	5,86	18,85	41,46	70,01	106,68
MDC	0,93	3,09	5,14	10,54	35,14
MDK	1,45	3,2	5,03	8,84	34,32
LT	2,17	7,24	14,76	24,98	35,77
LDC	0,26	0,52	0,57	0,8	1,96
LDK	0,28	0,5	0,57	0,78	1,99

Figure 2: Time complexity for redundancy level 5 (seconds)

Space complexity is even more significantly affected by the DCF heuristics. There is an interesting comparison of trivial and non-trivial DCF heuristics. Even BDK strategy brings significant reduction of resolvents amount, while LDK, LDC, MDK, MDC strategies have minimal necessary amount of kept resolvents during inference process. Performed experiments shows the significance of originally developed DCF strategies in combination with standard breadth-first search (important for general knowledge bases - **BDK**). We also outlined high efficiency for linear search based strategies (mainly **LDK**). Even this strategy is not fully complete and could be used only for guarded fragment of FDL, this problem is already known in classical (two-valued) logic programming and automated theorem proving. We also use these highly efficient linear search strategies, even they are not complete.

5 Conclusion, further research and applications

The Detection of Consequent Formulas algorithms family brings significant improvements in time and space efficiency for the best proof search. We have shown results indicating specific behavior of some combinations of the DCF and standard proof search (breadth-first and linear search). DCF strategies (BDC, BDK) have interesting results even for fully general fuzzy predicate logic with evaluated syntax, where the strategy makes the inference process practically manageable (in contrast to unrestricted "blind" proof-search). However it seems to be more promising for practical applications to utilize incomplete strategies with high time efficiency like LDK (even for large knowledge bases it has very short solving times). It conforms to another successful practical applications in two-valued logic like logic programming or deductive databases where we also use efficient incomplete strategies for fragments of fully general logics.

We have briefly shown some efficiency results for the presented automated theorem prover and inference strategies. They show the significant reduction of time and space complexity for the DCF technique. There is also prepared the experimental application FPLGERDS, which is obtainable from URL:

http://www1.osu.cz/home/habibal/files/gerds.zip.

The package contains current version of the application, source codes, examples and documentation.

Acknowledgement

The authors have been partially supported by the project MSM6198898701 and 1M0572 of the MSMT CR.

References

[1] BACHMAIR, L., GANZINGER, H. A theory of resolution. Technical report: Max-Planck-Institut, 1997.

[2] BACHMAIR, L., GANZINGER, H. Resolution theorem proving. In Handbook of Automated Reasoning, MIT Press, 2001.

[3] DUKIĆ, N., AVDAGIĆ, Z. Fuzzy Functional Dependency and the Resolution Principle. In Informatica, Vilnius: Lith. Acad. Sci. (IOSPRESS), 2005, Vol.16, No. 1, pp. 45 - 60, 2005.

[4] HABIBALLA, H. Non-clausal resolution - theory and practice. Research report: University of Ostrava, 2000, http://www.volny.cz/habiballa/files/gerds.pdf

[5] HABIBALLA, H., NOVÁK, V. Fuzzy General Resolution. In Proc. of Intl. Conf. Aplimat 2002. Bratislava, Slovak Technical University, 2002. pp. 199-206, also available as research rep. at http://ac030.osu.cz/irafm/ps/rep47.ps

[6] HABIBALLA, H. Non-clausal Resolution in Fuzzy Predicate Logic with Evaluated Syntax (background and implementation). In Proc. of Intl. Conf. The Logic of Soft Computing IV, Ostrava, pp. 51 - 54, 2005, also available as research rep. at http://ac030.osu.cz/irafm/ps/rep70.ps.gz

[7] HABIBALLA, H. Resolution Based Reasoning in Description Logic. In Proc. of Intl. Conf. ZNALOSTI 2006, Univ. of Hradec Kralove, 2006, also available as research rep. at http://ac030.osu.cz/irafm/ps/rep66.ps.gz.

[8] HABIBALLA, H. Fuzzy Predicate Logic Generalized Resolution Deductive System. Technical Report, Institute for Research and Application of Fuzzy Modeling, University of Ostrava, 2006.

[9] HÁJEK, P. Metamathematics of fuzzy logic. Kluwer Academic Publishers - Dordrecht, 2000.

[10] HÁJEK, P. Making fuzzy description logic more general. Fuzzy Sets and Systems 154(2005),pp. 1-15.

[11] NOVÁK, V., PERFILIEVA, I., MOČKOŘ, J. Mathematical principles of fuzzy logic. Kluwer, 1999.

Current address

Habiballa Hashim, RNDr. PaedDr., PhD.
Institute for research and application of fuzzy modeling Science
University of Ostrava,
30.dubna 22, 70200 Ostrava, Czech Republic
e-mail: hashim@seznam.cz

Journal of Applied Mathematics

volume IV (2011), number III

LINGUSTIC IF-THEN RULES FOR TIME SERIES PREDICTION

HABIBALLA Hashim, (CZ), PAVLISKA Viktor, (CZ)

Abstract. The article shows the system developed on University of Ostrava to provide new and efficient methods for time series prediction. It is based on Logical deduction through fuzzy lingustic rules and therefore it provide transparent predictions for following analysis.

Key words and phrases. time series, fuzzy logic.

Mathematics Subject Classification. Primary 37M10; Secondary 94D05.

1 Introduction

Time series prediction is an important task in many areas including economics, biology etc. There are many standard methods to solve this task like Winter's method for time series prediction. These methods have one essential imperfection - they are not transparent. It means if a prediction is produced it cannot be observed why and how such an result has been obtained or such explication of results is described in hardly-readable mathematical relations. We present a new method which overcomes this imperfection and it is based on already very successful formalism of fuzzy linguistic IF-THEN rules presented and developing by the group of Prof. Novak[1].

1.1 Fuzzy Linguistic Rules

The theory of linguistic term and variables is well-known approach in the fuzzy logic community. It enables to work with rules containing terms of natural language like small or big and modifiers like very, roughly etc. The rule interpretation is then done by logical deduction based on which is based on fuzzy set theory and fuzzy logic to enable to deduce conclusions on the basis of

imprecise description of the given situation using the linguistically formulated fuzzy IF-THEN rules [1]. The usage of this theory within a frame of time series prediction lies in the learning of these rules from the serie and then application to future (predicted) members of the serie. These learning algorithms are already prepared within the LFLC software [1], which is intended to perform logical deduction on linguistic rules. The core of the system serve also for the presented application.

1.2 Linguistic descriptions and their elaboration

The theoretical background of LFLC lays in formal fuzzy logic in broader sense (FLb), which is an extension of that in narrow sense (FLn) (for the detailed presentation of both logics see [2]). The theory provides elaboration of that part of the semantics, which consists of the so called *evaluating* and *conditional linguistic expressions*. The former are expressions such as "small, roughly medium, very big", etc. The latter are the well known fuzzy IF-THEN rules. These are usually gathered into sets called *linguistic descriptions* which take the form

$$\begin{aligned} \mathcal{R}_1 &:= \mathsf{IF}\ X\ \mathsf{is}\ \mathcal{A}_1\ \mathsf{THEN}\ Y\ \mathsf{is}\ \mathcal{B}_1 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathcal{R}_m &:= \mathsf{IF}\ X\ \mathsf{is}\ \mathcal{A}_m\ \mathsf{THEN}\ Y\ \mathsf{is}\ \mathcal{B}_m \end{aligned}$$

where $\mathcal{A}_j, \mathcal{B}_j$ are the mentioned evaluating linguistic expressions. They characterize property of some features of objects, for example size, volume, force, strength, etc. Since usually we are not interested in the concrete objects and their features, we replace them by some real numbers which are then represented by the variables X and Y. Thus, values of X and Y represent, e.g. values of temperature, pressure, price, etc. The linguistic expression of the form 'X is $\mathcal{A}$' is called the *evaluating linguistic predication*.

Fuzzy IF-THEN rules serve as a basis for *approximate reasoning*, which is a method for finding a conclusion on the basis of the imprecise initial information concentrated in the form of linguistic description and some new information. There are two fundamental approximate reasoning methods:

(a) *Linguistically based fuzzy logical deduction*, i.e. finding a formal conclusion when the fuzzy IF-THEN rules are treated as linguistically characterized logical implications.

(b) *Fuzzy approximation of a function*, i.e. finding a function which approximates some only imprecisely known function, whose course is estimated using the linguistic description.

The interpretation of the linguistic description significantly depends on the above chosen method.

The usual implementations of approximate reasoning focuses on the method (b). Our concept of LFLC implements both methods but its main strength lays in the method (a).

1.3 Fuzzy approximation of a function

In this case, each evaluating predication 'X is $\mathcal{A}$' is assigned some formula $A(x)$ of predicate fuzzy logic. The whole linguistic description is then assigned one of two special formulas called the disjunctive and conjunctive normal form.

The *disjunctive normal form* is the formula

$$\mathrm{DNF}(x,y) := \bigvee_{j=1}^{m} (A_j(x) \wedge B_j(y)). \tag{1}$$

In this case, each rule is assigned a conjunction of formulas $A_j(x)$ and $B_j(y)$ and all of them are joined by disjunction. We speak also about *functional interpretation* of the linguistic description.

The alternative possibility is the *conjunctive normal form*

$$\mathrm{CNF}(x,y) := \bigwedge_{j=1}^{m} (A_j(x) \Rightarrow B_j(y)). \tag{2}$$

In this case, each rule is assigned an implication between the formulas $A_j(x)$ and $B_j(y)$ and all of them are joined by conjunction. We speak about *logical* interpretation of the linguistic description. Recall, however, that the main goal is still fuzzy approximation of a function.

1.4 Linguistically based fuzzy logical deduction

The most specific feature of LFLC is the possibility to realize a *fuzzy logical deduction* when the rules are interpreted as *linguistically characterized* logical implications.

1.4.1 Linguistic aspect

In the concept of LFLC, we deal with the mentioned *evaluating linguistic expressions* (possibly with signs) which have the general form

$$\langle \text{linguistic modifier} \rangle \langle \text{atomic term} \rangle \tag{3}$$

where $\langle$atomic term$\rangle$ is one of the words "small, medium, big", or "zero" (possibly also arbitrary symmetric fuzzy number) and $\langle$linguistic modifier$\rangle$ is an intensifying adverb such as "very, roughly", etc.

The linguistic modifiers in (3) are of two basic kinds, namely those with narrowing and widening effect. *Narrowing* modifiers are, for example, "extremely, significantly, very" and *widening* ones are "more or less, roughly, quite roughly, very roughly". We will take these modifiers as canonical. Note that narrowing modifiers make the meaning of the whole expression more precise while widening ones do the opposite. Thus, "very small" is more precise than "small", which, on the other hand, is more precise than "roughly small".

The meaning of each linguistic expression $\mathcal{A}$ has two constituents: the *intension* $\mathrm{Int}(\mathcal{A})$ and *extension* $\mathrm{Ext}(\mathcal{A})$ in some *model* (this is often called the *possible world*).

Intension of the linguistic expression is a formal characterization of the property denoted by it on the level of formal syntax. It can be interpreted as a fuzzy set of special formulas[†)].

[†)] They have the form $\mathbf{A} := \{a_t/A_x[t] \mid t \in M\}$ where $A(x)$ is a formula, M is a set of constants and a_t is an evaluation of the instance $A_x[t]$. For the details — see [1].

However, it is a rather abstract concept, which in concrete situation (context) determines some fuzzy set of elements. Mathematically this means that a model w is given whose support is some set U (taken usually as a closed interval of real numbers). Then the extension of $\mathcal{A}$ is some fuzzy set of elements $\mathrm{Ext}_w(\mathcal{A}) \subsetneq U$, which is determined by its intension $\mathrm{Int}(\mathcal{A})$. Note that for each concrete situation, different model should be considered. However, intension is still the same.

Note that the concepts of intension and extension formalizes the following intuitive situation: we can speak about *high temperature*, *high pressure*, *high tree*, etc. But high temperature may mean 100°C at home or 1000°C in metal melting process, and similarly in other cases. This cannot be satisfactorily formalized without the mentioned concepts.

In the terminology used in LFLC, we speak about *linguistic context* in which the given evaluating expression is used since in the practice, it requires setting the *minimal* and *maximal* possible values which can be attained by the used variables.

Let us stress that the extensions of the evaluating expressions are fuzzy sets of the form of the so called S- and Π-curves, as is depicted on Figure 1. More on the formal theory of evaluating linguistic expressions can be found in [1].

1.4.2 Fuzzy logical deduction

Unlike fuzzy approximation, where we deal with fuzzy sets in a model (i.e. on the level of semantics), logical deduction must proceed on syntax. Instead of the detailed formal description, we will demonstrate the behaviour of the logical deduction on an example.

Let us consider a linguistic description consisting of two rules:

$$\mathcal{R}_1 := \mathsf{IF}\ X\ \mathsf{is\ small}\ \ \mathsf{AND}\ Y\ \mathsf{is\ small}\ \ \mathsf{THEN}\ Z\ \mathsf{is\ big}$$
$$\mathcal{R}_2 := \mathsf{IF}\ X\ \mathsf{is\ big}\ \ \mathsf{AND}\ Y\ \mathsf{is\ big}\ \ \mathsf{THEN}\ Z\ \mathsf{is\ small}.$$

These rules are assigned intensions $\mathrm{Int}(\mathcal{R}_1), \mathrm{Int}(\mathcal{R}_2)$, which can schematically be written as

$$\mathrm{Int}(\mathcal{R}_1) = (\mathbf{Sm}_x \wedge \mathbf{Sm}_y) \Rightarrow \mathbf{Bi}_z \tag{4}$$
$$\mathrm{Int}(\mathcal{R}_2) = (\mathbf{Bi}_x \wedge \mathbf{Bi}_y) \Rightarrow \mathbf{Sm}_z. \tag{5}$$

Furthermore, let X, Y, Z be interpreted in a model which will consist of three sets $U = V = W = [0, 1]$. Then small values are some values around 0.3 (and smaller) and big ones some values around 0.7 (and bigger). Of course, given the input, e.g. $X = 0.3$ and $Y = 0.25$ then we expect the result $Z \approx 0.7$ due to the rule $\mathcal{R}_1$. Similarly, for $X = 0.75$ and $Y = 0.8$ we expect the result $Z \approx 0.25$ due to the rule $\mathcal{R}_2$.

The value 0.3 is represented in the formal system by a certain intension $\mathbf{Sm}'_x$ and similarly, the value 0.25 is represented by $\mathbf{Sm}'_y$.

Then the inference rule of modus ponens is applied on $\mathbf{Sm}'_x$, $\mathbf{Sm}'_y$ and the implication (4). The result is the intension $\mathbf{Bi}'_z$. The latter is to be interpreted as some fuzzy set $B' \subsetneq W$.

To obtain one concrete value, the resulting fuzzy set B' should further be defuzzified. However, we deal with evaluating linguistic expressions, whose interpretation has always one of the three possible forms depicted on Figure 1. Therefore, standard defuzzification methods such as COG do not work properly. Instead, we have developed a special method, which we call

Figure 1: Form of fuzzy sets corresponding to the meaning of the evaluating linguistic expressions and the DEE defuzzification.

Defuzzification of Evaluating Expressions (DEE). This method classifies first the type of the membership function and then decides the defuzzification, as is depicted on Figure 1. There are two versions of the DEE method, namely *simple* which first classifies the resulting fuzzy sets in types "small", "medium" and "big" and then defuzzifies it using Last of Maxima, Center of Gravity or First of Maxima methods, respectively. The second one uses a sophisticated algorithm to choose a value close to these dependingly on the specific shape of the membership function.

In our case, when the input is $X = 0.3$ and $Y = 0.25$ then both values correspond to "small" and thus, with respect to the rule $\mathcal{R}_1$, the resulting linguistic corresponds to "big" and thus, after its interpretation in the model and defuzzification using the DEE method, we obtain the result $Z \approx 0.7$, i.e. a value being intuitively big. In other words, we obtain the result which, on the basis of the form of the given rules, should be expected. Similarly, the input values $X = 0.75$ and $Y = 0.8$ would lead to the value $Z \approx 0.25$ due to the rule $\mathcal{R}_2$.

To summarize: in the case of fuzzy approximation, we form the special formulas DNF or CNF on the level of syntax, interpret them in some model and then find the approximation on the level of semantics only. In the case of linguistically based fuzzy logical deduction we interpret the rules on the level of syntax, transform measurement also to this level, realize formal logical deduction and then interpret the result in some model.

2 Time Series Prediction Application

The linguistic inference is used to predict future values and the main advantage lies in explicit statement of the used rules for prediction. This makes the principle unique above standard methods. The IF-THEN rules are in the following form according to the theory stated above.

1	me	vr sm	$\longrightarrow$	ze
2	qr bi	me	$\longrightarrow$	ze
3	-ex bi	qr bi	$\longrightarrow$	ze
4	sm	-ex bi	$\longrightarrow$	ze
5	sm	sm	$\longrightarrow$	ze
6	me	sm	$\longrightarrow$	ze
7	-sm	me	$\longrightarrow$	ve sm
8	ex sm	-sm	$\longrightarrow$	sm
9	-ra sm	ex sm	$\longrightarrow$	ml me
10	ml me	-ra sm	$\longrightarrow$	ml me
11	-vr bi	ml me	$\longrightarrow$	qr bi
12	ml me	-vr bi	$\longrightarrow$	ml me
13	-ve bi	ml me	$\longrightarrow$	ml me
14	ro bi	-ve bi	$\longrightarrow$	ml me

3 Conclusion

The presented method is unique approach to predict time series. It enables to see what rules were used for prediction, which other methods cannot provide. The approach is moreover implemented like a software LFLForecaster, that is described in separated article.

Acknowledgement

The authors have been partially supported by the project MSM6198898701 and 1M0572 of the MSMT CR.

References

[1] DVORAK ET. AL.: The concept of LFLC 2000 - its specificity, realization and power of applications. Computers in Industry. 03/2003(51), Elsevier, Amsterdam, 2003, pp.269-280.
[2] KRIVY, I.: Time-series. University of Ostrava, Ostrava, 2005 (in czech).
[3] PERFILIEVA, I., NOVAK, V., DVORAK, A. Fuzzy transform in the analysis of data. INT J APPROX REASON. 2008, sv. 48, s. 36-46. ISSN 0888-613X..
[4] PERFILIEVA, I., NOVAK, V., PAVLISKA, V., DVORAK, A., STEPNICKA, M. Analysis and Prediction of Time Series Using Fuzzy Transform. WCCI 2008 Proceedings. Hong Kong: IEEE Computational Intelligence Society, 2008. s. 3875-3879.

Current address

HABIBALLA Hashim, RNDr., PaedDr., Ph.D., PhD.
Institute for Research and Application of Fuzzy Modeling, University of Ostrava, 30. dubna 22, 70103 Ostrava 1, CZ, tel.: +420608886130, email: hashim.habiballa@osu.cz

PAVLISKA, Viktor, RNDr., Ph.D.
Institute for Research and Application of Fuzzy Modeling, University of Ostrava, 30. dubna 22, 70103 Ostrava 1, CZ,

Journal of Applied Mathematics

volume IV (2011), number III

IJK – ALGORITHM TO CALCULATE THE INTERVAL RELIABILITY

KARPÍŠEK Zdeněk, (CZ), LACINOVÁ Veronika, (CZ)

Abstract. The paper presents a methodology and an IJK – algorithm to calculate an interval estimate of a system reliability function using statistical or expert interval estimators of the reliability functions of mutually independent system elements. The interval reliability calculations are based on interval arithmetic and the algorithm uses the method of list and the neighbourhood matrix of the graph of a given reliability system.

Keywords. reliability of system, interval arithmetic, interval reliability, IJK - algorithm

Mathematics Subject Classification: Primary 62N05, 65G40; Secondary 90B25.

1 Reliability of a system of mutually independent elements

A collection of objects used to carry out required activities is usually referred to as a ***system***. When analysed, complicated systems can be partitioned into functional units (***subsystems***) simpler in terms of the activities in question with these being further decomposed down to indivisible parts called system ***elements***. If the required activity is represented by the reliability of elements, we call this a ***reliability system*** [3].

The structure of a reliability system decomposed into elements is most often described by what is called a ***flow chart***. Further we assume a ***two-mode model*** with the system (an element) being either in a ***failure-free mode*** (logical value of 1) or in a ***failure mode*** (logical value of 0). If necessary, we can identify the system denotation with a logical variable A marking its mode with individual elements denoted by $A_1,\ldots,A_n$. The system structure can also be represented by a ***directed graph*** with the arcs of the graph corresponding to the system elements and the nodes to the connections of elements.

If the mode of element A_k does not influence the mode of element A_l and vice versa $(k \neq l)$, we say that the elements A_k, A_l are ***independent***. If the mode of a subset of elements has no effect on

the mode of any other subset of elements of the same system and the two sets are disjoint, we say that that the system elements are ***mutually independent***.

Generally, the mode of a (system) element depends on time t so that it is a function $A(t)$ assuming the values 1 and 0 with $t \in [0,\infty)$ and $A(0)=1$. We assume the mode $A(t)$ passing from 1 to 0 (not reversely) meaning that this is a model ***without renewal***. Next we assume that the time of failure-free operation is a non-negative random variable T and its ***reliability function*** (***survival function*** or ***reliability***) is $R(t)=P(T \geq t)=P(A(t)=1)$. With these assumptions, the mode of an element (system) is a random event.

Reliability systems with elements A_i, $i=1,\ldots,n$ used most frequently:

1. ***Serial system*** $A_S = A_1 \wedge \ldots \wedge A_n$.
2. ***Parallel system*** $A_P = A_1 \vee \ldots \vee A_n$.
3. ***Combined system*** A_K formed by serial and parallel subsystems repeatedly connected in series or in parallel.
4. ***k of n system*** $A_{k/n}$ is in failure-free mode if <u>at least</u> k elements of n elements of A_i is in failure-free mode ($k \leq n$). A special case of this is the serial system ($k=n$) and the parallel system ($k=1$).

Here $\wedge$ denotes the logical conjunction and $\vee$ the logical disjunction of element modes corresponding to the operations of intersection $\cap$ and union $\cup$ of random events.

2 Interval reliability

Definition 2.1. Let $[a,b], a \leq b, (a,b) \in \mathbb{R}^2$ be an ***interval number*** [1], then the following ***arithmetic operations with interval numbers*** are defined:

$$
\begin{aligned}
&[a,b]+[c,d]=[a+c,b+d],\\
&[a,b]-[c,d]=[a-d,b-c],\\
&[a,b]\cdot[c,d]=[\min\{ac,ad,bc,bd\},\max\{ac,ad,bc,bd\}],\\
&[a,b]/[c,d]=[a,b]\cdot[1/d,1/c] \text{ pro } 0 \notin [c,d].
\end{aligned}
$$

For $\forall a \in \mathbb{R}$ we put $a=[a,a]$. If $a>0$, we write $[a,b]>0$ etc.

Theorem 2.1. If J, K, L, M are interval numbers, we have:

$$
\begin{aligned}
&J+K=K+J,\ J+(K+L)=(J+K)+L,\\
&J\cdot K=K\cdot J,\ J\cdot(K\cdot L)=(J\cdot K)\cdot L,\\
&0+J=J,\ 1\cdot J=J,\\
&J\cdot(K+L)\subset(J\cdot K)+(J\cdot L).
\end{aligned}
$$

Specifically for $K \cdot L > 0$, we have

$$J \cdot (K + L) = J \cdot K + J \cdot L.$$

If $J \subset L$ and $K \subset M$, then

$$\begin{aligned} &J + K \subset L + M, \\ &J - K \subset L - M, \\ &J \cdot K \subset L \cdot M, \\ &J / K \subset L / M, \text{ if } 0 \notin M. \end{aligned}$$

For $J = [a,b] \geq 0$, $K = [c,d] \geq 0$, we have $J \cdot K = [ac,bd]$, and for $J = [a,b] \leq 0$, $K = [c,d] \leq 0$, we have $J \cdot K = [bd,ac]$.

Proof. See [1]. □

Definition 2.2. Let $f(x_1,...,x_n)$ be a real function and $I_1,...,I_n$ interval numbers, then by an ***interval value*** of this function in $(I_1,...,I_n)$, we mean the interval number

$$[\min f(x_1,...,x_n), \max f(x_1,...,x_n)]$$

with $(x_1,...,x_n) \in I_1 \times \cdots \times I_n$ speaking about an ***interval function*** $f(I_1,...,I_n)$.

Remark 2.1. If $f(x_1,...,x_n)$ is increasing for each of its independent variables, then evidently

$$[\min f(x_1,...,x_n), \max f(x_1,...,x_n)] = [f(\min I_1,...,\min I_n), f(\max I_1,...,\max I_n)].$$

Definition 2.3. ***The interval reliability*** of an element or system at time $t \in [0,\infty)$ is defined as an interval number $[R_1(t), R_2(t)]$ with $R_1(t), R_2(t)$ being the reliabilities (probabilities of a failure-free mode) of this element or system at time *t*.

Theorem 2.2. Let the elements A_i of a reliability system be mutually independent with interval reliabilities $[R_{i1}(t), R_{i2}(t)]$, $i = 1,\ldots,n$. Then the following is true:

1. The interval reliability of a serial system $A_S = A_1 \wedge \ldots \wedge A_n$ is

$$[R_1(t), R_2(t)] = \left[\prod_{i=1}^{n} R_{i1}(t), \prod_{i=1}^{n} R_{i2}(t)\right].$$

2. The interval reliability of a parallel system $A_P = A_1 \vee \ldots \vee A_n$ is

$$\left[R_1(t), R_2(t)\right] = \left[1 - \prod_{i=1}^{n}\left(1 - R_{i1}(t)\right), 1 - \prod_{i=1}^{n}\left(1 - R_{i2}(t)\right)\right].$$

3. The interval reliability of a combined system A_K with each element A_i, $i = 1, \ldots, n$ occurring in exactly one (arbitrary) subsystem can be obtained by gradually calculating the interval reliabilities of the serial and parallel subsystems using the equalities under 1 and 2.

4. The interval reliability of a *k* of *n* system $A_{k/n}$ is

$$\left[R_1(t), R_2(t)\right] =$$

$$= \left[\sum_{j=k}^{n}\sum_{\{i_1,\ldots,i_j\}} R_{i_1 1}(t)\cdots R_{i_j 1}\left(1 - R_{i_{j+1} 1}\right)\cdots\left(1 - R_{i_n 1}(t)\right), \sum_{j=k}^{n}\sum_{\{i_1,\ldots,i_j\}} R_{i_1 2}(t)\cdots R_{i_j 2}\left(1 - R_{i_{j+1} 2}\right)\cdots\left(1 - R_{i_n 2}(t)\right)\right]$$

where $\{i_1, \ldots i_j\}$ are all the class *j* combinations of the set of indices $\{1, \ldots, n\}$.

Proof. Assertions are a direct consequence of the reliabilities of the systems being increasing functions of the reliabilities of individual elements [2]. □

Example 2.1. The following simple problems were solved in Excel using the list method. In the problem description, the interval estimates of the elements are defined by the minimal and maximal values $[\text{MIN}, \text{MAX}]$ with the list containing, rather than the disjoint chains of the conjunctions of the possible system modes, directly the corresponding products of their probabilities expressed in terms of the reliabilities *R*1, *R*2, *R*3 of mutually independent elements. The interval reliabilities of the chains were calculated using the interval arithmetic, more specifically, as the product of interval numbers putting, however, $1 - [a, b] = [1 - a, 1 - b]$. This corresponds to the fact that the reliability of each system in question is an increasing function of each of its arguments, that is, the reliabilities of the given system elements. Table 2.1 and Table 2.2 record the setting and computed results.

Table 2.1. Definition and list for Excel

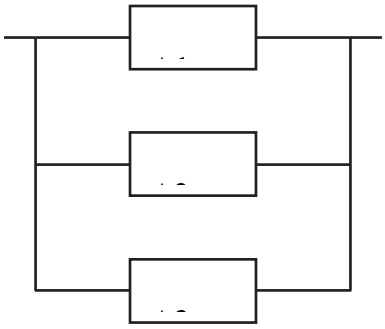
Elements	Relia-bility	MIN	MAX	*j*	List	MIN	MAX
*A*1	*R*1	0.4	0.8	1	*R*1 *R*2 *R*3	0.168	0.504
*A*2	*R*2	0.7	0.9	2	(1 - *R*1) *R*2 *R*3	0.252	0.126
*A*3	*R*3	0.6	0.7	3	*R*1 (1 - *R*2) *R*3	0.072	0.056
				4	*R*1 *R*2 (1 - *R*3)	0.112	0.216
				5	(1 - *R*1) (1 - *R*2) *R*3	0.108	0.014
				6	(1 - *R*1) *R*2 (1 - *R*3)	0.168	0.054
				7	*R*1 (1 - *R*2) (1 - *R*3)	0.048	0.024
				8	(1 - *R*1) (1 - *R*2) (1 - *R*3)	0.072	0.006
					Sum	1.000	1.000

Table 2.2. Computed interval reliability of systems

Serial system *S*:

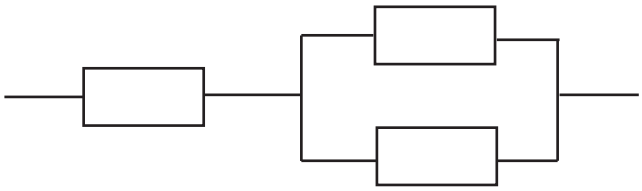
j	Mode *S*	MIN	MAX
1	1	0.168	0.504
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
Reliability	*R*(*S*)	0.168	0.504

Parallel system *P*:



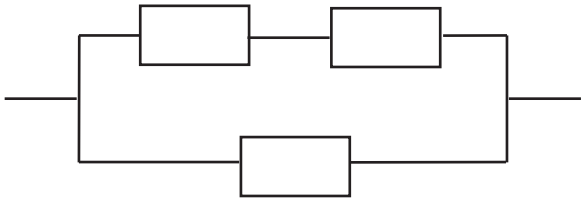
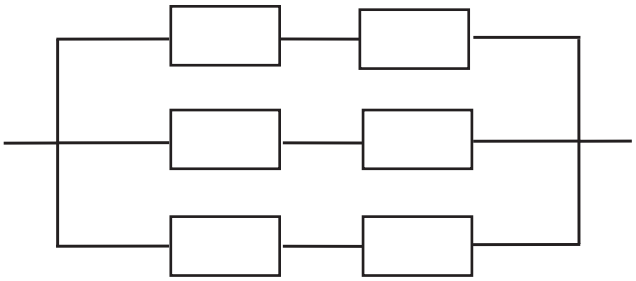
j	Mode *P*	MIN	MAX
1	1	0.168	0.504
2	1	0.252	0.126
3	1	0.072	0.056
4	1	0.112	0.216
5	1	0.108	0.014
6	1	0.168	0.054
7	1	0.048	0.024
8	0	0	0
Reliability	*R*(*P*)	0.928	0.994

Combined system *K*1:



j	Mode *K*1	MIN	MAX
1	1	0.168	0.504
2	0	0	0
3	1	0.072	0.056
4	1	0.112	0.216
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
Reliability	*R*(*K*1)	0.352	0.776

Table 2.2. Computed interval reliability of systems (continued)

Combined system *K*2: 	*j*	Mode *K*2	MIN	MAX
	1	1	0.168	0.504
	2	0	0	0
	3	0	0	0
	4	1	0.112	0.216
	5	1	0.108	0.014
	6	0	0	0
	7	0	0	0
	8	0	0	0
	Reliability	*R*(*K*2)	0.388	0.734
"2 of 3" system (i.e., at least 2 of 3) *A*2/3: 	*j*	Mode *A*2/3	MIN	MAX
	1	1	0.168	0.504
	2	1	0.252	0.126
	3	1	0.072	0.056
	4	1	0.112	0.216
	5	0	0	0
	6	0	0	0
	7	0	0	0
	8	0	0	0
	[1] **Reliability**	*R*(*A*2/3)	0.604	0.902

A direct calculation of the interval reliability of a combined system by can, however, only be used for not very large reliability systems. Then the following algorithm may be applied.

3 IJK – algorithm to calculate the interval reliability of a combined system

To calculate the interval reliability of a combined system, a ***JK-algorithm*** calculating the reliability of the system [4] and combining a ***list method*** and a ***method of paths*** was modified.

A list method sets up a method of all the logical events in a system to calculate the system reliability using disjoint random events. In our case, a path in a directed graph is defined as a sequence of arcs corresponding to system elements that connect nodes between the input and output graph nodes.

Suppose that a reliability system A is expressed by means of a simple acyclic directed graph with a ***neighbourhood matrix*** $\mathbf{A} = (a_{kl})_{k,l=1}^{m}$ where exactly one arc a_{kl} from node k to node l corresponds to a system element A_i. If the nodes k and l are connected with an arc, we put $a_{kl} = 1$ with $a_{kl} = 0$ otherwise. If the system graph is not simple (the system contains a parallel subsystem),

it can be transformed into a simple one using, for example, ***arc splitting*** [4]. Denoting by 1 the ***input node*** and by m the ***output node*** of the graph, we see that $1 \le m-1 \le n \le (m-1)m/2$.

To determine the interval reliability of a system A of mutually independent elements, we assume that the elements of a system A_i have interval reliabilities $\left[R_{i1}(t), R_{i2}(t)\right]$, $i = 1,...,n$. The interval reliability of a combined system can be determined for $\forall t \in [0,\infty)$ by implementing the steps of what is referred to as an ***IJK – algorithm***:

1. Generate a list of all possible modes of the elements of a given system in the form of a type $(2^n, n)$ matrix **V** with its rows formed by all class *n* variations with repetition of the two-element set {0;1} (they can be viewed as binary numbers from 0 to $2^n - 1$).
2. For each variation of the modes of system elements, calculate the mode of the system $S_j = \operatorname{sgn}(D_{m1})$, $j = 1,...,2^n$ using the algebraic complement D_{m1} of the matrix $\mathbf{D} = \mathbf{E} - \mathbf{A}$ where **E** denotes the unit matrix and substituting the value of the mode of the corresponding element A_i from matrix **V** for entry a_{kl} of the neighbourhood matrix **A**.
3. Determine the interval reliability of the system $\left[R_1(t), R_2(t)\right]$ at time *t* using the interval reliabilities $\left[R_{i1}(t), R_{i2}(t)\right]$ of elements $i = 1,...,n$, and system modes $S_j = \operatorname{sgn}(D_{m1})$, $j = 1,...,2^n$ using the equality

$$\left[R_1(t), R_2(t)\right] =$$

$$= \left[\sum_{j=1}^{2^n}\left\{S_j \prod_{i=1}^{n}\left[\left(1-R_{i1}(t)\right)^{1-A_i}\left(R_{i1}(t)\right)^{A_i}\right]\right\}, \sum_{j=1}^{2^n}\left\{S_j \prod_{i=1}^{n}\left[\left(1-R_{i2}(t)\right)^{1-A_i}\left(R_{i2}(t)\right)^{A_i}\right]\right\}\right],$$

putting $0^0 = 1$.

The interval reliability of the system thus calculated can be used to calculate further interval functional and numeric characteristics of the system: distribution function, probability density, failure intensity, mean time to failure, quantiles etc.

4 Conclusion

If the estimators of the reliability of system elements are expert ones or based on experience, there is no need to be concerned too much with the significance or weight of an interval estimator of the system reliability. If, however, the interval estimators of the reliability of mutually independent elements A_i have confidence levels (statistical reliabilities) $1-\alpha_i$, $i = 1,...,n$, then the confidence level of the interval estimator of the system reliability and of the system's other functional and numeric reliability characteristics is $\prod_{i=1}^{n}(1-\alpha_i)$. Specifically, for elements with identical confidence level $1-\alpha^*$, the confidence level of the system interval estimator is $(1-\alpha^*)^n$. Thus, if the

confidence level of the interval estimator of the system reliability (and other characteristics of the system) is to be equal to $1-\alpha$, the confidence level of the interval estimator of the reliability of each element needs to be chosen as $1-\alpha^* = (1-\alpha)^{1/n}$. Table 4.1 contains the representative values of confidence level.

Table 4.1. Confidence level $1-\alpha^*$ of element

n	$1-\alpha$		n	$1-\alpha$	
	0.95	0.99		0.95	0.99
1	0.9500000	0.9900000	10	0.9948838	0.9989955
2	0.9746794	0.9949874	50	0.9989747	0.9997990
5	0.9897938	0.9979920	100	0.9994872	0.9998995

Since $1-\alpha^* > 1-\alpha$ for $n>1$, the determining of the required confidence level of the interval estimator of the system reliability requires a higher confidence level of the interval estimator of each element. This, in turn, leads to an increase in the length of the interval estimate of the entire system.

The calculation illustrated in section 2 concerns the systems with an increasing reliability with respect to each system element. In the event of an "unnatural" reliability system with the system reliability decreasing as the reliability of an element increases, the resulting interval reliability of the system can be obtained by some of the optimization methods. This actually involves finding the absolute minimum and maximum of a certain multilinear (polynomial) function on the Cartesian product (hyperblock) of the element reliabilities. The IJK-algorithm described in section 3 is a special variant of the calculation of the fuzzy reliability of a combined system using what is referred to as the ***FJK – algorithm*** [5]. A computer program [6] for this algorithm was written also calculating the numeric values of the interval reliability function of a combined system with fuzzy Weibull probability distribution as well as the system's further interval characteristics from defined or estimated interval reliability functions of mutually independent system elements.

Acknowledgement

The paper was supported by a MSMT of the Czech Republic project no. 1M06047, „Center of Quality and Reliability of Production“, grant from the Grant Agency of the Czech Republic (Czech Science Foundation) reg. no. 103/08/1658, „Advanced Optimum Design of Composed Concrete Structures”, and research project no. 3, „Management Support of Small and Middle-Sized Firms Using Mathematical Methods” of Academy Sting, Business College in Brno.

References

[2] MOOR, R. E., KEARFOTT, R. B., CLOUD, M. J., *Introduction to Interval Analysis*. Philadelphia: SIAM 2009. ISBN 978-0-898716-69-6.

[3] ROSS, S. M., *Introduction to Probability Models.* 9th ed. Amsterodam: Academic Press 2007. ISBN-13: 978-0-12-598062-3.
[4] KARPÍŠEK, Z., Stochastické modelování spolehlivosti. In *Celostátní seminář Analýza dat 2005/II*. Lázně Bohdaneč 2005, pp. 75-98, ISBN 80-239-6552-2.
[5] KARPÍŠEK, Z., JELÍNEK, P., DOSTÁL, P., Určení spolehlivosti systému pomocí jedné věty z teorie grafů. In *Sborník z 10. semináře Moderní matematické metody v inženýrství v Dolní Lomné u Jablunkova 30.5. - 1.6.2001*. Ostrava 2001, pp. 91-95, ISBN 80-248-0013-6.
[6] KARPÍŠEK Z., Fuzzy spolehlivost. In *Conference REQUEST ´06*. Praha, 30. 1. - 1. 2. 2007, pp. 164-177, ISBN 978-80-01-03709-6.
[7] MARTIŠEK, K., KARPÍŠEK, Z., *System Fuzzy Reliability*. Authorized software. Centre of Quality and Reliability of Production, Brno University of Technology, Brno, 2010.

Current address

Karpíšek Zdeněk, Doc. RNDr. CSc.
Centre for Quality and Reliability of Production (CQR), Department of Statistics and Optimization, Institute of Mathematics, Faculty of Mechanical Engineering, Brno University of Technology, Technická 2, 616 69 Brno, karpisek@fme.vutbr.cz

Lacinová Veronika, Ing.
Department of Statistics and Optimization, Institute of Mathematics, Faculty of Mechanical Engineering, Brno University of Technology, Technická 2, 616 69 Brno, v.neradova@email.cz

Journal of Applied Mathematics

volume IV (2011), number III

FUZZY TECHNIQUES FOR TIME SERIES PREDICTION

KOTYRBA Martin, (CZ), VOLNA Eva, (CZ), JANOSEK Michal, (CZ), KOCIAN Vaclav, (CZ), HABIBALLA Hashim, (CZ)

Abstract. This paper presents a comparison of two methods for time series prediction, LFLC (Linguistic Fuzzy Logic Controller) and FIS (Fuzzy Interface System). It is essential that both are applied to the time series prediction of market development. It is very difficult to predict the market behavior because the overall trend is determined by many factors and unforeseen circumstances. The main experiment of this paper is to compare results of prediction of two different methods of ranking fuzzy theory, the challenge is to create a forecast of the stock market.

Key words. Fuzzy theory, prediction, FIS, LFLC, linguistic rules.

Mathematics Subject Classification: Primary 93C42, 37M10 ; Secondary 91F20.

1 Stock Market Forecast

Various methods can be used for time series prediction. In addition to traditional methods, such as Box-Jenkins methodology, we can use neural networks, genetic algorithms or fuzzy logic. Algorithmic effort prediction models are limited by their inability to cope with uncertainties. Lots of researchers have been involved in the topic of fuzzy logic prediction [2], [3], [4], [5]. In this paper, we present a software prediction based on an appropriate combination of two fuzzy logic approaches, e.g. LFLC (Linguistic Fuzzy Logic Controller) and FIS (Fuzzy Interface System).

Time series analysis and prediction is an important task that can be used in many areas of practice. The task of getting the best prediction to the given series may bring interesting engineering applications in a wide range of areas like economics, biology or industry. Sometimes time series analysis and prediction is performed using chaos theory. Economic systems are complex and may be described by deterministic or stochastic models. The discovery that simple non-linear systems can show complex and chaotic dynamics has attracted some economists to work in this field. It is well known that chaotic time series are not long-term predictable due to their sensitive dependence on initial values. However, it is short-term predictable and prediction of chaotic time series is very important in real-world applications such as cash-flow forecasting. Based on the reconstructed state

space, one may introduce various approaches for prediction of chaotic time series. Market systems appear to be chaotic. One of the essential characteristics of a chaotic system is its extreme sensitivity to initial conditions. A tiny change in values at the beginning of the time series produces drastic changes in behavior. However, it is obvious that financial markets do not exhibit such extreme sensitivity to initial conditions. For example [4], stocks are sold or bought based on their prices. The price depends on how much has been bought or sold. The feedback loop has both positive and negative effects. The law of supply and demand implies a negative feedback loop, because the higher the price, the lower the demand, which, in fact, causes a lower price in the future. However, a parallel speculation mechanism implies a positive feedback loop, because an increasing price makes an assumption that the price will increase in the future and thus motivate the traders to buy more stocks. As we do not know the delay the between these two effects we are not able to predict anything well. These nonlinear effects are common in the markets. Nevertheless, it is fair to say that the markets are not purely chaotic. Although a chaotic system is a collection of orderly, simple behaviors, modeling the market has turned out to be more difficult than anticipated. The problem is that chaotic systems can be unusually flexible and rapidly switch between their many different behaviors. One way to isolate these individual behaviors of the market might lie in a presumption that when the market is perturbed in just the right way (e.g. a large drop in price), it would exhibit one of its many regular behaviors for a short time. In the meantime, constructing leading indicators is difficult, that is why a forecast horizon for any market time series is limited. The only possibility we can do with such a system is to adapt its behavior as quickly as possible.

2 Linguistic Fuzzy Logic Controller

This chapter describes a tool that has been created at the Institute for Research and Applications of Fuzzy Modeling (IRAFM) at the University of Ostrava. This is a really a really powerful application giving good results in some cases. The usage of this tool within the frame of time series prediction lies in learning linguistic rules from the series and then application to future predicted members of the series. These leasing algorithms are already prepared within the LFLC (Linguistic Fuzzy Logic Controller) software [5], which is intended to perform logical deduction on linguistic rules. The core of the system also serves for the presented application.

LFLC is specialized software (Fig.1), which is based on deep results obtained in formal theory of fuzzy logic. It makes it possible to deduce conclusions on the basis of imprecise description of the given situation using fuzzy IF-THEN rules (1). The rules are interpreted either as fuzzy relations or they can be taken as genuine linguistic expressions such as small, very large, medium etc. The rule interpretation is then done by logical deduction based on the fuzzy set theory and fuzzy logic to enable to deduce conclusions on the basis of imprecise description of the given situation using linguistically formulated fuzzy IF-THEN rules [5].

The theory of linguistic term and variables is a well-known approach in the fuzzy logic community.

The fuzzy IF – THEN rules are usually put together to form linguistic descriptions.

$$\begin{aligned} &R_1 := IF\ \mathrm{X}_1\ is\ \mathrm{A}_{11}\ AND\ \ldots\ AND\ \mathrm{X}_n\ is\ \mathrm{A}_{1n}\ THEN\ \mathrm{Y}\ \text{is}\ \mathrm{B}_1 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &R_m := IF\ \mathrm{X}_1\ is\ \mathrm{A}_{m1}\ AND\ \ldots\ AND\ \mathrm{X}_n\ \text{is}\ \mathrm{A}_{mn}\ THEN\ \mathrm{Y}\ \text{is}\ \mathrm{B}_m \end{aligned} \tag{1}$$

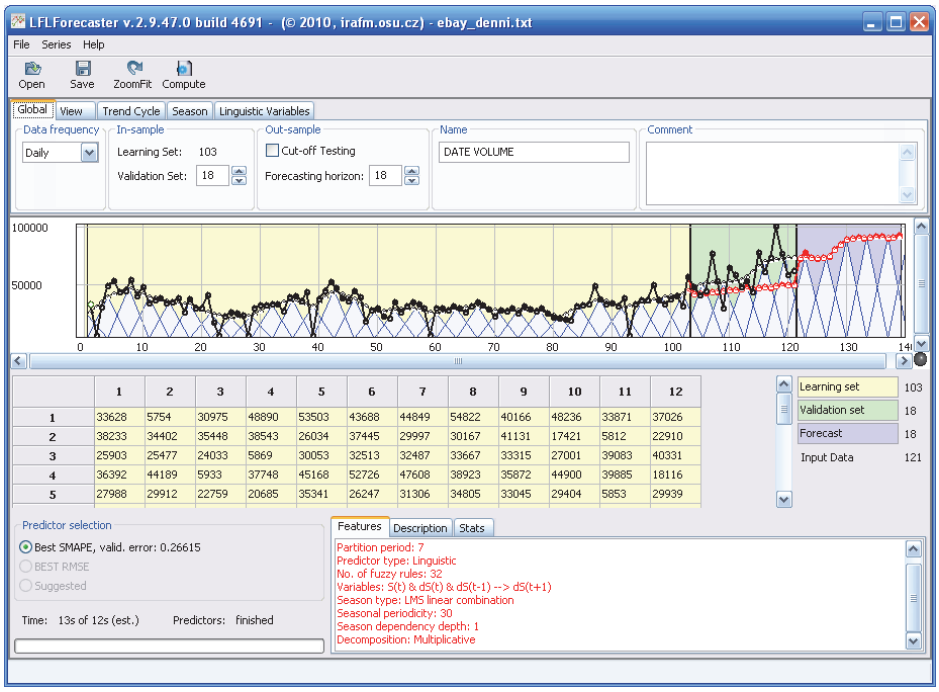


Figure 1. LFLC

3 Fuzzy Interface System

This chapter presents Fuzzy Inference System (FIS) that is included into Matlab - Fuzzy Logic Toolbox designed for time series prediction. FIS is based on the concepts of fuzzy set and fuzzy relations, which were defined by Lotfi A. Zadeh in 1965. Fuzzy sets generalize classical sets. Fuzzy Logic Toolbox contains Fuzzy Interface System (FIS) and allows working with fuzzy sets. It discusses the appropriate choice and use of the FIS (Sugeno type) for given time series prediction [1].

Fuzzy sets are sets whose elements have degrees of membership. Fuzzy sets introduce an extension of the classical notion of a set. In the classical set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition - an element either belongs (full membership in the set) or does not belong to the set (no membership in the set). A fuzzy set is a set which in addition to full or no membership allows partial membership. This means that the element belongs to the set with a certain degree of competence - level of competence. Function that assigns to each element of the universe is called the membership degree of membership functions. Given the classical set theory the degree of membership takes values in the range ⟨0, 1⟩.

Function μ_A is the membership functions of the fuzzy set A. Each element $x \in X$ assigns an element $\mu_A(x) \in \langle 0;1 \rangle$, which is called degree of membership of element x in fuzzy set A. If $\mu_A(x) = 0$ then x does not belong to fuzzy set A and if $\mu_A(x) = 1$ then x belongs to fuzzy set A. If $0 < \mu_A(x) < 1$ then x partially belongs to the fuzzy set A. Formal registration of a fuzzy set is the following (2):

$$A = \left(\mu_A(x_i) \Big/ x_i \right) \; for \; \forall x_i \tag{2}$$

There are more types of FIS. To solve our problem, we used FIS type P: u = R (e), where output values depend only on the size of the input values. The shape of the rules distinguishes between the FIS type: Mamdani and Sugeno. Mamdani FIS rules are described exclusively by means of fuzzy sets. To define the FIS we need to determine the following [1]:

- Number of the input variables (*n*)
- For each of them to determine the number and shape of the predefined input values that can be considered as model inputs
- Number of the output variables (*m*)
- For each of them to determine the predefined output value

Input and output values (which are considered in the form of fuzzy sets) are defined in the FIS rules follows (3).

$$\mathfrak{R}_k = \textbf{IF } x_1 \textbf{ is } A_{j,1}^k \textbf{ and } x_2 \textbf{ is } A_{j,2}^k \textbf{ and } \ldots \textbf{ and } x_n \textbf{ is } A_{j,n}^k \textbf{ then } z_k = B_j^k \tag{3}$$

Each rule determines the relationship between the selected input and output values. FIS can be regarded as fuzzy relations. When using FIS, input values are compared with the predefined input values. Based on this comparison and by FIS rules, we get the shape of the FIS output fuzzy sets [1]. Parameters that most affect the quality of the result are input variables, therefore very often used for debugging the matrix (4) that is used to create a language of values, rules, and to debug the FIS.

$$XY^L = \begin{pmatrix} x_{1,1}^L & \ldots & x_{1,n}^L & y_{1,1}^L & \ldots & y_{1,m}^L \\ x_{2,1}^L & \ldots & x_{2,n}^L & y_{2,1}^L & \ldots & y_{2,m}^L \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{K,1}^L & \ldots & x_{K,n}^L & y_{K,1}^L & \ldots & y_{K,m}^L \end{pmatrix} \tag{4}$$

4 Comparative experiments

Time series stock market development activities of company Ebay Corp. has been used in our prediction experiments. Data was downloaded from [http://www.forexrate.co.uk/historydates]. This is a 100-week data set depending on the volume, (see Fig. 2), and its following 20-week prediction. Results are reported in two different approaches that were compared mutually.

It is volume, after price, which is one of the most commonly quoted data points related to the stock market. Reflecting the overall activity in the stock or market, volume is the business of the market itself: buying and selling of shares. As such, volume is an important indicator for traders in analyzing market activity and planning strategy.

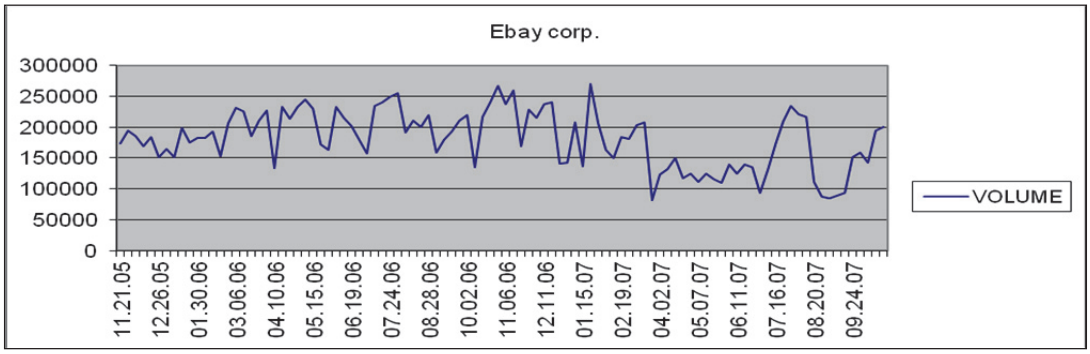


Figure 2. 100-week progress

We used the Sugeno type FIS for predicting time series volume indicators. FIS has been designed to predict one subsequent value on the basis of previous values. The requirement was to forecast the next 20 values. For 100 values were tuned Fuzzy Inference System as it was carried out using 20 projections. Clustering methods were used for week values in order to find a suitable base of fuzzy rules. There were approximately 20 estimates carried out on the data with using MAPE to select the most appropriate estimate to real values. The results are compared using MAPE (5), which gives a deviation of the predicted course from the real one.

$$MAPE = \frac{1}{K}\left(\sum_{h=1}^{K}(\mathrm{abs}(\,p_h - r_h\,)/r_h)\right) \tag{5}$$

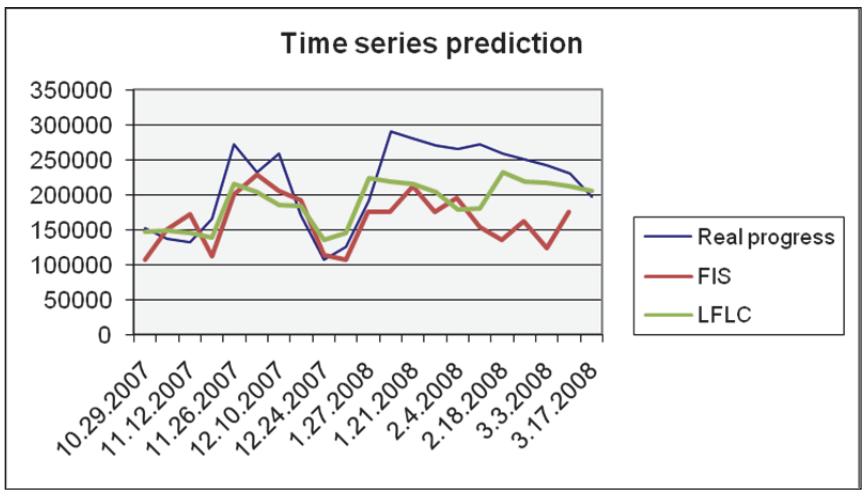


Figure 3. Comparison of prediction methods

LFLC predicts the development of series with MAPE = 0.19 and the FIS did not go for anything less and MAPE = 0.21 (Fig 3).

5 Conclusion

Our comparison is based on fuzzy technology and it will provide a very interesting look at the different possibilities of prediction. The aim of the article was to use FIS and LFLC methods based

on fuzzy logic to make difficult time series prediction of stock prices. Despite their different approaches, both methods show their ability to predict unusually large stock market development

Acknowledgements

The paper was supported by the University of Ostrava grant SGS/PřF/2011. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] ŽÁK, L Generalization of Fuzzy Clustering for Vaguely Defined Objects, *9th Zittau Fuzzy Colloquium*, Zittau, 2001, pp. 268-277, ISBN 3-9808089-0-4.
[2] ŽÁK, L. Fuzzy Inference System and Prediction. In International Conference on Soft Computing, Kunovice, Czech Republic: n, 2004. pp. 31-36. ISBN: 80-7314-025-X.
[3] KUO RJ, CHEN CH, HWANK YC:"An intelligent stock trading decision Support system Fuzzy neural network ", Fuzzy Sets and Systems, 118(1): pp. 21-45, 2001.
[4] PERFILIEVA, I., NOVÁK, V., PAVLISKA, V., DVOŘÁK, A., ŠTĚPNIČKA, M. Analysis and Prediction of Time Series Using Fuzzy Transform. WCCI 2008 Proceedings. Hong Kong: IEEE Computational Intelligence Society, 2008. pp. 3875-3879.
[5] DVOŘÁK, A., HABIBALLA H., NOVÁK, V., PAVLISKA, V. The concept of LFLC 2000. Computers in Industry. 03/2003(51), Elsevier, Amsterdam, 2003, pp.269-280.

Current address

Martin Kotyrba, Mgr.
University of Ostrava, Faculty of Science, Dept. of Informatics and Computers, 30.dubna 22,70103 Ostrava, Czech Republic, email: martin.kotyrba@osu.cz

Eva Volná, doc. RNDr. PaedDr. PhD.
University of Ostrava, Faculty of Science, Dept. of Informatics and Computers, 30.dubna 22,70103 Ostrava, Czech Republic, email: eva.volna@osu.cz

Michal Janošek, Mgr.
University of Ostrava, Faculty of Science, Dept. of Informatics and Computers, 30.dubna 22, 70103 Ostrava, Czech Republic, email: michal.janosek@osu.cz

Václav Kocian, Mgr.
University of Ostrava, Faculty of Science, Dept. of Informatics and Computers, 30.dubna 22, 70103 Ostrava, Czech Republic, email: vaclav.kocian@osu.cz

Hashim Habiballa, RNDr. PaedDr. PhD. Ph.D.
University of Ostrava, Faculty of Science, Dept. of Informatics and Computers, 30.dubna 22, 70103 Ostrava, Czech Republic, email: hashim.habiballa@osu.cz

Journal of Applied Mathematics

volume IV (2011), number III

ON AGGREGATION OF L-FUZZY REAL NUMBERS

ORLOVS Pavels, (LV)

Abstract. The paper is devoted to a general aggregation operator acting on L-fuzzy real numbers. The aim of our research is to analyze properties of a general aggregation operator depending on properties of the ordinary aggregation operator and the t-norm, which is used in the extension method. By using aggregation approach we describe some t-norm based operations with L-fuzzy real numbers and investigate their properties.

Key words and phrases. L-fuzzy real numbers, aggregation operator, T-extension.

Mathematics Subject Classification. 94D05, 03E72

1 Introduction

Aggregation of several input values into a single output value is an important tool of mathematics, physics, as well as of engineering, economical, social and other sciences. As the widely used examples of aggregation operators we can mention arithmetic and geometric mean, minimum and maximum operators, t-norms and others (see e.g. [4],[6]). The main object of our interest is an aggregation operator acting on non-negative L-fuzzy real numbers.

Our paper deals with a notion of L-fuzzy real numbers introduced by B. Hutton [2]. B. Hutton defined L-fuzzy numbers in the case, when L is the unit interval $[0, 1]$, but later some other authors ([7],[8],[9]) developed and extended his idea.

The notion of a general aggregation operator acting on fuzzy structures was introduced by A. Takaci in [3]. General aggregation operator is defined by using a t-norm T as a T-extension of an ordinary aggregation operator. The aim of our research is to analyze properties of the general aggregation operator $\tilde{A}$ acting on L-fuzzy real numbers depending on properties of the ordinary aggregation operator and the t-norm. In particular we consider such properties as associativity, symmetry, idempotence, existence of a neutral element. By using the extended

aggregation operator we consider t-norm based operations with L-fuzzy real numbers such as addition, maximum, minimum and investigate their properties. Initial results on this topic were presented at the conference FSTA in 2010 in Slovakia.

2 L-fuzzy real numbers

Let $L = (L, \wedge, \vee, 0_L, 1_L)$ be a completely distributive lattice, equiped with a t-norm T, where 0_L an 1_L are the least and the greatest elements of L.

Definition 2.1 *An L-fuzzy real number is a function $z\colon \mathbb{R} \to L$ such that*

(N1) z is non-increasing: $x_1 \geq x_2 \Longrightarrow z(x_1) \leq z(x_2)$;

(N2) z is bounded: $\bigwedge\limits_x z(x) = 0_L$, $\bigvee\limits_x z(x) = 1_L$;

(N3) z is left semi-continuous: $\bigwedge\limits_{t<x} z(t) = z(x)$.

The set of all L-fuzzy real numbers is called the L-fuzzy real line and it is denoted by $\mathbb{R}(L)$. In this paper we consider non-negative L-fuzzy real line $\mathbb{R}_+(L) = \{z \mid z(0) = 1_L,\ z \in \mathbb{R}(L)\}$, as well as extended non-negative L-fuzzy real line $\overline{\mathbb{R}_+}(L) = \mathbb{R}_+(L) \cup \{\tilde{1}\}$, where $\tilde{1} \equiv 1_L$. We denote the minimal element of $\overline{\mathbb{R}_+}(L)$ by

$$\tilde{\theta}(x) = \begin{cases} 1_L, & x \leq 0, \\ 0_L, & x > 0. \end{cases}$$

Operations with L-fuzzy real numbers such as addition $\oplus_T$ and multiplication by a positive real number $k \in \mathbb{R}_+$ are defined as following:

$$(z_1 \oplus_T \ldots \oplus_T z_n)(x) = \bigvee_{x=x_1+\ldots+x_n} T(z_1(x_1), \ldots, z_n(x_n)) \quad \text{and} \quad kz(x) = z\left(\frac{x}{k}\right),\ k > 0.$$

3 Aggregation operator

We start with the classical notion of an aggregation operator [4],[6]. Let us denote $\mathbb{I} = [0, 1]$ and consider the following definition.

Definition 3.1 *A mapping $A : \bigcup_n \mathbb{I}^n \to \mathbb{I}$ is called an aggregation operator if the following conditions hold:*

(A1) $A(0, \ldots, 0) = 0$;

(A2) $A(1, \ldots, 1) = 1$;

(A3) $\forall x_1,\ldots,x_n,y_1,\ldots,y_n \in \mathbb{I} : x_i \le y_i, i=1,\ldots,n \Longrightarrow A(x_1,\ldots,x_n) \le A(y_1,\ldots,y_n)$.

Conditions (A1) and (A2) are called boundary conditions of A, but (A3) means the monotonicity of A. One can consider a case, when instead of $\mathbb{I}$ an arbitrary closed interval $[a,b] \subset [-\infty,+\infty]$ is used.

Next we define a general aggregation operator $\tilde{A}$ acting on L^X, where L^X is the set of all L-fuzzy subsets of a set X [3]. Let $P_1, P_2, \ldots, P_n$ are L-sets, i.e. $P_i \colon X \to L$, $i=1,\ldots,n$. We denote the order on L^X by $\preceq$, but the least and the greatest elements of this order are denoted respectively by $\tilde{0}$ and $\tilde{1}$.

Definition 3.2 *A mapping* $\tilde{A} \colon \bigcup_n (L^X)^n \to L^X$ *is called a general aggregation operator if the following conditions hold:*

($\tilde{A}$*1)* $\tilde{A}(\tilde{0},\ldots,\tilde{0}) = \tilde{0}$*;*

($\tilde{A}$*2)* $\tilde{A}(\tilde{1},\ldots,\tilde{1}) = \tilde{1}$*;*

($\tilde{A}$*3)* $\forall P_1,...,P_n,Q_1,...,Q_n \in L^X : P_i \preceq Q_i, i=1,\ldots,n \Longrightarrow \tilde{A}(P_1,\ldots,P_n) \preceq \tilde{A}(Q_1,\ldots,Q_n)$.

There exist several approaches to construct a general aggregation operator $\tilde{A}$ based on an ordinary aggregation operator A. We use the notion of a T-extension [3] of A, which idea comes from the classical extension principle [5]. To apply this principle we take X equals to an interval on which A is acting.

Definition 3.3 $\tilde{A}$ *is called a* T*-extension of an aggregation operator* A *if*

$$\tilde{A}(P_1,\ldots,P_n)(x) = \bigvee_{x=A(x_1,\ldots,x_n)} T(P_1(x_1),\ldots,P_n(x_n)),$$

where $P_1, P_2, \ldots, P_n \in L^X$, $x, x_1, x_2, \ldots, x_n \in X$.

4 Aggregation of non-negative L-fuzzy real numbers

We introduce an aggregation operator on the extended non-negative L-fuzzy real line $\overline{\mathbb{R}_+}(L)$ by using the T-extension of an ordinary aggregation operator A. We assume that $A \colon \bigcup_n \overline{\mathbb{R}_+}^n \to \overline{\mathbb{R}_+}$ is a continuous aggregation operator and a t-norm T is continuous too. Let us also assume that A takes the value $+\infty$ if at least one of the arguments is infinite.

We define the operator $\tilde{A} \colon \bigcup_n (\overline{\mathbb{R}_+}(L))^n \to \overline{\mathbb{R}_+}(L)$ by the formula

$$\tilde{A}(z_1,\ldots,z_n)(x) = \bigvee_{x=A(x_1,\ldots,x_n)} T(z_1(x_1),\ldots,z_n(x_n)), \text{ if } x \ge 0,$$

$$\text{and } \tilde{A}(z_1,\ldots,z_n)(x) = 1_L, \text{ if } x < 0,$$

where $z_1, z_2, \ldots, z_n \in \overline{\mathbb{R}_+}(L)$, $x, x_1, x_2, \ldots, x_n \in \overline{\mathbb{R}_+}$.

Let us show that properties $(\tilde{A}1) - (\tilde{A}3)$ hold for the operator $\tilde{A}$, then we can be sure that $\tilde{A}$ is a general aggregation operator.

Proof of property $(\tilde{A}1)$*:* To prove the equality $\tilde{A}(\tilde{\theta}, \ldots, \tilde{\theta}) = \tilde{\theta}$ we consider

$$\tilde{A}(\tilde{\theta}, \ldots, \tilde{\theta})(x) = \bigvee_{x=A(x_1,\ldots,x_n)} T(\tilde{\theta}(x_1), \ldots, \tilde{\theta}(x_n)).$$

If $x = 0$, then

$$\tilde{A}(\tilde{\theta}, \ldots, \tilde{\theta})(0) \geq T(\tilde{\theta}(0), \ldots, \tilde{\theta}(0)) = 1_L.$$

If $x > 0$, then $A(x_1, ..., x_n) > 0$, and there exists such $i \in \{1, \ldots, n\}$ that $x_i > 0$. Hence

$$T(\tilde{\theta}(x_1), \ldots, \tilde{\theta}(x_{i-1}), \tilde{\theta}(x_i), \tilde{\theta}(x_{i+1}), \ldots, \tilde{\theta}(x_n)) = T(\tilde{\theta}(x_1), \ldots, \tilde{\theta}(x_{i-1}), 0_L, \tilde{\theta}(x_{i+1}), \ldots, \tilde{\theta}(x_n)) = 0_L.$$

Proof of the property $(\tilde{A}2)$*:*

$$\tilde{A}(\tilde{1}, \ldots, \tilde{1})(x) = \bigvee_{x=A(x_1,\ldots,x_n)} T(\tilde{1}(x_1), \ldots, \tilde{1}(x_n)) = T(1_L, \ldots, 1_L) = \tilde{1}(x).$$

Proof of property $(\tilde{A}3)$*:* By the monotonicity of a t-norm, for all $x_1, \ldots, x_n$, we obtain

$$z_i \leq y_i, i = 1, 2, \ldots, n \Longrightarrow T(z_1(x_1), \ldots, z_n(x_n)) \leq T(y_1(x_1), \ldots, y_n(x_n)) \Longrightarrow$$

$$\Longrightarrow \bigvee_{x=A(x_1,\ldots,x_n)} T(z_1(x_1), \ldots, z_n(x_n)) \leq \bigvee_{x=A(x_1,\ldots,x_n)} T(y_1(x_1), \ldots, y_n(x_n)) \Longrightarrow$$

$$\Longrightarrow \tilde{A}(z_1, \ldots, z_n) \leq \tilde{A}(y_i, \ldots, y_n).$$

Now it is important to show that by using $\tilde{A}$ in the result we get a non-negative L-fuzzy real number. It means that we must check properties $(N1)$ – $(N3)$ for $\tilde{A}(z_1, \ldots, z_n)$.

Proposition 4.1 *For all* $z_1, \ldots, z_n \in \overline{\mathbb{R}_+}(L)$ *the function* $\tilde{A}(z_1, \ldots, z_n)$ *is non-increasing:*

$$x_1 \leq x_2 \Longrightarrow \tilde{A}(z_1, \ldots, z_n)(x_1) \geq \tilde{A}(z_1, \ldots, z_n)(x_2).$$

Proof. We should prove that

$$\bigvee_{x_1=A(\tau_1,\ldots,\tau_n)} T(z_1(\tau_1), \ldots, z_n(\tau_n)) \geq \bigvee_{x_2=A(t_1,\ldots,t_n)} T(z_1(t_1), \ldots, z_n(t_n)).$$

Let us fix arbitrary $t_1, \ldots, t_n$ such that $x_2 = A(t_1, ..., t_n)$. Considering the continuity of A and using the intermediate value theorem, we obtain that

$$\exists(\tau_1, \ldots, \tau_n) \in \prod_{i=1}^{n} [0, t_i] : \ A(\tau_1, ..., \tau_n) = x_1.$$

So if $\tau_i \leq t_i$, then $z_i(\tau_i) \geq z_i(t_i)$, $i = 1, \ldots, n$. Thereby

$$T(z_1(\tau_1), \ldots, z_n(\tau_n)) \geq T(z_1(t_1), \ldots, z_n(t_n)),$$

$$\tilde{A}(z_1,\dots,z_n)(x_1)=\bigvee_{x_1=A(u_1,\dots,u_n)} T(z_1(u_1),\dots,z_n(u_n))\geq T(z_1(t_1),\dots,z_n(t_n)),$$

$$\tilde{A}(z_1,\dots,z_n)(x_1)\geq \bigvee_{x_2=A(t_1,\dots,t_n)} T(z_1(t_1),\dots,z_n(t_n))=\tilde{A}(z_1,\dots,z_n)(x_2).$$

Proposition 4.2 *For all $z_1,\dots,z_n\in\overline{\mathbb{R}_+}(L)$ the function $\tilde{A}(z_1,\dots,z_n)$ is bounded:*

$$\bigwedge_x \tilde{A}(z_1,\dots,z_n)(x)=0_L,\quad \bigvee_x \tilde{A}(z_1,\dots,z_n)(x)=1_L.$$

Proof.

Taking into account that $\tilde{A}(z_1,\dots,z_n)$ is non-increasing, we obtain

$$\bigvee_{x\in[0,+\infty]}\bigvee_{x=A(x_1,\dots,x_n)} T(z_1(x_1),\dots,z_n(x_n))\geq T(z_1(0),\dots,z_n(0))=1_L;$$

$$\bigwedge_{x\in[0,+\infty]}\bigvee_{x=A(x_1,\dots,x_n)} T(z_1(x_1),\dots,z_n(x_n))=\bigvee_{+\infty=A(x_1,\dots,x_n)} T(z_1(x_1),\dots,z_n(x_n)).$$

The equality $A(x_1,...,x_n)=+\infty$ implies that there exists such i that $x_i=+\infty$. Then

$$T(z_1(x_1),\dots,z_n(x_n))=T\left(z_1(x_1),\dots,z_{i-1}(x_{i-1}),z_i(+\infty),z_{i+1}(x_{i+1}),\dots,z_n(x_n)\right)=0_L.$$

Therefore

$$\bigwedge_{x\in[0,+\infty]}\bigvee_{x=A(x_1,\dots,x_n)} T(z_1(x_1),\dots,z_n(x_n))=0_L.$$

Proposition 4.3 *For all $z_1,\dots,z_n\in\overline{\mathbb{R}_+}(L)$ the function $\tilde{A}(z_1,\dots,z_n)$ is left semi-continuous:*

$$\bigwedge_{x<x_0}\tilde{A}(z_1,\dots,z_n)(x)=\tilde{A}(z_1,\dots,z_n)(x_0).$$

Proof. Let us denote $\tilde{A}(z_1,\dots,z_n)(x)=z(x)$. Then we should prove that $\bigwedge\limits_{x<x_0} z(x)=z(x_0)$. Let us assume the opposite, i.e. $\bigwedge\limits_{x<x_0} z(x)=y(x_0)\neq z(x_0)$. By the monotonicity we have $y(x_0)>z(x_0)$.

Basing on the fact that L is completely distributive [1], we will have

$$\bigwedge_{x<x_0}\tilde{A}(z_1,\dots,z_n)(x)=\bigwedge_{x<x_0}\bigvee_{x=A(t_1,\dots,t_n)} T(z_1(t_1),\dots,z_n(t_n))=\bigvee_{f\in F}\bigwedge_{x<x_0} T(z_1(t_1^{f,x}),\dots,z_n(t_n^{f,x})),$$

where F is the set of choice functions f such that

$$\text{for all } x<x_0\ f(x)=(t_1^{f,x},\dots,t_n^{f,x}) \text{ and } A(t_1^{f,x},\dots,t_n^{f,x})=x.$$

By the assumption

$$\bigvee_{f\in F}\bigwedge_{x<x_0} T(z_1(t_1^{f,x}),\ldots,z_n(t_n^{f,x})) > z(x_0)$$

one can find such choice function f that

$$\bigwedge_{x<x_0} T(z_1(t_1^{f,x}),\ldots,z_n(t_n^{f,x})) \not\leq z(x_0).$$

Let us denote

$$\alpha(x_0) = \bigwedge_{x<x_0} T(z_1(t_1^{f,x}),\ldots,z_n(t_n^{f,x})).$$

The value $\alpha(x_0)$ is either greater than $z(x_0)$ or is incomparable with $z(x_0)$.

Now let us take $x = x_0 - \frac{1}{m}$ and denote

$$t_i^{f,x_0-\frac{1}{m}} = t_i^m,\ i = 1,\ldots,n, \quad t^m = (t_1^m,\ldots,t_n^m),\ m\in\mathbb{N}.$$

The sequence $(t^m)_{m\in\mathbb{N}}$ is such that

$$T(z_1(t_1^m),\ldots,z_n(t_n^m)) \geq \alpha(x_0).$$

We select a subsequence $(t^{m_k})_{k\in\mathbb{N}}$, which has the limit, and denote this limit by t^0: $\lim\limits_{k\to\infty} t^{m_k} = t^0$. By the continuity of A:

$$A(t_1^0,\ldots,t_n^0) = A(\lim_{k\to\infty} t_1^{m_k},\ldots,\lim_{k\to\infty} t_n^{m_k}) = \lim_{k\to\infty} A(t_1^{m_k},\ldots,t_n^{m_k}) = x_0.$$

We consider the set $D = \{(t_1,\ldots,t_n)\in[0,\infty] \mid T(z_1(t_1),\ldots,z_n(t_n)) \geq \alpha(x_0)\}$ and note the following properties:

- if $t\in D$ and $\tau\leq t$, i.e. for all $i\in\{1,\ldots,n\}$ $\tau_i\leq t_i$, then $\tau\in D$;
- if a point $\tau=(\tau_1,\ldots,\tau_n)$ is such that $A(\tau_1,\ldots,\tau_n)=x_0$, then $\tau\notin D$;
- if a point $\tau=(\tau_1,\ldots,\tau_n)$ is such that $A(\tau_1,\ldots,\tau_n)>x_0$, then $\tau\notin D$.

Let us consider a line K from the product $\prod\limits_{i=1}^{n}[0,t_i^0]$, which connects point zero and point t^0. Such a line can be described by the equations $t_i=\tau_i(u)$ for all $i\in\{1,\ldots,n\}$:

- if $t_i^0<+\infty$, then $\tau_i(u)=ut_i^0$, $u\in[0,1]$;
- if for some $t_i^0=+\infty$, then $\tau_i(u)=\frac{u}{1-u}$, for all $u\in[0,1[$ and $\tau_i(1)=+\infty$.

All points $\tau=\tau(u)$ of the line K, when $u\in[0,1[$, belong to the set D: if we fix some point $\tau^0\in K$, $\tau^0\neq t^0$, then in every neighborhood of the point t^0 one can find such a point $t^{m_{k_0}}$ from the sequence $(t^{m_k})_{k\in\mathbb{N}}$ that for all $i\in\{1,\ldots,n\}$ $\tau_i^0 < t_i^{m_{k_0}}$. Now taking into account that $t^{m_{k_0}}\in D$, we get $\tau^0\in D$.

For every $m \in \mathbb{N}$ on the line K one can find a point τ^m such that $A(\tau_1^m, \ldots, \tau_n^m) = x_0 - \frac{1}{m}$. This point can be found by using the continuity of A and the intermediate value theorem. Then

$$\lim_{m\to\infty} A(\tau_1^m, \ldots, \tau_n^m) = \lim_{m\to\infty} \left(x_0 - \frac{1}{m} \right) = x_0.$$

For the subsequence $(\tau^{m_k})_{k\in\mathbb{N}}$ we have

$$\lim_{k\to\infty} \tau^{m_k} = t^0, \text{ since } \lim_{k\to\infty} \tau^{m_k} \in K \text{ and } \lim_{k\to\infty} A(\tau_1^{m_k}, \ldots, \tau_n^{m_k}) = x_0.$$

Now let us take the limit in $T(z_1(\tau_1^{m_k}), \ldots, z_n(\tau_n^{m_k}))$:

$$\bigwedge_k T(z_1(\tau_1^{m_k}), \ldots, z_n(\tau_n^{m_k})) = T\left(\bigwedge_k z_1(\tau_1^{m_k}), \ldots, \bigwedge_k z_n(\tau_n^{m_k}) \right) = T(z_1(t_1^0), \ldots, z_n(t_n^0)) = z(x_0).$$

Here we have got a contradiction, since $\bigwedge_k T(z_1(\tau_1^{m_k}), \ldots, z_n(\tau_n^{m_k})) \geq \alpha(x_0)$. Thereby our assumption was wrong and $\bigwedge_{x<x_0} z(x) = z(x_0)$ or

$$\bigwedge_{x<x_0} \tilde{A}(z_1, \ldots, z_n)(x) = \tilde{A}(z_1, \ldots, z_n)(x_0).$$

5 Properties of aggregation operator $\tilde{A}$

Now we consider some properties of aggregation operator $\tilde{A}$ acting on non-negative L-fuzzy real numbers. As $\tilde{A}$ is based on an ordinary aggregation operator A and a t-norm T, it is natural to investigate the properties of $\tilde{A}$ depending on the properties of A and T.

Proposition 5.1 *If operator A is **associative**, then operator $\tilde{A}$ is **associative**:*

$$\forall z_1, z_2, z_3 \in \overline{\mathbb{R}_+}(L) \quad \tilde{A}(z_1, \tilde{A}(z_2, z_3)) = \tilde{A}(\tilde{A}(z_1, z_2), z_3).$$

Proof.

$$\tilde{A}(z_1, \tilde{A}(z_2, z_3))(x) = \bigvee_{x=A(x_1,x_2)} T\left(z_1(x_1), \tilde{A}(z_2, z_3)(x_2) \right) =$$

$$= \bigvee_{x=A(x_1,x_2)} T\left(z_1(x_1), \bigvee_{x_2=A(x_3,x_4)} T(z_2(x_3), z_3(x_4)) \right) =$$

$$= \bigvee_{x=A(x_2,x_4)} T\left(\bigvee_{x_2=A(x_1,x_3)} T(z_1(x_1), z_2(x_3)), z_3(x_4) \right) =$$

$$= \bigvee_{x=A(x_2,x_4)} T\left(\tilde{A}(z_1, z_2)(x_2), z_3(x_4) \right) = \tilde{A}(\tilde{A}(z_1, z_2), z_3)(x).$$

Proposition 5.2 *If operator A is **commutative**, then operator $\tilde{A}$ is **commutative**:*

$$\forall z_1, z_2 \in \overline{\mathbb{R}_+}(L) \quad \tilde{A}(z_1, z_2) = \tilde{A}(z_2, z_1).$$

Proof. $\tilde{A}(z_1, z_2)(x) = \bigvee_{x=A(x_1,x_2)} T\left(z_1(x_1), z_2(x_2)\right) = \bigvee_{x=A(x_2,x_1)} T\left(z_2(x_2), z_1(x_1)\right) = \tilde{A}(z_2, z_1)(x).$

Proposition 5.3 *If operator A is **idempotent** and T is a minimum t-norm T_M, then operator $\tilde{A}$ is **idempotent**:*

$$\forall z \in \overline{\mathbb{R}_+}(L) \quad \tilde{A}(z, \ldots, z) = z.$$

Proof. First of all let us show that $\tilde{A}(z, \ldots, z)(x) \geq z(x)$:

$$\tilde{A}(z, \ldots, z)(x) = \bigvee_{x=A(x_1,\ldots,x_n)} T_M(z(x_1), \ldots, z(x_n)) \geq \bigvee_{x=A(x,\ldots,x)} T_M(z(x), \ldots, z(x)) = z(x).$$

Now we should prove that $\tilde{A}(z, \ldots, z)(x) \leq z(x)$. For an ordinary aggregation operator A the idempotence is equivalent to the compensation property [6]:

$$\min(x_1, ..., x_n) \leq A(x_1, ..., x_n) \leq \max(x_1, ..., x_n).$$

Using this property we obtain

$$x = A(x_1, ..., x_n) \leq \max(x_1, ..., x_n),$$

$$z(x) \geq T_M(z(x_1), ..., z(x_n)).$$

Therefore

$$z(x) \geq \bigvee_{x=A(x_1,\ldots,x_n)} T_M(z(x_1), \ldots, z(x_n)) = \tilde{A}(z, \ldots, z)(x).$$

Proposition 5.4 ***The neutral element*** *of operator $\tilde{A}$ in the case, when 0 is the neutral element of operator A, is element θ:*

$$\tilde{A}(z_1, \ldots, z_{i-1}, \tilde{\theta}, z_{i+1}, \ldots, z_n) = \tilde{A}(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n).$$

Proof. If $x_i > 0$ and $x = A(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$, then

$$T(z_1(x_1), \ldots, z_{i-1}(x_{i-1}), \tilde{\theta}(x_i), z_{i+1}(x_{i+1}), \ldots, z_n(x_n)) =$$

$$= T(z_1(x_1), \ldots, z_{i-1}(x_{i-1}), 0_L, z_{i+1}(x_{i+1}), \ldots, z_n(x_n)) = 0_L.$$

Thereby

$$\tilde{A}(z_1, \ldots, z_{i-1}, \tilde{\theta}, z_{i+1}, \ldots, z_n)(x) =$$

$$= \bigvee_{x=A(x_1,\ldots,x_n)} T(z_1(x_1), \ldots, z_{i-1}(x_{i-1}), \tilde{\theta}(x_i), z_{i+1}(x_{i+1}), \ldots, z_n(x_n)) =$$

$$= \bigvee_{x=A(x_1,\ldots,x_{i-1},0,x_{i+1},\ldots,x_n)} T(z_1(x_1),\ldots,z_{i-1}(x_{i-1}),\tilde{\theta}(0),z_{i+1}(x_{i+1}),\ldots,z_n(x_n)) =$$

$$= \bigvee_{x=A(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)} T(z_1(x_1),\ldots,z_{i-1}(x_{i-1}),z_{i+1}(x_{i+1}),\ldots,z_n(x_n)) =$$

$$= \tilde{A}(z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_n)(x).$$

Proposition 5.5 *If operator A is **homogeneous** by means of multiplication with a positive real number, then $\tilde{A}$ is **homogeneous** by means of multiplication with a positive real number:*

$$\tilde{A}(kz_1,\ldots,kz_n) = k\tilde{A}(z_1,\ldots,z_n), \quad k>0.$$

Proof. $\tilde{A}(kz_1,\ldots,kz_n)(x) = \bigvee_{x=A(x_1,\ldots,x_n)} T\left(kz_1(x_1),\ldots,kz_n(x_n)\right) =$

$$= \bigvee_{x=A(x_1,\ldots,x_n)} T\left(z_1\left(\frac{x_1}{k}\right),\ldots,z_n\left(\frac{x_n}{k}\right)\right) = \bigvee_{\frac{x}{k}=\frac{1}{k}A(x_1,\ldots,x_n)} T\left(z_1\left(\frac{x_1}{k}\right),\ldots,z_n\left(\frac{x_n}{k}\right)\right) =$$

$$= \bigvee_{\frac{x}{k}=A\left(\frac{x_1}{k},\ldots,\frac{x_n}{k}\right)} T\left(z_1\left(\frac{x_1}{k}\right),\ldots,z_n\left(\frac{x_n}{k}\right)\right) = \tilde{A}(z_1,\ldots,z_n)\left(\frac{x}{k}\right) = k\tilde{A}(z_1,\ldots,z_n)(x).$$

6 Operations with L-fuzzy real numbers

In this section we consider such t-norm based operations with non-negative L-fuzzy real numbers as addition, minimum and maximum. We can rewrite the formula of addition by using the general aggregation operator of arithmetic mean $\tilde{A}_M$, which is based on the ordinary aggregation operator of arithmetic mean A_M:

$$z_1 \oplus_T \ldots \oplus_T z_n = n\tilde{A}_M(z_1,\ldots,z_n).$$

This formula is an equivalent of the classical formula for addition of L-fuzzy real numbers:

$$n\tilde{A}_M(z_1,\ldots,z_n)(x) = \tilde{A}_M(z_1,\ldots,z_n)\left(\frac{x}{n}\right) = \bigvee_{\frac{x}{n}=A_M(x_1,\ldots,x_n)} T(z_1(x_1),\ldots,z_n(x_n)) =$$

$$= \bigvee_{\frac{x}{n}=\frac{1}{n}(x_1+\ldots+x_n)} T(z_1(x_1),\ldots,z_n(x_n)) = \bigvee_{x=x_1+\ldots+x_n} T(z_1(x_1),\ldots,z_n(x_n)).$$

Some properties of the operator $\oplus_T$ can be obtained from the corresponding properties of $\tilde{A}_M$. As the ordinary A_M operator is associative, commutative, and homogeneous, the operator $\oplus_T$ is associative, commutative and homogeneous as well. The element $\tilde{\theta}$ is the neutral element of $\oplus_T$, i.e. $z \oplus_T \tilde{\theta} = z$.

The property of distributivity $(\alpha+\beta)z = \alpha z \oplus_T \beta z$, $\alpha, \beta > 0$ does not hold for an arbitrary t-norm. For example, in the case of product t-norm T_P: $z \oplus_{T_P} z \neq 2z$ for some $z \in \overline{\mathbb{R}_+}(L)$. The distributivity holds, when the extension of operator preserves the idempotence. Really, taking into account the equality

$$\tilde{A}_M((\alpha+\beta)z, (\alpha+\beta)z) = \tilde{A}_M(2\alpha z, 2\beta z),$$

which holds in the case of minimum t-norm T_M, the distributivity can be reduced to

$$\tilde{A}_M((\alpha+\beta)z, (\alpha+\beta)z) = (\alpha+\beta)z$$

(this last equality means the idempotence property). To prove the equality we consider

$$\tilde{A}_M((\alpha+\beta)z, (\alpha+\beta)z)(x) = \bigvee_{x_1+x_2=2x} T_M\left(z\left(\frac{x_1}{\alpha+\beta}\right), z\left(\frac{x_2}{\alpha+\beta}\right)\right) = z\left(\frac{x}{\alpha+\beta}\right),$$

$$\tilde{A}_M(2\alpha z, 2\beta z)(x) = \bigvee_{x_1+x_2=2x} T_M\left(z\left(\frac{x_1}{2\alpha}\right), z\left(\frac{x_2}{2\beta}\right)\right) = z(x_0),$$

where $x_0 = \frac{x}{\alpha+\beta}$ is obtained as the solution of the following system of linear equations:

$$\begin{cases} \frac{x_1}{2\alpha} = \frac{x_2}{2\beta} = x_0, \\ x_1 + x_2 = 2x. \end{cases}$$

We define the operations of minimum and maximum of L-fuzzy real numbers by the following formulas:

$$MIN(z_1, \ldots, z_n)(x) = \bigvee_{x=\min(x_1,\ldots,x_n)} T(z_1(x_1), \ldots, z_n(x_n)),$$

$$MAX(z_1, \ldots, z_n)(x) = \bigvee_{x=\max(x_1,\ldots,x_n)} T(z_1(x_1), \ldots, z_n(x_n)).$$

As the ordinary operators of minimum and maximum are associative, commutative, homogeneous and idempotent, the extended operations with L-fuzzy real numbers MIN and MAX will be associative, commutative, homogeneous and idempotent (in the case of minimum t-norm) as well. But it is worth to mention that the result of these operations depends on the choise of a t-norm. For example the result of MIN operation in the case of minimum t-norm T_M will be just an ordinary minimum of functions, but the result in the case of product t-norm T_P can be different from the ordinary one.

Acknowledgement

The paper was supported by ESF project Nr.2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004.

References

[1] B. A. DAVEY, H. A. PRIESTLEY *Introduction to Lattices and Order*, Cambridge University Press, Second Edition, 2002.

[2] B. HUTTON *Normality in Fuzzy Topological Spaces*, J.Math.Anal.Appl., **50**, 1975, 74 – 79.

[3] A. TAKACI *General Aggregation Operators Acting on Fuzzy Numbers Induced by Ordinary Aggregation Operators*, Novi Sad J. Math., **33**(2), 2003, 67 – 76.

[4] M. DETYNIECKI *Fundamentals on Aggregation Operators*, Berkeley, 2001.

[5] E. P. KLEMENT, R. MESIAR, E. PAP *Triangular Norms*, Kluwer Academic Publishers, Dordrecht, 2000.

[6] T. CALVO, G. MAYOR, R. MESIAR *Aggregation Operators*, Physical-Verlag, Heidelberg, 2002.

[7] R. LOWEN *On* $(\mathbb{R}(L), \oplus)$, Fuzzy Sets and Systems, **10**, 1983, 203 – 209.

[8] S. E. RODABAUGH *Fuzzy Addition and the Fuzzy Real Lines*, Fuzzy Sets and Systems, **8**, 1982, 39 – 52.

[9] S. E. RODABAUGH *Complete Fuzzy Topological Hyperfields and Fuzzy Multiplication in the Fuzzy Real Lines*, Fuzzy Sets and Systems, **15**, 1988, 285 – 310.

Current address

Pavels Orlovs, M.Sc.
University of Latvia, Department of Mathematics
Zellu street 8, Riga, LV-1002, Latvia
Tel.: +371 29681122
e-mail: pavels.orlovs@gmail.com

Journal of Applied Mathematics

volume IV (2011), number III

ON AN L-FUZZY VALUED INTEGRAL WITH RESPECT TO AN L-FUZZY VALUED T_M-MEASURE

RUZA Vecislavs, (LV)

Abstract. We continue to develop a theory of an L-fuzzy valued T-measure and consider an L-fuzzy valued integral of a real valued non-negative measurable function over L-set with respect to an L-fuzzy valued T-measure. The main purpose of the present paper is to introduce the concept of an L-fuzzy valued integral and to investigate it basic properties.

Key words and phrases. L-sets, L-fuzzy real numbers, L-fuzzy valued measure, L-fuzzy valued integral.

Mathematics Subject Classification. 94D05, 03E72.

1 Introduction

One can find a lot of works regarding fuzzy approach to measure and integral. The biggest scope on this subject can be found in [1],[2],[3]. Our interest is in developing a theory where not only sets are fuzzy, but also measure and integral take fuzzy real values. To realize this we need a concept of a fuzzy real number. From majority of different definitions we give our preference to the fuzzy real numbers as they were first defined by B. Hutton [4], and then studied thoroughly in a series of papers [5],[6],[7].

In our previous works [8],[9] we suggested the construction of an L-fuzzy valued T-measure of L-sets by extension a measure defined on a σ-algebra of crisp sets to a T-measure on a T-tribe where T is the minimum t-norm $T_M(x, y) = x \wedge y$ and L is a completely distributive lattice.

The main purpose of the present paper is to define an L-fuzzy valued integral based on described above L-fuzzy fuzzy valued measure and to show some basic properties of it. Initial results on this topic were presented at the conference FSTA in Slovakia in 2010.

2 L-sets and L-fuzzy real numbers

Given a (crisp) universe X and a completely distributive lattice $L(\wedge, \vee, 0_L, 1_L)$, an *L-subset* A of X (or, briefly, an L-set A) is a function $A : X \longrightarrow L$. The class of all L-subsets of X is denoted L^X. The operations for L-sets A, B and for a sequence of L-sets $(A_n)_{n\in\mathbb{N}}$ are defined by using the minimum triangular norm T_M, its corresponding co-norm S_M and involution N:

$$(A \wedge B)(x) = T_M(A(x), B(x)),\ (A \vee B)(x) = S_M(A(x), B(x)),\ A^c(x) = N(A(x)),$$
$$\bigwedge_{n=1}^{\infty} A_n = \bigwedge_{n\in\mathbb{N}} (A_1 \wedge A_2 \wedge ... \wedge A_n) \text{ and } \bigvee_{n=1}^{\infty} A_n = \bigvee_{n\in\mathbb{N}} (A_1 \vee A_2 \vee ... \vee A_n).$$

A finite family of L-sets $A_1, A_2, \ldots, A_n$ is said to be T_M-*disjoint* (see e.g.[1]) iff for each $k \in \{1, \ldots, n\}$ we have $(\bigvee_{j=1, j\neq k}^{n} A_j) \wedge A_k = \emptyset$. A countable family of L-sets is said to be T_M-disjoint iff every finite subfamily of this family is T_M-disjoint.

In order to define an L-fuzzy valued T_M-measure we consider a class of L-sets called T_M-tribe (see e.g. [1]) and L-fuzzy real numbers defined by B. Hutton [4].

Definition 2.1 *A subclass $\Sigma \subset L^X$ is called a T_M-tribe on X if the following properties are satisfied:*

- $\emptyset \in \Sigma$;
- *for all $A \in \Sigma$ we have $A^c \in \Sigma$;*
- *for all sequences $(A_n)_{n\in N} \subset \Sigma$ we have $\bigwedge_{n=1}^{\infty} A_n \in \Sigma$.*

Definition 2.2 *An L-fuzzy real number is a function $z : \mathbb{R} \to L$ such that*

- *z is non-increasing;*
- $\bigwedge_t z(t) = 0_L$, $\bigvee_t z(t) = 1_L$;
- *z is left semi-continuous, i.e. $\bigwedge_{t<t_0} z(t) = z(t_0)$.*

The set of all L-fuzzy real numbers is called *the L-fuzzy real line* and it is denoted by $\mathbb{R}(L)$. A fuzzy number z is called *non-negative* if $z(0) = 1_L$. The set of all non-negative L-fuzzy real numbers we denote by $\mathbb{R}_+(L)$.

Operations with L-fuzzy real numbers such as addition $\oplus$ and multiplication by a real positive number are defined as following:

$$(z_1 \oplus z_2)(t) = \bigvee_{\tau} \{z_1(\tau) \wedge z_2(t-\tau)\}, (\bigoplus_{n\in\mathbb{N}} z_n)(t) = \bigvee_{n\in\mathbb{N}} (z_1 \oplus z_2 \oplus ... \oplus z_n)(t) \text{ and } (zr)(t) = z(\frac{t}{r}).$$

The supremum of $F \subset \mathbb{R}_+(L)$ defined by formula (see e.g. [10]):

$$(Sup\ F)(t) = \wedge\{z(t) \mid z \text{ is an L-fuzzy number and } z \geq z' \text{ for all } z' \in F\}.$$

If F is bounded from above (i.e. there exists $z_0 \in \mathbb{R}(L)$ such that $z \leq z_0$ for all $z \in F$), then $SupF$ is an L-fuzzy real number, otherwise the condition $\bigwedge\limits_t SupF(t) = 0_L$ not necessarily holds.

For $a \in \mathbb{R}_+$ and $\alpha \in L$ by $z(a, \alpha)$ we denote a special type of non-negative L-fuzzy real numbers

$$(z(a,\alpha))(t) = \begin{cases} 1, & t \leq 0, \\ \alpha, & 0 < t \leq a, \\ 0, & t > a, \end{cases}$$

that will play an important role in our work. Note that:

$$a_1, a_2 \in \mathbb{R}_+ \Rightarrow z(a_1,\alpha) \oplus z(a_2,\alpha) = z(a_1 + a_2, \alpha); c \in \mathbb{R}_+ \Rightarrow cz(a,\alpha) = z(ca,\alpha);$$

$$a_i \in \mathbb{R}_+, i \in J \Rightarrow Sup\{z(a_i,\alpha) \mid i \in J\} = z(\sup\{a_i \mid i \in J\}, \alpha).$$

3 L-fuzzy valued T_M-measure

In this section we consider a measure that defined on a T_M-tribe and takes values in $\mathbb{R}_+(L)$.

Definition 3.1 *Let Σ be a T_M-tribe. A function $\mu : \Sigma \to \mathbb{R}_+(L)$ is called an L-fuzzy valued T_M-measure if it satisfies the following conditions:*

- $\mu(\emptyset) = z(0, 1_L)$;
- *μ is T_M-valuation, i.e. for all $A, B \in \Sigma$ it holds $\mu(A \wedge B) \oplus \mu(A \vee B) = \mu(A) \oplus \mu(B)$;*
- *μ is left T_M-continuous, i.e. $\bigvee\limits_{n\in\mathbb{N}} \mu(A_n) = \mu(A)$, where $(A_n)_{n\in N} \subset \Sigma$, $\bigvee\limits_{n\in\mathbb{N}} A_n = A \in \Sigma$.*

For a given σ-algebra $\Phi \subset 2^X$ and a finite measure $\nu : \Phi \to \mathbb{R}_+$ an L-fuzzy valued T_M-measure can be obtained by the following schema (see e.g. [8], [9]):

- For $M \in \Phi$, $\alpha \in L$ we define an L-fuzzy set

  $$(A(M,\alpha))(x) = \begin{cases} \alpha, x \in M, \\ 0, x \notin M. \end{cases}$$

  All these L-sets form a class of L-sets that we denote by $\wp$.

- Next we define an L-fuzzy valued function $m : \wp \to \mathbb{R}_+(L)$ by

  $$m(A(M,\alpha)) = z(\nu(M),\alpha),$$

  and we extend it to the L-fuzzy valued function $m^* : L^X \to \mathbb{R}_+(L)$ as following:

  $$m^*(E) = \bigwedge\{\bigoplus_{n=1}^{\infty} m(E_n) \mid (E_n)_{n\in\mathbb{N}} \subset \wp : E \leq \bigvee_{n=1}^{\infty} E_n\}.$$

- We denote by Σ the T_M-tribe of all so called m^*-measurable L-sets $B \in L^X$ such that for all $E \in L^X$ it holds

$$m^*(B) \oplus m^*(E) = m^*(B \wedge E) \oplus m^*(B \vee E),$$
$$m^*(B^c) \oplus m^*(E) = m^*(B^c \wedge E) \oplus m^*(B^c \vee E).$$

- We consider μ as the restriction of m^* to Σ. Then μ is an L-fuzzy valued T_M-measure such that $\mu/_{\wp} = m$.

4 L-fuzzy valued integral

Our aim is to define an L-fuzzy valued integral $\int\limits_E f \; d\mu$, where $E \in \Sigma$ and $f : X \to \mathbb{R}$ is a non-negative measurable function with respect to σ-algebra Φ.

By analogy with the classical case ([11]) we define an L-fuzzy valued integral stepwise, first considering the case of simple non-negative measurable functions (for short SNMF):

$$\int\limits_E (\sum_{i=1}^{n} c_i \chi_{C_i}) \; d\mu = \bigoplus_{i=1}^{n} (c_i \; \mu(C_i \wedge E)),$$

whenever $c_i \in \mathbb{R}_+, C_i \in \Phi, \chi_{C_i}$ is the characteristic function, $i \in \{1, ..., n\}$, and $C_1, ..., C_n$ are pairwise disjoint sets.

Then considering the case for non-negative measurable function f (for short NMF):

$$\int\limits_E f \; d\mu = Sup\{ \int\limits_E g \; d\mu \mid g \leq f \text{ and } g \text{ is SNMF}\}.$$

For $\mathbb{I}_f = \int\limits_E f \; d\mu$ due to properties of the supremum of a set of L-fuzzy numbers, we have

- $\mathbb{I}_f$ is non-increasing,
- $\bigvee\limits_t \mathbb{I}_f(t) = 1_L$,
- $\mathbb{I}_f$ is left semi-continuous, i.e. $\bigwedge\limits_{t<t_0} \mathbb{I}_f(t) = \mathbb{I}_f(t_0)$.

Definition 4.1 *We say that a non-negative measurable function f is L-fuzzy integrable iff*

$$\bigwedge_t \mathbb{I}_f(t) = 0_L.$$

5 Properties of an L-fuzzy valued integral

In this section we consider properties of an L-fuzzy valued integral of L-fuzzy integrable functions.

(𝕀1) $r \in \mathbb{R}_+ \Rightarrow \int\limits_E rf d\mu = r \int\limits_E f d\mu$

Proof. In the case when integrand $f = \sum\limits_{i=1}^{n} c_i \chi_{C_i}$ is SNMF the equality follows from

$$\bigoplus_{i=1}^{n}((rc_i)\ \mu(C_i \wedge E)) = r \bigoplus_{i=1}^{n}(c_i\ \mu(C_i \wedge E)).$$

Considering the case when f is NMF and $r > 0$ (if $r = 0$ then the equality is obvious) we have

$$\int\limits_E rf d\mu = Sup\{\int\limits_E g\ d\mu \mid g \leq rf, \text{ and } g \text{ is SNMF}\} =$$

$$= rSup\{\int\limits_E \frac{g}{r}\ d\tilde{\mu} \mid \frac{g}{r} \leq f, \text{ and } g \text{ is SNMF}\} = r \int\limits_E f d\mu.$$

(𝕀2) $f_1 \leq f_2 \Rightarrow \int\limits_E f_1 d\mu \leq \int\limits_E f_2 d\mu$

Proof. From

$$\{\int\limits_E g\ d\mu \mid g \leq f_1 \text{ and } g \text{ is SNMF}\} \subset \{\int\limits_E g\ d\mu \mid g \leq f_2 \text{ and } g \text{ is SNMF}\}$$

it follows

$$Sup\{\int\limits_E g\ d\mu \mid g \leq f_1 \text{ and } g \text{ is SNMF}\} \leq Sup\{\int\limits_E g\ d\mu \mid g \leq f_2 \text{ and } g \text{ is SNMF}\}.$$

(𝕀3) $E_1 \subset E_2 \Rightarrow \int\limits_{E_1} f\ d\mu \leq \int\limits_{E_2} f\ d\mu$

Proof. The inequality

$$Sup\{\int\limits_{E_1} g\ d\nu \mid g \leq f \text{ and } g \text{ is SNMF}\} \leq Sup\{\int\limits_{E_2} g\ d\nu \mid g \leq f \text{ and } g \text{ is SNMF}\}$$

holds due to $\int\limits_{E_1} g d\mu \leq \int\limits_{E_2} g d\mu$, where $g = \sum\limits_{i=1}^{n} c_i \chi_{C_i}$ is SNMF. The last inequality is equivalent to

$$\bigoplus_{i=1}^{n} c_i\ \mu(C_i \wedge E_1) \leq \bigoplus_{i=1}^{n} c_i\ \mu(C_i \wedge E_2),$$

that holds because of monotonicity of μ.

(𝕀4) $(E_k)_{k\in\mathbb{N}} : E_k \leq E_{k+1}$ and $\bigvee\limits_{k\in\mathbb{N}} E_k = E \Rightarrow \int\limits_E f d\mu = Sup\{\int\limits_{E_k} f d\mu\}$

Proof. We start with the case when f is SNMF:

$$\int\limits_E f\ d\mu = \bigoplus_{i=1}^{k}(c_i\ \mu(C_i \wedge E)) = Sup\{\bigoplus_{i=1}^{k}(c_i\ \mu(C_i \wedge E_k)) \mid k \in \mathbb{N}\} = Sup\{\int\limits_{E_k} f d\mu \mid k \in \mathbb{N}\}.$$

Now to prove the equality when f is NMF we show both inequalities "$\geq$" and "$\leq$".
The inequality $\int\limits_E f\ d\mu \geq \int\limits_{E_k} f\ d\mu$ for all $k \in \mathbb{N}$ implies the inequality

$$\int\limits_E f\ d\mu \geq Sup\{\int\limits_{E_k} f d\mu \mid k \in \mathbb{N}\}.$$

Taking into account that for all functions g (g is SNMF and $g \leq f$) we have

$$\int\limits_E g\ d\mu = Sup\{\int\limits_{E_k} g\ d\mu \mid k \in \mathbb{N}\} \leq Sup\{\int\limits_{E_k} f\ d\mu \mid k \in \mathbb{N}\},$$

it follows that

$$\int\limits_E f\ d\mu = Sup\{\int\limits_E g\ d\mu \mid g \text{ is SNMF and } g \leq f\} \leq Sup\{\int\limits_{E_k} f\ d\mu \mid k \in \mathbb{N}\}.$$

(𝕀5) $(f_n)_{n\in\mathbb{N}} : f_n \leq f_{n+1}$ and $\lim\limits_{n\to\infty} f_n = f \Rightarrow Sup\{\int\limits_E f_n\ d\mu \mid n \in \mathbb{N}\} = \int\limits_E f\ d\mu$

Proof. To prove the equality we show that both inequalities "$\leq$" and "$\geq$".
The inequality $\bigvee\limits_{n\in\mathbb{N}} (\int\limits_E f_n\ d\mu) \leq \int\limits_E f\ d\mu$ holds due to property (𝕀2). To show the opposite inequality by using a function g (g is SNMF and $g \leq f$) and a number $c \in (0,1)$ we define

$$M_n = \{x \in X \mid f_n(x) \geq cg(x),\ c \in (0,1)\} \text{ and } E_n = E \wedge M_n,\ n \in \mathbb{N}.$$

Obviously, $M_n \leq M_{n+1}$ and $\bigcup\limits_{n\in\mathbb{N}} M_n = Dom(f)$. Now we have

$$\int\limits_E f_n d\mu \geq \int\limits_{E\wedge M_n} f_n d\mu \geq \int\limits_{E\wedge M_n} cg\ d\mu,$$

$$Sup\{\int\limits_E f_n\ d\mu \mid n \in \mathbb{N}\} \geq c\ Sup\{\int\limits_{E\wedge M_n} g\ d\mu \mid n \in \mathbb{N}\} = c \int\limits_{\bigvee\limits_{n\in\mathbb{N}} E_n} g\ d\mu.$$

It follows

$$Sup\{\int\limits_E f_n \ d\mu \mid n \in \mathbb{N}\} \geq c \int\limits_E f \ d\mu.$$

And finally,

$$Sup\{\int\limits_E f_n \ d\mu \mid n \in \mathbb{N}\} \geq \int\limits_E f \ d\mu.$$

(𝕀6) $\int\limits_E (f_1 + f_2)d\mu = \int\limits_E f_1 d\mu \oplus \int\limits_E f_2 d\mu$

Proof. Again we start with the case when integrands are SNMF:

$$f_1(x) = \sum_{i=1}^{n} c_i \ \chi_{C_i}, \ f_2(x) = \sum_{j=1}^{k} b_j \ \chi_{B_j}, \ (f_1 + f_2)(x) = \sum_{l=1}^{p} a_l \ \chi_{A_l}.$$

We suppose that $\bigcup\limits_{i=1}^{n} C_i = \bigcup\limits_{j=1}^{k} B_j = \bigcup\limits_{l=1}^{p} A_l = X$. Then

$$\int\limits_E (f_1 + f_2)d\mu = \bigoplus_{l=1}^{m}(a_l \ \mu(A_l \wedge E)) = \bigoplus_{i=1}^{n}\bigoplus_{j=1}^{k}\bigoplus_{l=1}^{m}(a_l \ \mu(A_l \wedge C_i \wedge B_j \wedge E)) =$$

$$= \bigoplus_{i=1}^{n}\bigoplus_{j=1}^{k}\bigoplus_{l=1}^{m}((c_i + b_j) \ \mu(A_l \wedge C_i \wedge B_j \wedge E)) =$$

$$= \bigoplus_{i=1}^{n} c_i \bigoplus_{j=1}^{k}\bigoplus_{l=1}^{m}\mu(A_l \wedge C_i \wedge B_j \wedge E)) \oplus \bigoplus_{j=1}^{k} b_j \bigoplus_{i=1}^{n}\bigoplus_{l=1}^{m}\mu(A_l \wedge C_i \wedge B_j \wedge E)) =$$

$$= \bigoplus_{l=1}^{n} c_i \ \mu(C_i \wedge E) \oplus \bigoplus_{j=1}^{k} b_j \ \mu(B_j \wedge E) = \int\limits_E f_1 d\mu \oplus \int\limits_E f_2 d\mu.$$

Now we consider the case when integrands are NMF. Then we can find non-increasing sequences of SNMF $(g_n)_{n\in\mathbb{N}}$ and $(h_n)_{n\in\mathbb{N}}$: $\lim\limits_{n\to\infty} g_n = f_1$ and $\lim\limits_{n\to\infty} h_n = f_2$.
Hence,

$$\int\limits_E f_1 \ d\mu \oplus \int\limits_E f_2 \ d\mu = Sup\{\int\limits_E g_n \ d\mu \mid n \in \mathbb{N}\} \oplus Sup\{\int\limits_E h_n \ d\mu \mid n \in \mathbb{N}\} =$$

$$= Sup\{\int\limits_E (g_n + h_n) \ d\mu \mid n \in \mathbb{N}\} = \int\limits_E (f_1 + f_2) \ d\mu.$$

(𝕀7) $E_1 \wedge E_2 = \emptyset \Rightarrow \int\limits_{E_1 \vee E_2} f d\mu = \int\limits_{E_1} f d\mu \oplus \int\limits_{E_2} f d\mu$

Proof. To prove the equality we define

$$f_1(x) = \begin{cases} f(x), \text{ when } x \in Supp(E_1), \\ 0, \text{ otherwise.} \end{cases} \quad f_2(x) = \begin{cases} f(x), \text{ when } x \in Supp(E_2), \\ 0, \text{ otherwise.} \end{cases}$$

Now we obtain

$$\int\limits_{E_1 \vee E_2} f d\mu = \int\limits_{E_1 \vee E_2} (f_1 + f_2) d\mu = \int\limits_{E_1 \vee E_2} f_1 d\mu \oplus \int\limits_{E_1 \vee E_2} f_2 d\mu =$$

$$= \int\limits_{E_1} f_1 d\mu \oplus \int\limits_{E_2} f_2 d\mu = \int\limits_{E_1} f d\mu \oplus \int\limits_{E_2} f d\mu.$$

(𝕀8) $A(M, \alpha) \in \wp \Rightarrow \int\limits_{A(M,\alpha)} f d\mu = z(\int\limits_M f \ d\nu, \alpha)$

Proof. For $f = \sum\limits_{i=1}^{n} c_i \chi_{C_i}$ we get

$$\int\limits_E \sum_{i=1}^{n} c_i \chi_{C_i} d\mu = \bigoplus_{i=1}^{n} (c_i \ \tilde{\mu}(C_i \wedge A(M, \alpha))) = \bigoplus_{i=1}^{n} (c_i \ z(\nu(M \cap C_i), \alpha)) =$$

$$= z(\sum_{i=1}^{n} c_i \ \nu(M \cap C_i), \alpha) = z(\int\limits_M f d\nu, \alpha).$$

In the case when f is NMF we have

$$\int\limits_E f \ d\mu = Sup\{z(\int\limits_M g \ d\nu, \alpha) \mid g \leq f \text{ and } g \text{ is SNMF}\} =$$

$$= z(sup\{\int\limits_M g \ d\nu, \alpha \mid g \leq f \text{ and } g \text{ is SNMF}\}, \alpha) = z(\int\limits_M f d\nu, \alpha).$$

6 Indefinite L-fuzzy valued integral

For a given real valued non-negative L-fuzzy integrable function f we define $\mu_f : \Sigma \to \mathbb{R}_+(L)$ as following:

$$\mu_f(E) = \int\limits_E f \ d\mu.$$

It is easy to show that μ_f is an L-fuzzy valued measure. And really, due to integral property ($\mathbb{I}4$) μ_f is left T_M-continuous. We need to prove that

$$\mu_f(E_1 \vee E_2) \oplus \mu_f(E_1 \wedge E_2) = \mu_f(E_1) \oplus \mu_f(E_1) \text{ for all } E_1, E_2 \in \Sigma.$$

Considering $M_1 = \{x \in X \mid E_1(x) \geq E_2(x)\}$ and $M_1 = \{x \in X \mid E_1(x) < E_2(x)\}$ we can split L-sets E_1, E_2 into two T_M-disjoint L-sets as following:

$$E_1 = (E_1 \wedge M_1) \vee (E_1 \wedge M_2) \text{ and } E_2 = (E_2 \wedge M_1) \vee (E_2 \wedge M_2).$$

Obviously,

$$E_1 \vee E_2 = (E_1 \wedge M_1) \vee (E_2 \wedge M_2) \text{ and } E_1 \wedge E_2 = (E_1 \wedge M_2) \vee (E_2 \wedge M_1).$$

Hence,

$$\int\limits_{E_1} f\ d\mu \oplus \int\limits_{E_2} f\ d\mu = \int\limits_{(E_1 \wedge M_1) \vee (E_1 \wedge M_2)} f\ d\mu \oplus \int\limits_{(E_2 \wedge M_1) \vee (E_2 \wedge M_2)} f\ d\mu =$$

$$= \int\limits_{E_1 \wedge M_1} f\ d\mu \oplus \int\limits_{E_1 \wedge M_2} f\ d\mu \oplus \int\limits_{E_2 \wedge M_1} f\ d\mu \oplus \int\limits_{E_2 \wedge M_2} f\ d\mu = \int\limits_{E_1 \vee E_2} f\ d\mu \oplus \int\limits_{E_1 \wedge E_2} f\ d\mu.$$

By analogy with the classical case μ_f is called an *indefinite L-fuzzy valued integral*.

We can also show that μ_f can be obtained in another way described by the following diagram:

$$\begin{array}{ccc} \nu & \longrightarrow & \mu \\ \downarrow & & \downarrow \\ \nu_f & \rightarrow & \mu_f \end{array}$$

First, let us consider measure ν_f (the indefinite integral of f with respect to measure ν) as following:

$$\nu_f(M) = \int\limits_M f\ d\nu, \quad M \in \Phi.$$

Now by using the construction described in the third section we obtain measure $\tilde{\mu}_f$. Taking into consideration

$$\mu_f(A(M, \alpha)) = z(\int\limits_M f\ d\nu, \alpha) \text{ and } \tilde{\mu}_f(A(M, \alpha)) = z(\nu_f(M), \alpha)$$

we obtain that values μ_f and $\tilde{\mu}_f$ are equal for any L-fuzzy set from $\wp$, that proves equality $\mu_f = \tilde{\mu}_f$.

Acknowledgement

The paper was supported by ESF project Nr.2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004.

References

[1] KLEMENT E. P., BUTNARIU D., *Triangular norm based measures.* In: E. Pap, *Handbook of Measure Theory*, Kapitel 23, Seite(n) 947-1010, Elsevier, Amsterdam, 2002.

[2] WANG Z., KLIR G.J., *Fuzzy measure theory.* Plenum Press, New York, 1992.

[3] MESIAR R., *Fuzzy measures and integrals.* Fuzzy Sets and Systems 156 (2005), 365-370.

[4] HUTTON B., *Normality in fuzzy topological spaces.* J.Math.Anal.Appl., 50 (1975), 74-79.

[5] LOWEN R., *On (R(L);⊕).* Fuzzy Sets and Systems, 10, 1983, 203 - 209.

[6] RODABAUGH S.E., *Fuzzy addition and the fuzzy real lines.* Fuzzy Sets and Systems, 8, 1982, 39 - 52.

[7] RODABAUGH S.E., *Complete fuzzy topological hyperfields and fuzzy multiplication in the fuzzy real lines.* Fuzzy Sets and Systems, 15, 1988, 285 - 310.

[8] RUZHA V., ASMUSS S., *A construction of a fuzzy valued measure based on minimum t-norm.* In New Dimensions in Fuzzy Logic and Related Technologies. Proceedings of the 5th EUSFLAT Conference, Ostrava, 2007, p. 175-178.

[9] Ruzha V., Asmuss S., *A construction of an L-fuzzy valued measure of L-fuzzy sets.* Proceedings of the IFSA-EUSFLAT 2009 Conference, Lisbon, 2009, p. 1735-1739.

[10] ASMUSS S.V. ŠOSTAK A.P., *Extremal problems of approximation theory in fuzzy context.* Fuzzy Sets and Systems, 105, 1999, 249 -257.

[11] HALMOSH P., *Measure theory*, Moscow, IL (1953) (in russian).

Current address

Vecislavs Ruza, PhD student
Zellu street 8, Riga, LV-1002, Latvia, +371 26326042,
e-mail: vecislavs.ruza@ge.com

Journal of Applied Mathematics

volume IV (2011), number III

FUZZY SEMANTIC NETWORKS

ŽÁČEK Martin, (CZ)

Abstract. Semantic networks have origins in linguistics, but this conceptual oriented mean appeared in computer science. Semantic network is easy and intelligible mean of modeling on conceptual level, because they rank among the sophisticated systems. Paper should characterize the basic possibility of semantic networks. Although their development is already closed, I am going to continue in their development in field of fuzzy modeling. Implementation of fuzzy logic into semantic networks increases expressivity of semantic networks.

Key words. Semantic networks, Fuzzy logic, Fuzzy predicate logic, Fuzzy semantic networks.

Mathematics Subject Classification: 03B52, 18C50.

1 First Section

In 1968 M. R. Quillian booted associative networks for the purpose of modeling the semantics of English sentences. The associative networks have only double figure predicate in contrast to 1st order logic, only those relations can be represented as the network that has its edges and nodes. Edges carry labels as their predicate symbols, nodes carry symbols permitted in predicate attributes, i.e. terms denotative objects represented world - the reference system.
Basic (atomic) statements of networks have the character of vector
(<subject><have a property><object>) or (<subject><predicate><object>).

Figure 1: The graph of vector.

Definition 1
The semantic network is a weighted graph consisting of nodes, labeled terms, and edges, labeled binary predicate symbols, where edges connect some pair of nodes.

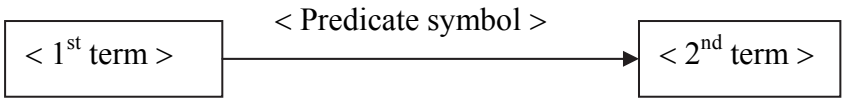


Figure 2: The semantic network.

Example 1
The statement "David likes strawberry ice cream." has the scheme like(<who><what><what kind>). This statement can express ternary predicate and can be represented with binary predicate like(<who><what>), flavor(<what><what kind>) the following graph.

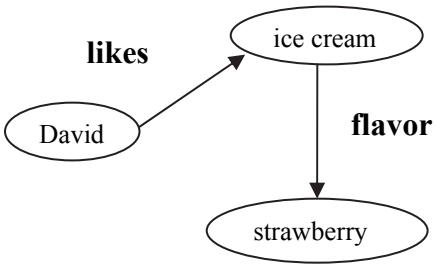


Figure 3: The statement of example 1.

The semantic networks don´t dispose of means for representation universal and existential quantification. Formula of first order logic must adjust to the special clausal form. There is a problem, which the associative networks share with clausal form logic.

Statements in the knowledge bases of associative networks are represented these types of networks:

- unconditional networks – universal and base networks,
- conditional networks – universal and base networks.

Conditions in associative networks are, just as in clausal form or predicate logic, rules with antecedent and consequent. General form of conditional is in the associative networks:

$$\mathbf{q} \text{ if } \mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_n \tag{1}$$

Conditions in the networks may have more than one atom in the antecedent, but only one (opposed to clausal form logic) in consequent. Conditions are drawn so that the atoms of antecedent are inscribed with interrupted line, consequent solid line.

Example 2
The base condition represents the statement "Dagmar is a first lady when she married to the President Vaclav."

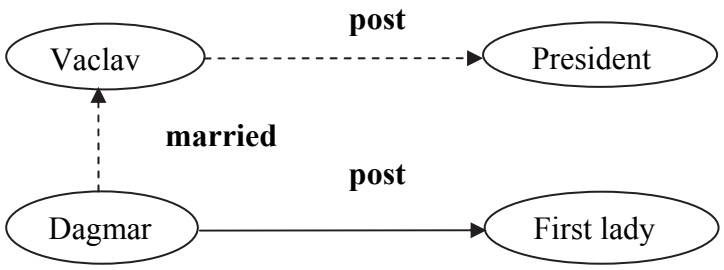


Figure 4: The base condition of example 2

To negate the semantic network a special statement is used. It stands for the false consequent in the implication with a true antecedent. Due to no existence of a definition for a contradiction semantic network, a new special symbol of network has been created, called falsum, (notation $\otimes$), which is false in all interpretations.

Example 3
Let we have statement: "David doesn't like strawberry ice cream." See following figure.

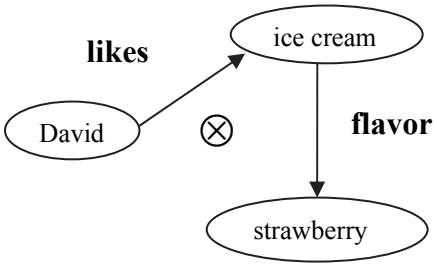


Figure 5: The statement of example 1.

2 Fuzzy Predicate Logic

The fuzzy predicate logic with evaluated syntax is a flexible and fully complete formalism, which will be used for the below presented extension [5]. In order to use an efficient form of the resolution principle we have to extend the standard notion of a proof (provability value and degree) with the notion of reputational proof (refutation degree). Propositional version of the fuzzy resolution principle has been already presented in [3]. We suppose that set of truth values is Łukasiewicz algebra. Therefore we assume standard notions of conjunction, disjunction etc. to be bound with Łukasiewicz operators.
We will assume Łukasewicz algebra to be

$$\mathcal{L}_{Ł} = \langle [0,1], \wedge, \vee, \otimes, \rightarrow, 0, 1 \rangle \tag{2}$$

where [0, 1] is the interval of real's between 0 and 1, which are the smallest and greatest elements respectively. Basic and additional operations are defined as follows:

$$a \otimes b = 0 \vee (a + b - 1) \tag{3}$$

$$a \rightarrow b = 1 \wedge (1 - a + b) \tag{4}$$

$$a \oplus b = 1 \wedge (a + b) \tag{5}$$

$$\neg a = 1 - a \tag{6}$$

The equivalence operation $\leftrightarrow$ could be defined $a \leftrightarrow b =_{df} (a \to b) \wedge (b \to a)$, where $\wedge$ is infimum operation. The following properties of $\mathcal{L}_{\mathrm{L}}$ will be used in the sequel:

- $a \otimes 1 = a$, $a \otimes 0 = 0$,
- $a \oplus 1 = 1$, $a \oplus 0 = a$,
- $a \to 1 = 1$, $a \to 0 = \neg a$,
- $1 \to a = a$,
- $0 \to a = 1$.

The syntax and semantics of fuzzy predicate logic is following:

- terms $t_1, \ldots, t_n$ are defined as in FOL (First Order Logic),
- predicates with $p_1, \ldots, p_m$ are syntactically equivalent to FOL ones. Instead of 0 we write $\perp$ and instead of 1 we write $\top$, connectives - & (Łukasiewicz conjunction), ∇ (Łukasiewicz disjunction), $\rightarrow$ (implication), $\neg$ (negation), $\forall$ (universal quantifier), $\exists$ (existential quantifier) and furthermore by F_J we denote set of all formulas of fuzzy logic in language J,
- FPL formulas have the following semantic interpretations (D is the universe): Interpretation of terms is equivalent to FOL, $D(p_i(t_{i_1}, \ldots, t_{i_n})) = P_i(D(t_{i_1}), \ldots, D(t_{i_n}))$ where P_i is a fuzzy relation assigned to p_i, $D(a) = a$ for a$\in$[0, 1], $D(A \,\&\, B) = D(A) \otimes D(B)$, $D(A \nabla B) = D(A) \oplus D(B)$, $D(A \Rightarrow B) = D(A) \to D(B)$, $D(\neg A) = \neg D(A)$, $D(\forall X(A)) = \wedge D(A[x/d] \mid d \in D)$, $D(\exists X(A)) = \vee D(A[x/d] \mid d \in D)$,
- for every subformula defined above *Sub, Sup, Pol, Lev, Qnt, Sbt, Sig* and other derived properties defined in section 2 hold (where the classical FOL connective is presented the Łukasiewicz one has the same mapping value).

Graded fuzzy predicate calculus assigns grade to every axiom, in which the formula is valid. It will be written as ^{a}/A where A is a formula and a is a syntactic evaluation.

3 Fuzzy semantic networks

Of the above is clear that the semantic networks have only double figure predicate in contrast to 1st order logic, only those relations can be represented as the network. Fuzzy semantic network created evaluated of these double figure predicates by fuzzy predicate logic.

Definition 2
The fuzzy semantic network is a weighted graph consisting of nodes, labeled terms, and edges, evaluated binary predicate symbols, where edges connect some pair of nodes.

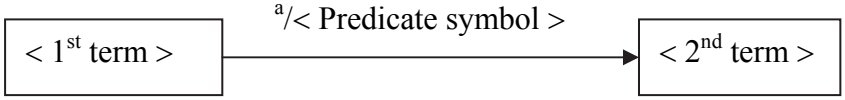


Figure 5: The fuzzy semantic network.

Classical semantic network represents a statement (see Figure 3) or negation statement (Figure 5), corresponding first order logic and these statements are evaluated either true or false, 1 or 0. For example, we have statement "David likes strawberry ice cream." where this statement is evaluated as true (or 1). And we have statement "David doesn't like strawberry ice cream." where is evaluated as false (or 0).

If we have a fuzzy semantic network, so we can evaluated double figure predicate values from the interval <0, 1> (see fuzzy predicate logic).

Example 4
We have statement "David likes an ice cream." that is evaluated value 0.6 [$^{0.6}$/likes(David, ice cream)], see following figure:

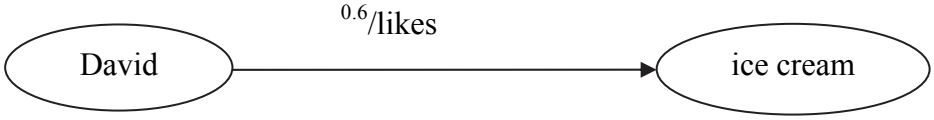


Figure 6: Semantic network of Example 4.

Example 5
Let we have statement "David likes strawberry and chocolate ice cream." and have question "What flavor of ice cream does David like the most?" Now we must adjust this statement to form of Figure 3 and evaluate double figure predicates about flavor.
Evaluated predicates:

- $^{0.9}$/likes(David, ice cream) = a,
- $^{0.7}$/flavor(ice cream, chocolate) = b,
- $^{0.3}$/flavor(ice cream, strawberry) = c.

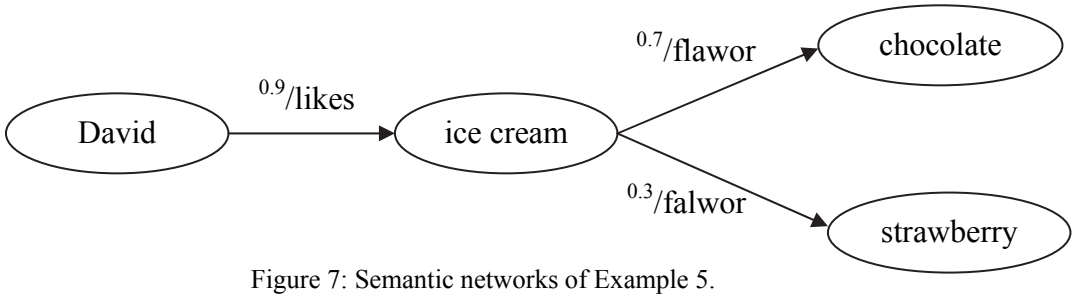


Figure 7: Semantic networks of Example 5.

So that we can answer on our question, we must evaluate base on operations (3) – (6) each network of sentences: "David likes chocolate ice cream." and "David likes strawberry ice cream."

- For first sentence holds: $a \otimes b = 0 \vee (a + b - 1) = 0 \vee (0.9 + 0.7 - 1) = 0 \vee 0.6 = 0.6$
- For second sentence holds: $a \otimes c = 0 \vee (a + c - 1) = 0 \vee (0.9 + 0.3 - 1) = 0 \vee 0.2 = 0.2$

Now we can answer on our question so that David likes chocolate ice cream the most.

Example 6
Consider following statement: "Exist child who is happy, if has mother and father at the same time."

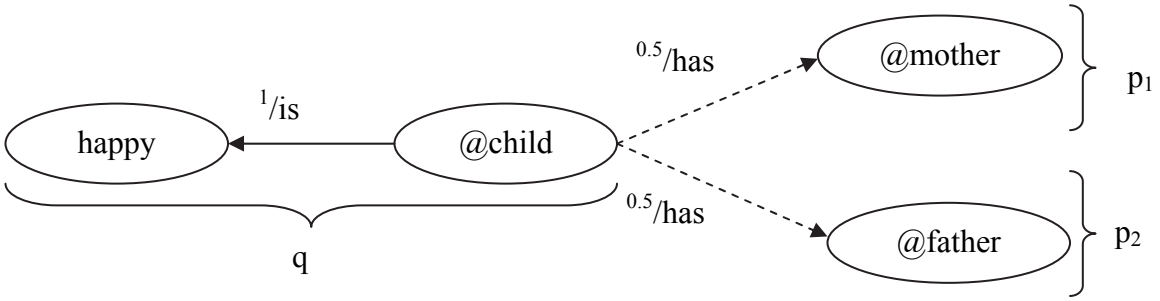


Figure 8: Semantic networks of Example 6.

Because we have conditional universal networks, therefore hold the rule (1), where:

- q is ^{1}/is(@child, happy),
- p_1 is $^{0.5}$/has(@child, @mother),
- p_2 is $^{0.5}$/has(@child, @father),

For rule (1) holds: q if p_1, p_2 and this rule we can rewrite to form: $q \rightarrow p_1 \,\&\, p_2$.
Now we can use this modified rule for evaluated whole networks:

1. $q \rightarrow p_1 \,\&\, p_2$
2. $1 \rightarrow 0.5 \,\&\, 0.5$ substitute
3. $0.5 \,\&\, 0.5 \Rightarrow 0.5 \oplus 0.5 = 1 \wedge \left(0.5 + 0.5\right) = 1$ apply rule (5)
4. $1 \rightarrow 1 \Rightarrow 1$ apply rule implication
5. Child is happy. conclusion

These steps are proof that exist child who is happy if has both parents.

References

[1.] LUKASOVÁ, A.: Formální logika v umělé inteligenci. Computer Press, 2003.
[2.] LUKASOVÁ, A.: Reprezentace znalostí v Asociativních sítích. Znalosti 2001.
[3.] HABIBALLA, H.: Non-clausal Resolution Theorem Proving for Fuzzy Description Logic. SOFSEM 2006: Theory and practice of computer science 32nd conference on current trends in theory and practice of computer science. Praha: Institute of Computer Science, Czech Academy of Sciences, 2006. s. 1-12. ISBN 80-903298-4-5
[4.] QUILLIAN, M. R.: Semantic memory. In Minsky, M. (ed.): Semantic Information processing, MA: MIT Press, pp. 27-70, 1968.
[5.] NOVÁK, V., PERFILIEVA, I., Močkoř, J.: Mathematical Principles of Fuzzy Logic. Kluwer, Boston 1999., Kap. 5, paper 5.3, 5.4.
[6.] NOVÁK, V.: Základy fuzzy modelování, BEN, Praha 2000.

Current address

Martin Žáček, Mgr.
University of Ostrava, Faculty of Science,
Department of Informatics and Computers, 30. dubna 22, 702 00,
Phone: +420 597 092 147
e-mail : Martin.Zacek@osu.cz

Journal of Applied Mathematics

volume IV (2011), number III

PATERNITIES SEARCH WITH OBJECT-ORIENTED BAYESIAN NETWORKS

ANDRADE Marina, (PT), FERREIRA, Manuel Alberto M., (PT)

Abstract. Paternity dispute problems are examples of situations in which forensic approach the DNA profiles study is a common procedure. To implement this approach an efficient tool are the object-oriented Bayesian networks (OOBN). Along this paper are presented the various OOBN adequate to solve the simple paternity dispute and more complex paternity dispute problems with incomplete DNA profiles data about the putative father such as: only putative grandfather information, only putative uncle information, only putative father 's uncle information and only simultaneously putative uncle and putative father's uncle information. Here it is exhibited an algebraic treatment, for the simple problem and with those the use of the object-oriented Bayesian networks is shown. Then the most complex kind of problems that may occur is presented. Although these are not the most common cases there is notice of its occurrence at least in Portuguese courts.

Key words: Bayesian networks, DNA profiles, paternity dispute problems.

Mathematics Subject Classification: Primary 62C10; Secondary 62P99.

1 Introduction

The use of Bayesian networks in forensic identification problems has raised more and more attention, even for the social impact of these problems. It is usually recognized that the paternity dispute problems approach using Bayesian networks began with the works of Dawid et al. (2002) and Lauritzen (2003). In Andrade (2007) the use of this tool in paternity dispute and criminal cases is discussed. Some of the paternity dispute cases discussed here, although not the more frequent in courts, already occurred. And given its specificity justify the discussion and the use of Bayesian networks in the computation of a measure of the available evidence.

In the developed countries the application of the forensic identification statistics approach has grown significantly. The use of DNA evidence in forensic identification problems tries essentially

to look for answers to the logical and computational challenges that may occur in more complex situations such as, for instance, incomplete data.

The OOBN adequate to solve the simple paternity dispute is presented first jointly with the alternative algebraic treatment for checking purpose. Then the OOBN for the more complex paternity dispute problems with incomplete DNA profiles data about the putative father such as:

- only putative grandfather information,
- only putative uncle information,
- only putative father 's uncle information,
- only simultaneously putative uncle and putative father's uncle information

are shown. In these cases an algebraic treatment is out of question being the computational procedure imperative.

2 Simple Paternity Dispute

In a disputed paternity decision problem there are formally two challenging hypotheses (prosecution and defense):

H_P: The true father is the putative father.

vs

H_D: The true father is another individual randomly drawn from the population, and not genetically related with the mother or the putative father.

The court has to decide about the paternity of the child, and so, after Bayes' Law

$$\frac{P(H_P \mid E)}{P(H_D \mid E)} = \frac{P(E \mid H_P)}{P(E \mid H_D)} \times \frac{P(H_P)}{P(H_D)} \qquad (1),$$

with E the vector containing the available evidence, genetic information of the mother (*mgt*), of the child (*cgt*) and of the putative father (*pfgt*), being the algebraic approach simple.

It is needed to assess the likelihood function over the hypotheses as to the true father, i.e., to evaluate the likelihood ratio:

$$LR = \frac{P(E \mid H_P)}{P(E \mid H_D)} \qquad (2).$$

Naturally the court has to answer to the truly paternity of the child. So it has to evaluate the ratio of the hypotheses in dispute. Admitting that $P(H_P) = P(H_D)$ then (1) becomes

$$\frac{P(H_P \mid E)}{P(H_D \mid E)} = \frac{P(E \mid H_P)}{P(E \mid H_D)} \qquad (3).$$

In fact, knowing that the markers are in different chromosomes (*linkage equilibrium*) and assuming random mating (*Hardy-Weinberg equilibrium*) there is independence between and within markers. Thus, it is possible to obtain the *LR* for each marker separately and finally multiply the values to determine the overall likelihood ratio based on the data available for all markers.

To determine algebraically the probability of the triplet *E*, under the two hypotheses, it is reasonable to consider that before knowing any data on the child it is reasonable to assume that the identity of the true father is independent of the mother's and the putative father's. And supported on that, it is easily seen that it is possible to determine the conditional probability of the child's genotype, given the other two available genotypes. Thus, to determine $P(E \mid H_P)$ one has only to apply Mendel's laws. But the calculus of $P(E \mid H_D)$ necessarily demands the knowledge of the population allele frequencies for the considered markers.

If for a certain marker the triplet $E = (mgt, cgt, pfgt)$ is $E = ((A,B);(B,B);(A,B))$, and p_A and p_B are the population allele frequencies then

$$\begin{aligned} P(E \mid H_P) &= P[(mgt; cgt; pfgt) \mid (mgt; pfgt)] \\ &= P[cgt \mid (mgt; pfgt)] \\ &= 0.5 \times 0.5 \end{aligned}$$

and

$$\begin{aligned} P(E \mid H_D) &= P[(mgt; cgt; pfgt) \mid (mgt; rgt)] \\ &= P[cgt \mid (mgt; rgt)] \\ &= 0.5 \times p_B \end{aligned}$$

where *rgt* assigns the genotype of a random individual of the population, not related to the mother or the putative father.

Therefore,

$$LR = \frac{0.5}{p_B}.$$

The considered problem is, as shown, easily algebraically solved. It is used to illustrate the simplicity and the advantages of this tool in more complex situations. Given the freedom of choice for the variables to include in the graphical approach, different representations can be obtained. Some of them simpler than others. To get a 'good' representation is very important to the efficiency and the viability of the computational routines. These are extremely sensible to the organization of the graphical structure. The first step consists on the identification and definition of the nodes for all the variables of interest to the problem.

Then the graphical representation can be obtained. According to Dawid *et al*. (2002), *in order to maximize the efficiency of the calculations as well as the logical clarity of the representation we chose to disaggregate each individual's genotype into its constituent, unobserved, paternally and maternally inherited genes*.

Figure 1 exhibits the OOBN for a paternity case as the discussed above considering a single marker. Each node (instance) in the network represents itself a Bayesian network.

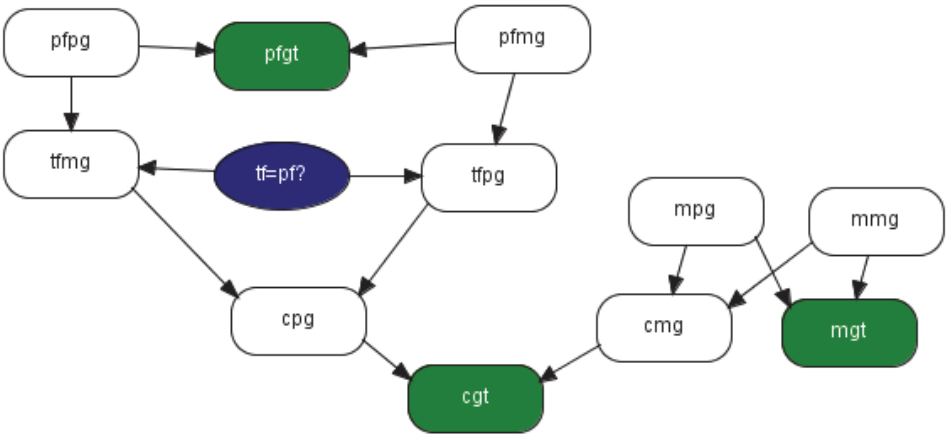


Figure 1: **Simple paternity network.**

In this simple paternity case instances **pfmg**, **pfpg**, **mpg** and **mmg** are all of class **founder**, a single node *gene,* having for its space of states all the possible alleles that can be presented for the specific case, and the correspondent population gene frequencies. Instances **mgt**, **cgt** and **pfgt** are of class **genotype**, an unordered pair of alleles inherited from paternal, *pg*, and maternal, *mg*, genes, here represented by *gtmin:=min*{*pg*, *mg*} and *gtmax:=max*{*pg*, *mg*}, where *pg* and *mg* are input nodes identical to the *gene* node of **founder**. Instances **tfmg** and **tfpg** are of class **whom**, describing the true father's allele origin. If *tf=pf?* has true for value then the true father's allele, *tfg*, will be identical with the putative father's, *pfg*, otherwise the true father's allele is randomly chosen from another man in the population. And **cpg** and **cmg** instances are of class **inherit**, modelling the Mendel's inheritance in which the child's allele is chosen at random from the two parents, *pg* and *mg*, here as the sequence of the observed outcome of a fair coin toss.

For illustration according to Dawid *et al*. (2002), the data for marker FES are child genotype *cgt* = {*B, B*}, mother's genotype *mgt* = {*A, B*} and putative father's genotype = {*A, B*}. The population allele frequencies are $p_A = 0.28425$ and $p_B = 0.25942$.

After specifying the network, put it to run and then insert the evidence. Considering equal prior probabilities for the query node representing the hypotheses, the likelihood is got after inserting the evidence. The likelihood ratio, based on the data for this marker, is obtained from the marginal posterior distribution of the query node. Thus, $P(tf = pf\,? := true \mid E) = 0.6584$ and $P(tf = pf\,? := false \mid E) = 0.3416$, and $LR = 1.9274$, being these results in agreement with the algebraic approach (note that $0.5/0.25942 \cong 1.9274$).

3 Paternities search in more uncommon situations

When the data E are not in the form $(mgt, cgt, pfgt)$ it is not possible to determine in algebraic form the likelihood function for the various hypotheses, i.e. to determine the weight of the genetic connection of the child with the putative father ancestor(s). The use of Bayesian networks allow to overcome these problems. These networks are a good tool to compute the likelihood functions.

Forwarding and backwarding the information a measure of the "strength" of the information available in each case is obtained.

In the sequence the networks for the uncommon cases described in the introduction are presented each one together with a numerical example.

The data considered are the same for the whole cases and are in Table 1 where five different markers are considered and the respective genotypes for the mother, the child, the grandfather, the uncle and the grandfather brother, where * indicates rare alleles, and (*a*) signs alleles considered as good discriminate markers, with more than 10 alleles in each marker.

Table 1:

Marker	*mgt*	*cgt*	*gfgt*	*ungt*	*gfbgt*
D3S1358	16, 18	13*, 16	13*, 17	13*, 16	13*, 15
VWA	16, 17	13*, 16	13*, 16	16, 18	13*, 15
D16S539	11, 12	12, 12	9, 12	10, 12	12, 13
D8S1179	12, 13	13, 17*	14, 17*	14, 15	12, 17*
D21S11(a)	29, 31.2	29, 31.2	29, 31.2	28, 31.2	29, 30

Genetic profiles

In Table 2 the respective allelic frequencies are presented: p_i is the i allele frequency in the population.

Marker	**Frequencies**				
D3S1358	p_{13}	p_{15}	p_{16}	p_{17}	p_{18}
	0.0032	0.2611	0.2477	0.2065	0.1606
VWA	p_{13}	p_{15}	p_{16}	p_{17}	p_{18}
	0.0023	0.1216	0.2300	0.2649	0.1859
D16S539	p_{9}	p_{10}	p_{11}	p_{12}	p_{13}
	0.1431	0.0545	0.3009	0.2876	0.1654
D8S1179	p_{12}	p_{13}	p_{14}	p_{15}	p_{17}
	0.1351	0.3028	0.2178	0.1223	0.0031
D21S11(a)	p_{28}	p_{29}	p_{30}	$p_{31.2}$	
	0.1674	0.2136	0.2437	0.1138	

Table 2: **Allele frequencies**

The allelic frequencies used were collected in
www.uni.duesseldorf.de/WWW/MedFak/Serology/dna.htm
for Portugal (Azores and Madeira archipelagos not included).

4 Only putative grandfather information

Bayesian networks for more complex problems can be built out of the same fundamental local modules that we have already described for the simple paternity dispute problem, Dawid et al. (2002).

The object-oriented Bayesian network for the "only putative grandfather information" case is shown in Figure 2. Note, for example, the node gfgt (grandfather genotype) and the respective connections with the other nodes.

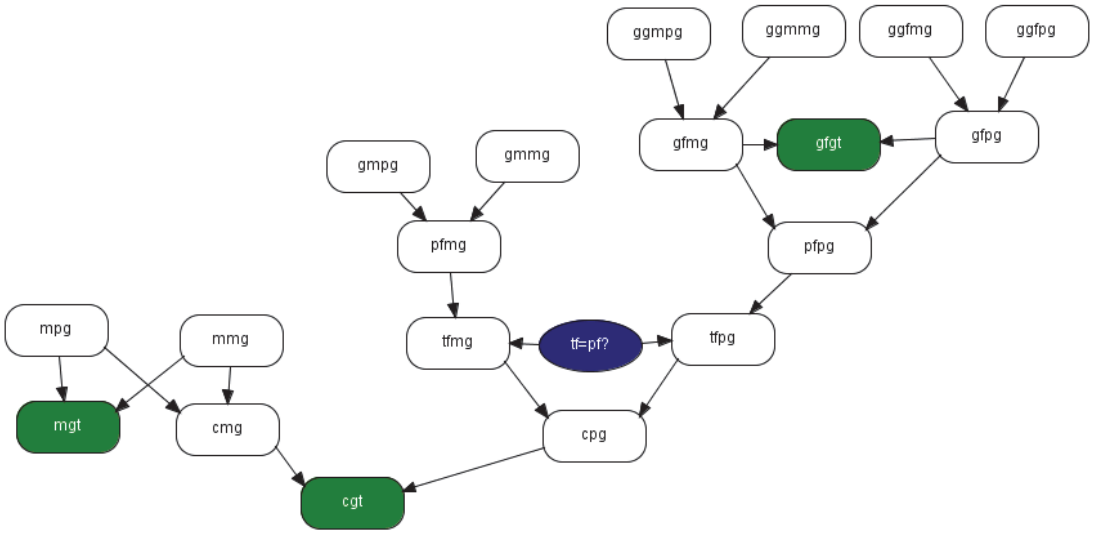


Figure 2**: Only putative grandfather network**

The results obtained are in Table 3. In the last column **Rescaled** – corrected so that the sum of the entries is equal to 1 – is presented the result for the 5 markers. Since the markers are independent the final result is obtained by multiplying the result obtained for each marker.

	D3S1358	**VWA**	**D16S539**	**D8S1179**	**D21S11**	**Rescaled**
$P(H_P\|E)$	0.9874	0.9909	0.5779	0.9878	0.6255	0,999999
$P(H_D\|E)$	0.0126	0.0091	0.4221	0.0122	0.3745	6,33E-07

Table 3: **Analysis results with only putative grandfather information**

5 Only putative uncle information

The object-oriented Bayesian network for the "only putative uncle information" case is shown in Figure 3.

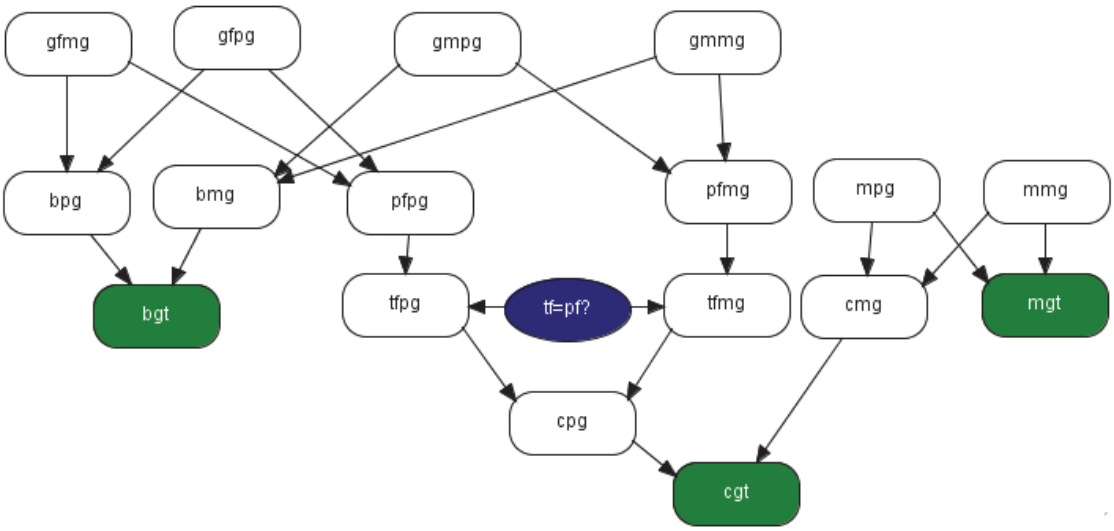


Figure 3**: Only putative uncle network**

The results obtained are in Table 4 following the same methodology as in section 5.

	D3S1358	**VWA**	**D16S539**	**D8S1179**	**D21S11**	**Rescaled**
$P(H_P\|E)$	0.9874	0.3333	0.5779	0.3333	0.5582	0,97133
$P(H_D\|E)$	0.0126	0.6667	0.4221	0.6667	0.4418	0,02867

Table 4: **Analysis results with only putative uncle information**

6 Only putative father ‘s uncle information

In Figure 4 the object-oriented Bayesian network for the “only putative father’s uncle information” case is shown.

It is a network more complex than the former ones owing to the further parentage relationship considered, that implies more complex genetic connections.

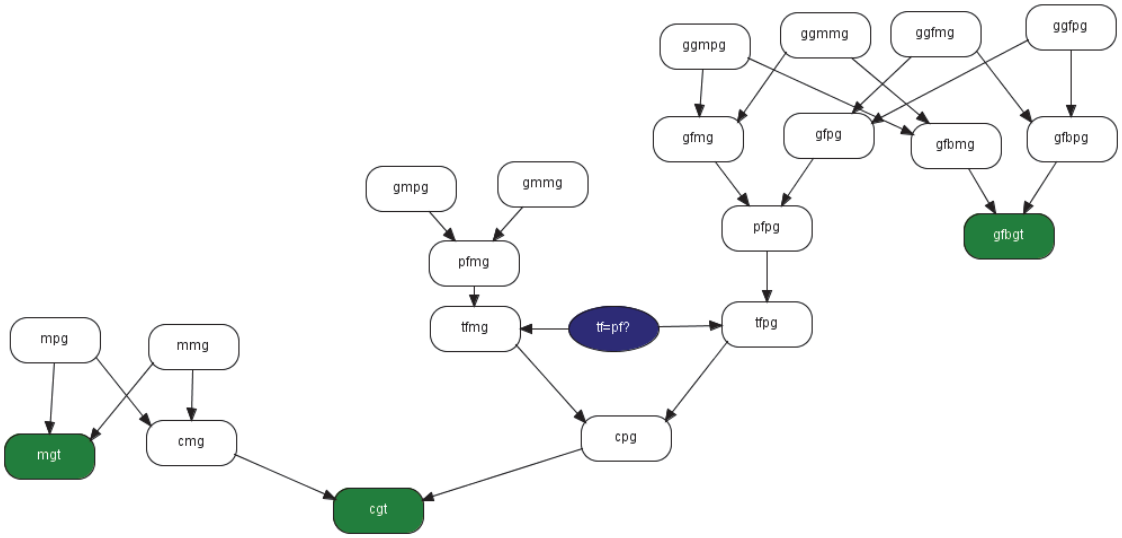


Figure 4**: Only putative father 's uncle network**

The results obtained are in Table 5.

	D3S1358	**VWA**	**D16S539**	**D8S1179**	**D21S11**	**Rescaled**
$P(H_P\|E)$	0.9755	0.9822	0.5423	0.9762	0.5309	0,999992
$P(H_D\|E)$	0.0245	0.0178	0.4577	0.0238	0.4691	8,28E-06

Table 5: **Analysis results with only putative father 's uncle information**

7 Only simultaneously putative uncle and putative father's uncle information

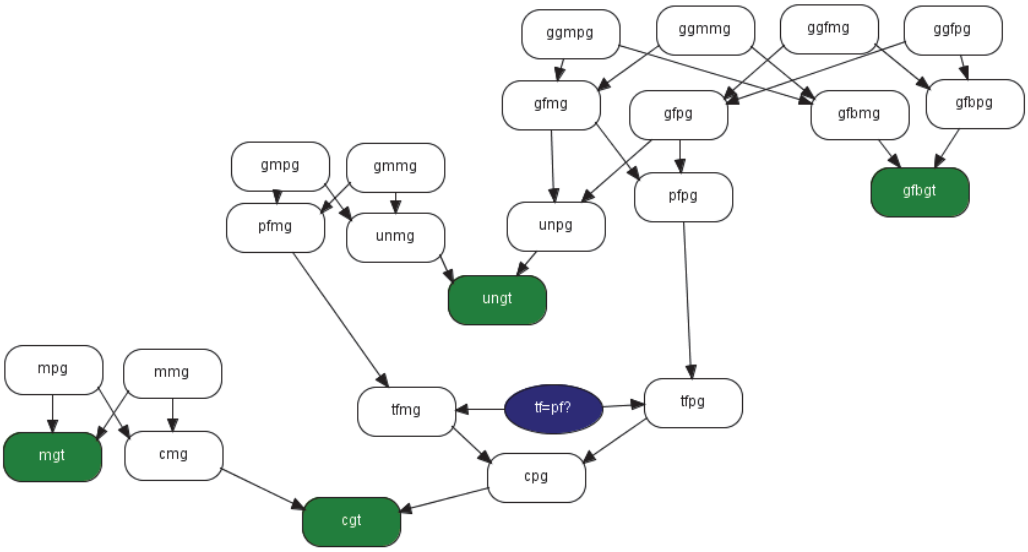


Figure 5**: Only simultaneously putative uncle and putative father's uncle**

The "only simultaneously putative uncle and putative father's uncle information" case network is the last one presented (in Figure 5) and the results are presented in Table 6.

	D3S1358	VWA	D16S539	D8S1179	D21S11	Rescaled
$P(H_P\|E)$	0.9875	0.9650	0.5764	0.9536	0.5707	0,999988
$P(H_D\|E)$	0.0125	0.0350	0.4236	0.0464	0.4293	1,23E-05

Table 6: **Analysis results with only simultaneously putative uncle and putative father's uncle information**

8 Conclusions

The paternities search in more uncommon cases demands the calculation of probabilities in the context of numerous and complex successive uses of Bayes Law. This situation is impossible to be treated algebraically. It was shown that the object-oriented Bayesian networks are a very powerful tool, very simple to use, that allows the referred calculations in an efficient way.

The major problem is to build the network taking in account the various and complex connections that may occur in parentage relationships. Then the use of an adequate software as Hugin or SPSS makes easy to apply it in practical cases. In this work Hugin was the chosen.

Inspecting the tables of results one can note that, as expected, rare alleles shared lead to greater probabilities of true paternity. On the contrary, more frequent alleles shared lead to lesser probabilities.

With the particular data used the final probabilities for true paternity were in general great.

References

[1] ABRANTES, D., PONTES, M. L., PINHEIRO, M. F., ANDRADE, M. and FERREIRA, M. A. M.: *Towards a systematic probabilistic evaluation of parentage casework in forensic genetics: A modest attempt to define a general standardized approach to simple and complex cases*. Forensic Science International: Genetics Supplement Series 1, pp. 635-637, 2008.

[2] ANDRADE, M.: *A Estatística Bayesiana na Identificação Forense – análise e avaliação de vestígios de DNA com redes Bayesianas*. PhD Thesis, ISCTE, Lisboa, 2007.

[3] ANDRADE, M.: *A Note on Foundations of Probability*. Journal of Mathematics and Technology, vol.. 1 (1), pp 96-98, 2010.

[4] ANDRADE, M., FERREIRA, M. A. M. and FILIPE, J. A.: *Evidence evaluation in DNA mixture traces*. Journal of Mathematics, Statistics and Allied Fields (Scientific Journals International-Published online), vol. 2 (2), 2008.

[5] ANDRADE, M., FERREIRA, M. A. M., FILIPE, J. A. and COELHO, M.: *Paternity dispute: is it important to be conservative?*. Aplimat – Journal of Applied Mathematics, vol. 1 (2), 2008.

[6] ANDRADE, M. and FERREIRA, M. A. M.: *Bayesian networks in forensic identification problems*. Aplimat - Journal of Applied Mathematics, vol. 2 (3), pp. 13-30, 2009.
[7] ANDRADE, M. and FERREIRA, M. A. M.: *Civil Identification Problems with Bayesian Networks Using Official DNA Databases*. Aplimat-Journal of Applied Mathematics, vol. 3 (3), pp. 155-162, 2010.
[8] ANDRADE, M. e FERREIRA, M. A. M.: *Solving civil identification cases with DNA profiles databases using Bayesian nerworks*. Journal of Mathematics and Technology, 1(2), pp. 37-40, 2010.
[9] ANDRADE, M. e FERREIRA, M. A. M.: *Evaluation of Paternities with less usual Data using Bayesian Networks*. *IEEE Xplore (BMEI 2010 IEEE Catalog Number CFP1093D-PRT, ISBN: 978-1-4244-6496-8)*, 2010.
[10] ANDRADE, M., FERREIRA, M. A. M.,ABRANTES, D., PONTES, M. L. e PINHEIRO, M. F.: *Object-oriented Bayesian Networks in the evaluation of paternities in less usual environments*. Journal of Mathematics and Technology, 1(1), pp. 161-164, 2010.
[11] DAWID, A. P., MORTERA, J., PASCALI, V. L. and van BOXEL, D. W.: *Probabilistic expert systems for forensic inference from genetic markers.* Scandinavian Journal of Statistics vol. 29, pp. 577-595, 2002.
[12] FERREIRA, M. A. M. and ANDRADE, M.: *A note on Dawnie Wolfe Steadman, Bradley J. Adams, and Lyle W. Konigsberg, Statistical Basis for Positive Identification in Forensic Anthropology. American Journal of Physical Anthropology 131: 15-26 (2006)*. International Journal of Academic Research, vol. 1 (2), pp. 23-26, 2009.
[13] LAURITZEN, S. L.: *Bayesian networks for forensic identification Problems*. Tutorial 19th Conference on Uncertainty in Artificial Intelligence, Mexico, 2003.

Current address

Marina Andrade, Professor Auxiliar
ISCTE – Lisbon University Institute
UNIDE - IUL
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
UNIDE - IUL
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

Journal of Applied Mathematics

volume IV (2011), number III

THE BANKING EFFICIENCY MEASUREMENT USING THE FRONTIER ANALYSIS TECHNIQUES

ARSHINOVA Tatyana, (LV)

Abstract. The research focus of the scientific paper is on the problem of performance measurement of credit institutions. The author recommends applying frontier analysis techniques such as Data Envelopment Analysis and Stochastic Frontier Approach to the efficiency analysis of Decision Making Units. Using modern computer technologies, the author has calculated dynamics of efficiency score of ten Latvian banks on the basis of DEA CCR approach, provided recommendations concerning optimal input volumes and established hidden development reserves using SFA method.

Key words: DEA (Data Envelopment Analysis), Decision Making Units (DMUs), efficiency measurement, SFA (Stochastic Frontier Approach)

Mathematics Subject Classification: 90B30, 90B50, 90C08, 90C15, 90C30, 90C90.

Introduction

The remaining uncertainty in development of Latvian national economy and the absence of significant improvement of economical situation impacts activities of all economical subjects. Due to this fact, credit institutions that are performing redistribution of income inside of the country are especially vulnerable. In June 2010 21 banks and seven branches of foreign banks were functioning in Latvia that is indicative of increasing level of competition in the banking sector. Due to the absence of improvement in quality of credit portfolio, the volumes of reserves for non-performing loans in the first current half-year have reached 1,671 million LVL. Operating profits of Latvian banks continue decreasing, mainly because of credit impairments that made 49,3% of total banking losses (430,3 million LVL) in the end of June 2010. These negative macroeconomical trends that have impact on activities of credit institutions are indicative of necessity of strong control over banking performance.

Actually the estimation of the level of operating efficiency in the most Latvian banks is realized on the basis of quantitative approach of ratio analysis. Ratios measure the relationship between two variables chosen to provide insights into different aspects of the banks multifaceted operations, such as liquidity, profitability, capital adequacy, asset quality, risk management, and many others.

Although the traditional ratio measures are attractive to analysts due to their simplicity, there are several limitations that must be considered. For example, the analysis assumes comparable units, which implies constant returns to scale (Smith 1990). Each of the indicators yields a one-dimensional measure by examining only a part of the organization's activities, or combining the multiple dimensions into a single, unsatisfactory number. Moreover, the seemingly unlimited number of ratios that can be created from financial statement data are often contradictory, thus ineffective for the assessment of overall performance. This overly simplistic analytical approach offers no objective means of identifying inefficient units and requires a biased separation of the inefficient and efficient levels. [1, 350-351]

Methods of frontier analysis ensure a principally different approach to the problem of efficiency measurement. They provide an opportunity of complex analysis of banking efficiency level for a certain period of time and comparison of it among investigated banks. This multidimensional approach meets the requirements to the banking performance evaluation methodology. The objective of the author's research is to improve and supplement the methodology of efficiency measurement of Latvian banks on the basis of methods of frontier analysis.

In the circumstances of unstable macroeconomical environment and competition, profitability is one of the most important indicators of stability and development of credit institutions. In this connection, the author analyzed the performance of a set of Latvian banks, assuming operating profits as an output. The objects of the research are members of Latvian banking sector; their efficiency level is analyzed over the time period from 2003 to 2009. Evaluating the performance on the basis of frontier approaches, the author included into the set of investigated objects banks that take leading positions on Latvian market (according to the volumes of total assets): JSC "Swedbank", JSC "DnB Nord Banka", JSC "Aizkraukles Banka", JSC "Parex banka" (currently JSC "Citadeles Banka"), JSC "SEB Banka", JSC "Latvijas Krājbanka", JSC "Mortgage Bank", JSC "Rietumu Banka", JSC "Norvik Banka", JSC "GE Money Bank" with the exception of branches of foreign banks.

1 Methods of Frontier Data Analysis

The progress of production technology and increase of production volumes have stimulated the development of performance measurement methodology. In the second part of the 20th century there were introduced methods of frontier data analysis that provided a qualitatively different approach to the problem. According to the methodology of methods of frontier data analysis, the efficiency score of investigated DMUs is calculated as a distance from the point that defines the production process of a Decision Making Unit (DMU) to the certain efficiency frontier. Entities that are functioning on the efficiency frontier are considered to be absolutely technically efficient; inefficiency of other DMUs is increasing together with extension of the distance to the efficiency frontier.

Methods of frontier analysis may be divided into two groups: parametric (Stochastic Frontier Approach (SFA), Distribution-Free Approach (DFA), Thick Frontier Approach (TFA)) and non-parametric (Data Envelopment Analysis (DEA), Free Disposal Hull (FDH)) methods. In accordance with parametric approaches, the efficiency frontier is constructed on the basis of econometric modeling, usually in form of Cobb-Douglas (log-linear) production function. Econometric analyses include two error components: an error term that captures inefficiency (u_i) and a random error (v_i). Parametric methods have significant advantages – they provide the possibilities to use panel data, to distinguish the random noise from inefficiency and to calculate the standard error of efficiency

measurement results. Nevertheless, the stochastic approaches of performance measurement presume the comparison of investigated DMUs efficiency to the theoretically developed benchmark frontier; therefore the optimal combinations of inputs and outputs sometimes are not achievable practically. The application of parametric methods also requires observance of the restrictions imposed on the distributional assumptions on the inefficiencies and random error. In contrast to the econometric approaches, non-parametric methods are based on the hypothesis that the efficiency frontier is generated from the empirical results of the most efficient DMUs i.e. benchmarks that „float" on the piecewise linear frontier. The level of technical efficiency of these DMUs is 100%. While mathematical, non-parametric methods require few assumptions when specifying the best-practice frontier, they generally do not account for random errors. [8, 93]

1.1 Data Envelopment Analysis (CCR DEA Model)

The CCR DEA model was developed by Charnes, Cooper and Rhodes in 1978 to evaluate the performance of Decision Making Units (DMUs). To allow for applications to a wide variety of activities, the term DMU might be used to refer to any entity that is to be evaluated in terms of its abilities to convert inputs into outputs. These evaluations can involve governmental agencies and non-profit organizations as well as business firms, hospitals and educational institutions.
The production process might be aimed either at minimization of resources or maximization of production volumes. The orientation of the model should be aimed at controllable variables. In context of banking, volumes of resources are usually over control of management; therefore only input-oriented model will be examined in the paper.
The measurement of comparative efficiency is based on the assumption that the performance of each DMU is calculated in comparison to $\boldsymbol{n}$ investigated DMUs. Each DMU consumes varying amounts of $\boldsymbol{m}$ different inputs to produce $\boldsymbol{s}$ different outputs. Specifically, $\boldsymbol{DMU_j}$ consumes amount $\boldsymbol{x_{ij}}$ of input $\boldsymbol{i}$ and produces amount $\boldsymbol{y_{rj}}$ of output $\boldsymbol{r}$. It is necessary to assume that $\boldsymbol{x_{ij}} \geq 0$ and $\boldsymbol{y_{rj}} \geq 0$ and further to assume that each DMU has at least one positive input and one positive output value. Primarily the DEA model was expressed in fractional, i.e. ratio-form. In this form the ratio of outputs to inputs is used to measure the relative efficiency of the $DMU_j = DMU_0$ to be evaluated relative to the ratios of all of the $\boldsymbol{j = 1,2, ..., n}$ DMU_j. The CCR construction can be interpreted as the reduction of the multiple-output/multiple-input situation (for each DMU) to that of a single 'virtual' output and 'virtual' input. For a particular DMU the ratio of this single virtual output to single virtual input provides a measure of efficiency that is a function of the multipliers. In mathematical programming parlance, this ratio, which is to be maximized, forms the objective function for the particular DMU being evaluated. A set of normalizing constraints (one for each DMU) reflects the condition that the virtual output to virtual input ratio of every DMU, including $DMU_j = DMU_0$, must be less than or equal to unity. The mathematical programming problem may thus be stated as (1):

$$\begin{aligned} &\max\ h_0(u,v) = \sum_r u_r y_{ro} \Big/ \sum_i v_i x_{io} \\ &subject\ to \\ &\sum_r u_r y_{rj} \Big/ \sum_i v_i x_{ij} \leq 1\ \ for\ \ j = 1,\ldots, n, \\ &u_r, v_i \geq 0\ \ for\ all\ i\ and\ r, \end{aligned} \tag{1}$$

where

h_0 – the function of virtual output and virtual input ratio of DMU_0;

u_r – the output multiplier of DMU_0;
v_i – the input multiplier of DMU_0;
y_{r0} – the output of DMU_0;
x_{i0} – the input of DMU_0;
y_{rj} – outputs of 1,2…n DMUs;
x_{ij} – inputs of 1,2…n DMUs.

The above ratio form yields an infinite number of solutions; if ***(u*, v*)*** is optimal, then ***(αu*, αv*)*** is also optimal for ***α > 0***. However, the transformation developed by Charnes and Cooper (1962) for linear fractional programming selects a representative solution [the solution ***(u, v)*** for which $\sum_{i=1}^{m} v_i x_{io} = 1$ and yields the equivalent linear programming problem in which the change of variables from ***(u, v)*** to ***(μ, v)*** is a result of the Charnes-Cooper transformation (2):

$$\max z = \sum_{r=1}^{s} \mu_r y_{ro}$$

subject to

$$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0$$

$$\sum_{i=1}^{m} v_i x_{io} = 1$$

$$\mu_r, v_i \ge 0, \qquad (2)$$

μ_r – the output multiplier of DMU_0;
v_i – the input multiplier of DMU_0;
y_{r0} – the output of DMU_0;
x_{i0} – the input of DMU_0;
y_{rj} – outputs of 1,2…n DMUs;
x_{ij} – inputs of 1,2…n DMUs.

Model that is expressed by (2) can be solved by its dual problem (3):

$$\theta^* = \min \theta$$

subject to

$$\sum_{j=1}^{n} x_{ij} \lambda_j \le \theta x_{io} \qquad i = 1,2,..., m;$$

$$\sum_{j=1}^{n} y_{rj} \lambda_j \ge y_{ro} \qquad r = 1,2,..., s;$$

$$\lambda_j \ge 0 \qquad j = 1,2,..., n, \qquad (3)$$

where
θ^* – the optimal value of dual variable θ of DMU_0;
θ, λ_j – dual variables of DMU_0;
y_{r0} – the output of DMU_0;
x_{i0} – the input of DMU_0;

y_{rj} – outputs of 1,2…n DMUs;
x_{ij} – inputs of 1,2…n DMUs.

By virtue of the dual theorem of linear programming we have $z^* = \theta$. Hence either problem may be used. One can solve the dual linear program, to obtain an efficiency score. Setting $\theta = 1$ and $\lambda_k^* = 1$ with $\lambda_k = \lambda_o^*$ and all other $\lambda_k^* = 0$, a solution of dual problem (see Formula 3) always exists. Moreover this solution implies $\theta^* \leq 1$. The optimal solution, θ^*, yields an efficiency score for a particular DMU.
The process is repeated for each DMU. i.e., solving the model, expressed by Formula 3, with $(X_o, Y_o) = (X_k, Y_k)$, where (X_k, Y_k) represent vectors with components x_{ik} , y_{rk} and, similarly (X_o, Y_o) has components x_{ok} , y_{ok}. DMUs for which $\theta^* < 1$ are inefficient, while DMUs for which $\theta^* = 1$ are boundary points. Some boundary points may be "weakly efficient" because we have non-zero slacks. This may appear because alternate optima may have non-zero slacks in some solutions, but not in others. However, we can avoid this effect by invoking the following linear program in which the slacks are taken to their maximal values (4).

$$\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta^* x_{io} \qquad i = 1,2,...,m;$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{ro} \qquad r = 1,2,...,s;$$

$$\lambda_j, s_i^-, s_r^+ \geq 0 \;\forall\; i, j, r, \tag{4}$$

s_r^+ – output slacks;
θ^* – the optimal value of dual variable θ of DMU_0;
λ_j – the dual variable of DMU_0;
y_{r0} – the output of DMU_0;
x_{i0} – the input of DMU_0;
y_{rj} – outputs of 1,2…n DMUs;
x_{ij} – inputs of 1,2…n DMUs.

It shall be noted that the choices of s_i^- and s_r^+ do not affect the optimal θ^* which is determined from model expressed by (3). These developments lead to the following definitions of DEA efficiency:
DEA Efficiency: The performance of DMU_0 is fully (100%) efficient if and only if both (i) $\theta^* = 1$ and (ii) all slacks $s_i^{-*} = s_r^{+*} = 0$.
Weakly DEA Efficiency: The performance of DMU_0 is weakly efficient if and only if both (i) $\theta^* = 1$ and (ii) $s_i^{-*} \neq 0$ and/or $s_r^{+*} \neq 0$ for some i and r in some alternate optima. [2, 8-12]
The CCR efficiency score is indicative of the overall efficiency level of investigated DMUs.

2 The Stochastic Frontier Approach

The Stochastic Frontier Approach (SFA) is the econometric method of efficiency estimation. It presumes a certain functional form for the description of the process of production. The performance of investigated banks is estimated on the basis of stochastic Cobb-Douglas production

function. The SFA method allows estimating the level of technical efficiency. Assuming that the production function is depending on several factors $x_1,...,x_n$, its functional form is $y = F(x_1,...,x_n)$, and the functioning bank using a similar volume of resources can produce at least the same volume of production if: $y = F(x_1,...,x_n)\ exp(-u) \leq F(x_1,...,x_n)$ where $u > 0$. $Exp(-u)$ expresses the level of technical inefficiency of investigated objects. Due to the competition in the banking sector, the author recommends to use the stochastic model with time-varying technical efficiency for panel data (5):

$$\ln y_{it} = \beta_{0t} + \sum_n \beta_n \ln x_{nit} + v_{it} - u_{it}, \tag{5}$$

where

y_{it} – panel data of production (output) volumes;
β_{ot} – frontier intercept (constant);
β_n – vector of of technological parameters;
x_{nit} – panel data of resources (input) volumes;
v_{it} – random error term for panel data;
u_{it} – inefficiency error term for panel data.

The Stochastic Frontier Approach model includes two error components: an error term that captures inefficiency (u_{it}) and a random error (v_{it}). It is impossible to calculate the precise value of inefficiency, because of its composite structure. Due to this, the result of efficiency measurement is the conditional expectation of its value (6):

$$\hat{u}_{it} = E(u_{it} \mid v_{it} - u_{it} = \hat{e}_{it}) \tag{6}$$

where

$\hat{u}_{it}$ – modeled inefficiency error term for panel data;
u_{it} – inefficiency error term for panel data;
v_{it} – random error term for panel data;
$\hat{e}_{it}$ – modeled random error term for panel data without inefficiency error term.

3 The Application of Multistage Approach to the Efficiency Measurement of Latvian Banks

3.1 Methodology of the Research

It is possible to view stochastic frontier regressions as competing with DEA. Carried to an extreme the two approaches, DEA vs. Stochastic Frontier Regressions, can be regarded as mutually exclusive – as in Schmidt (1985). An alternative view is also possible in which the two approaches can be used in complementary fashion. Ferrier and Lovell (1990), for example, use two approaches to cross-check each other. In this approach, the objective is to avoid what Charnes, Cooper and Sueyoshi (1988) refer as „methodological bias". Indeed, going a step further, it is possible to join the two approaches in the multistage methodology of efficiency evaluation. [6, 292-293]
The problem of keeping profitability is especially actual and important in the circumstances of unstable macroeconomical environment. In this connection, there is developed a concept of efficiency measurement of Latvian banks in the research, assuming operational profit to be an output while interest expense, personnel costs and credit impairments are defined as inputs.

The first stage of the performance evaluation will be completed on the basis of DEA CCR approach that allows calculating overall efficiency score and optimal volumes of inputs. The second stage of the research is realized using SFA method that provides a possibility of cross-checking and identification of hidden reserves of development.

3.2 Efficiency Measurement Results of Latvian Banks on the Basis of CCR DEA Approach

The results of banking performance evaluation on the basis of CCR input-oriented model, assuming operating profit as an output, are represented in Figure 1.

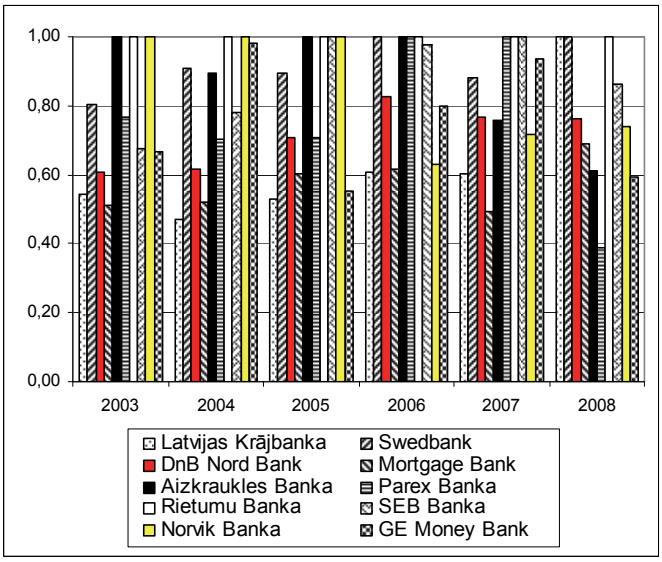


Fig. 1. Dynamics of banking CCR efficiency score, (%)

The application of DEA approach requires the determination of assumptions, concerning orientation measures of the model and the concept of returns to scale (RTS). The banking production process may be aimed either at minimization of resources (input-oriented) or maximization of production volumes (output-oriented). It is emphasized in the international researches that the orientation of the model should be aimed at controllable variables. Usually volumes of resources are considered to be over control of banking management, therefore there is applied the assumption of input orientation in the research. Since the constant returns to scale CRS approach represents the total (overall) efficiency level, CCR DEA model is considered to be the basic concept of the research.

The results of the efficiency measurement approve that the average level of performance of investigated banks has diminished to 76,54% in 2008. The highest overall efficiency (84,58%) was observed in 2006 that is concerned with the increase of crediting activity in the banking sector. The redistribution of leaders' positions among investigated objects is among important trends that are characterizing the dynamics of the efficiency score.

JSC "Rietumu Banka" has demonstrated the best result, operating on the efficiency frontier during all periods of the observation. The long-term stability of the efficiency level of JSC "Rietumu Banka" is indicative of its ability to maximize the volume of output using minimal volumes of inputs and to ensure optimal proportions of output and inputs in the process of production, thus of both 100% technical and scale efficiency in comparison to the set of investigated banks. The maximal overall efficiency of "Rietumu Banka" is confirmed by its successful strategy and

consistent management activities. The target customer group of JSC "Rietumu Banka" consists of legal entities and private customers with high level of income, mainly nonresidents. Other financial indicators justify high efficiency level of JSC "Rietumu Banka": its profit for the nine months of 2010 is 4 million euro, the capital adequacy ratio is 18.17% and liquidity ratio – 53.85%.
To improve the performance of inefficient banks, it is important to determine, which proportions of resources will maximize the overall efficiency level (see Table 1).

Table 1

CCR virtual input volumes in 2003 (thsd. LVL)

	Virtual input (Personnel costs), reduction (%)	**Virtual input (Interest expense), reduction (%)**	**Virtual input (credit impairments), reduction (%)**
Latvijas Krājbanka	3 046,43 ***(45,81%)***	1 596,09 ***(56,72%)***	560,30 ***(45,81%)***
Swedbank	8 164,16 ***(19,53%)***	9 550,00 ***(32,93%)***	3 657,19 ***(21,60%)***
DnB Nord Bank	2 246,81 ***(39,11%)***	1 178,92 ***(59,25%)***	248,43 ***(39,11%)***
Mortgage Bank	2 485,19 ***(48,94%)***	2 907,04 ***(58,43%)***	1 113,26 ***(54,69%)***
Aizkraukles Banka	3 375,00 ***(0,00%)***	1 768,00 ***(0,00%)***	643,00 ***(0,00%)***
Parex Banka	11 394,03 ***(23,23%)***	13 328,13 ***(24,35%)***	5 104,04 ***(57,13%)***
Rietumu Banka	5 308,00 ***(0.00%)***	2 788,00 ***(0.00%)***	319,00 ***(0.00%)***
SEB Banka	7 600,51 ***(32,31%)***	8 890,67 ***(34,99%)***	3 404,70 ***(33,89%)***
Norvik Banka	2 056,00 ***(0,00%)***	2 405,00 ***(0,00%)***	921,00 ***(0,00%)***
GE Money Bank	2 359,88 ***(33,22%)***	1 239,08 ***(45,96%)***	182,97 ***(33,22%)***

JSC "Swedbank" and JSC "Latvijas Krājbanka" have improved their performance during the investigated period significantly. In 2008 both credit institutions were functioning on the efficiency frontier, having 100% CCR efficiency score. Despite of growing interest expenses and impairment losses JSC "Swedbank" and JSC "Latvijas Krājbanka" have succeeded in keeping and increasing volumes of operating revenues. JSC "Aizkraukles Banka" and JSC "Norvik Banka" were fully efficient in 2003 and 2005, but both credit institutions have lost their leading positions by the end of the investigation period.
The DEA approach provides a possibility to calculate the volumes of virtual optimal inputs. According to the data in Table I, the CCR projection requires significant reduction in inputs, especially for JSC „DnB Nord Bank", JSC "Latvijas Krājbanka", JSC „Mortgage Bank" and JSC „Parex Banka". It is necessary to emphasize that interest expense and impairment losses have a stronger impact on the overall efficiency level than personnel costs. The average efficiency in the year 2008 has remained on the same level as in 2003. Nevertheless, the improvement of profitability of two banks could be achieved only after dramatic decrease of input volumes: JSC "Parex Banka" should cut its interest expense and impairment losses by 61,34% and 68,51% respectively in order to operate on the efficiency frontier (see Table 2). However, the optimization of JSC "GE Money Bank" input volumes is concerned with diminishing of its personnel costs by 71,72% and credit impairment costs by 61,96%.

Table 2

CCR virtual input volumes in 2008 (thsd. LVL)

	Virtual input (Personnel costs), reduction (%)	Virtual input (Interest expense), reduction (%)	Virtual input (credit impairments), reduction (%)
Latvijas Krājbanka	11 369,00 ***(0,00%)***	22 260,00 ***(0,00%)***	5 200,00 ***(0,00%)***
Swedbank	34 582,00 ***(0,00%)***	184 127,00 ***(0,00%)***	52 807,00 ***(0,00%)***
DnB Nord Bank	8 814,21 ***(23,60%)***	46 930,03 ***(39,88%)***	13 459,37 ***(34,88%)***
Mortgage Bank	6 605,38 ***(30,83%)***	23 945,69 ***(45,66%)***	6520,30 ***(30,83%)***
Aizkraukles Banka	9 042,89 ***(42,46%)***	18 575,23 ***(38,78%)***	11 437,03 ***(38,78%)***
Parex Banka	19 416,46 ***(61,34%)***	48 615,17 ***(61,34%)***	34 154,96 ***(68,51%)***
Rietumu Banka	12 526,00 ***(0,00%)***	26 515,00 ***(0.00%)***	22 433,00 ***(0.00%)***
SEB Banka	19 229,25 ***(13,76%)***	93 696,12 ***(13,76%)***	30 077,98 ***(16,14%)***
Norvik Banka	7 375,69 ***(25,88%)***	16 022,55 ***(25,88%)***	9 769,03 ***(25,88%)***
GE Money Bank	2 621,86 ***(71,72%)***	5 549,95 ***(40,41%)***	4 695,53 ***(61,96%)***

3.3 Efficiency Measurement of Latvian banks on the Basis of Stochastic Frontier Approach

The first stage of the performance analysis provides information concerning both DEA overall efficiency scores and optimal input volumes. In accordance with the approach of Data Envelopment Analysis, the efficiency frontier is generated from the empirical results of the most efficient Decision Making Units i.e. benchmarks that „float" on the piecewise linear frontier; therefore some of investigated objects are 100% efficient. Nevertheless, it is important to identify hidden efficiency reserves even for fully CCR-efficient DMUs. It is achievable on the basis of Stochastic Frontier Approach technique. Due to this method, the efficiency frontier is constructed using principles of econometrical modeling that allows estimating the performance in comparison to the theoretically developed efficiency frontier.

The application of SFA method requires specifying the functional form of the efficiency estimation model. The author has used the log-linear model specification, as in (5). In accordance with the basic concept of the research, operational profits are presumed to be an output while interest expense, personnel costs and credit impairments are defined as inputs. Due to the composite structure of the model error, there are applied Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE) methods in the efficiency analysis. v_i and u_i are assumed to be normal and half-normal distributed, respectively. The calculations are accomplished using the FRONTIER 4.1. program (see Table 3).

Table 3

Coefficients of the efficiency measurement model on the basis of Stochastic Frontier Approach

Parameters of the model	Coefficient	Standard errors
Beta 0 (constant)	0.28	0.70
Beta 1 (personnel costs)	-0.77	0.12
Beta 2 (interest expense)	0.88	0.69
Beta 3 (impairment losses)	0.72	0.13

On the one hand, it is possible that on the basis of MLE algorithm due to distributional assumptions concerning composite error terms there is calculated a local extremum of the log-likelihood function. On the other hand, the results of efficiency measurement correspond to the economical intuition: after optimization of input volumes 100% CCR-efficient banks proved to be 83-86% efficient on the basis of SFA method, demonstrating minimal volatility of the efficiency score during the period of the research (see Figure 2).

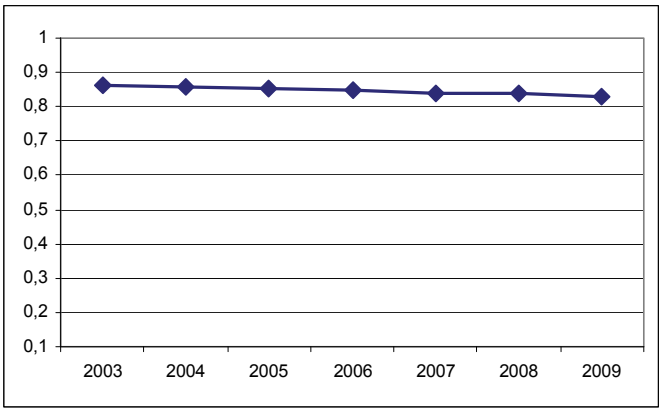


Fig. 2. The average technical efficiency score on the basis of Stochastic Frontier Approach, (%)

Conclusions

The scientific paper is devoted to the performance measurement problem of Latvian credit institutions. The standard methods of efficiency measurement, such as quantitative ratio analysis, ratings and regression analysis do not provide the possibility of multidimensional efficiency evaluation. The methodology of frontier methods is considered to be a sophisticated tool for efficiency measurement that allows the investigation of complex production processes among a set of Decision Making Units (DMUs). According to the information that is available to the author, Latvian banks to the efficiency measurement currently do not apply methods of frontier data analysis.

The author has implemented the multistage approach, analyzing efficiency scores of a set of Latvian banks using CCR DEA and SFA techniques. JSC "Rietumu Banka" has demonstrated the best result, operating on the efficiency frontier during all periods of the observation. There were calculated volumes of weighted optimal inputs that provide the possibility to maximize the overall efficiency score, using DEA technique. Using the Stochastic Frontier Approach, it was stated that

even after optimization of input volumes there is still a possibility to improve the banking performance by 17% in the year 2009.
According to the obtained results of the research, the author recommends to improve the methodology of efficiency measurement of Latvian banks on the basis of multistage application of frontier analysis methods.

References

[1] Assessing Bank and Bank Branch Performance: Modeling Considerations and Approaches. Paradi J.C., Vela S., Yang Z. Handbook on Data Envelopment Analysis. – Boston: Kluwer, 2004, pp. 349-400.

[2] Data Envelopment Analysis: History, Models and Interpretations. Cooper W.W., Seiford L.M., J. Zhu. Handbook on Data Envelopment Analysis. – Boston: Kluwer, 2004, pp.1-39.

[3] Returns to Scale in DEA. Banker R.D., Cooper W.W., Seiford L.M., J. Zhu. Handbook on Data Envelopment Analysis. – Boston: Kluwer, 2004, pp. 41-74.

[4] FARRELL, M. J. (1957): The Measurement of Productive Efficiency, Journal of the Royal Statistical Society, Vol. 120, pp. 253-281.

[5] CHARNES, A., COOPER, W., LEWIN, A.Y., and SEIFORD, L.M. (eds.). Data Envelopment Analysis - Theory, Methodology and Applications. – Dordrecht: Kluwer, 1997, 513 pp.

[6] COOPER, W.W., SEIFORD, L.M., and TONE, K. (2000). Data Envelopment Analysis - A Comprehensive Text with Models, Applications, References and DEA-Solver Software. – New York: Springer, 2007, 490 pp.

[7] FARE, R., GROSSKOPF, S., and LOVELL, C.A.K. Production Frontiers. – Cambridge: University of Cambridge Press, 1994, 296 pp.

[8] T. ARSHINOVA, "The Application of Data Envelopment Analysis Approach to the Efficiency Measurement of Latvian Banks", *Scientific Journal of Riga Technical University*, Computer Science, Vol.5, no. 39, p.92-104, 2009

Current address

Tatyana Arshinova, Mg.oec.
Department of Probability Theory and Mathematical Statistics,
Faculty of Computer Science and Information Technologies,
Riga Technical University, 1/4 Meza Street, Riga, LV-1048
Phone: (+371)29855850
e-mail: tatjana.arsinova@rtu.lv

Journal of Applied Mathematics

volume IV (2011), number III

DETECTION AND CORRECTION OF CALENDAR EFFECTS: AN APPLICATION TO INDUSTRIAL PRODUCTION INDEX OF ÁLAVA

ARTECHE Josu, (E), MAJOVSKÁ Renata, (CZ), MARIEL Petr, (E), ORBE Susan, (E)

Abstract. Calendar effects concern special days such as Christmas and Eastern and usually are associated with economic activity fluctuations. In analysis in which time series are seasonally adjusted it is necessary to detect and correct these calendar effects using suitable procedures. This article compares different methods of processing of these effects using spectral analysis, recursive estimation, information criteria and t -statistics. The proposed procedure is applied to time series of Industrial Production Index of the Spanish province of Álava. Results indicate that simpler models can achieve better results, but preclude the identification of particular calendar effects. Furthermore, adjustment of classical calendar effects by simple models may be insufficient in some time series, because their duration and impact may vary between countries and sectors.

Key words. Calendar effect, TRAMO, time series, seasonal adjustment, spectral analysis, recursive estimation, information criteria

Mathematics Subject Classification: Primary 62M10, 91B84; Secondary 91B82.

1 Introduction

Time series of many economic variables are significantly influenced by various factors related to the calendar. These factors include non-working days, leap years, holidays etc. There is no general definition or uniform procedure for processing of these events, which are known as calendar effects. Calendar effects can be divided into two groups. The first group includes the effects of working (or trading) days and the second group deals with special calendar effects, such as Christmas, Easter or holidays. These effects have to be taken into account during the data processing and various methods of seasonal adjustment related with these effects must be applied.

In this paper we compare different methods of correction of the calendar effects using various statistical procedures which are applied to time series of Industrial Production Index (IPI) of the Spanish province of Álava (Basque). They are based heavily on free software package TSW which

consists of two parts - TRAMO and SEATS. This software is used for seasonal adjustment of time series by National Statistical Institutes (NSI) in many European countries.
This article is organized as follows. In Section 2 different methodological approaches for the adjustment of calendar effects are described and in Section 3 alternative models for adjustment of time series are presented. In Section 4 we describe obtained results and in Section 5 these results are summarized.

2 Methodology

There are many different methods for correction of calendar effects. Our analysis is focused on the linear regression method because Eurostat recommends European NSI to use this method. Corrections of working (or trading) days are derived from the estimation of the linear regression. Therefore the calendar structure can be modelled using some explanatory variables according to the following model:

$$\begin{aligned} y_t &= x_t'\beta + v_t, \\ \varphi(L)\delta(L)v_t &= \theta(L)\varepsilon_t \quad t = 1, 2, \ldots, T, \end{aligned} \tag{1}$$

where y_t is the observed time series, v_t are error terms which follow an ARIMA process, $\varphi(L)$, $\delta(L)$ and $\theta(L)$ are finite polynomials of the lag operator[1] L, x_t is a vector $(K \times 1)$ of K relevant explanatory variables, β is a vector $(K \times 1)$ of unknown parameters and ε_t is an error term defined as white noise.
Component $x_t'\beta$ represents nonstochastic effects which are subtracted from the original series before applying the ARIMA methodology for decomposition of the time series into trend/cycle, seasonal and irregular component. The simplest nonstochastic effect, which is one of many that are subtracted from the original series, is the average (regression constant). More complex effects are for example the intervention variables, atypical observations or calendar effects.
Each time period (month, quarter, etc.) is characterized by different number of Mondays, Tuesdays, Sundays, therefore economic activity can be affected by this fact. For example, one working day guarantees a certain level of production of an enterprise, but lower income for firms related to tourism.
The working days effect distinguishes working days from weekend days, and that is why special variable (we_t) is usually used in the literature to express the weighted difference between the number of working days (w_t) and non-working days (nw_t) during the period t. This is defined as

$$we_t = \left(w_t - \frac{5}{2} nw_t \right), \tag{2}$$

where the number of non-working days is multiplied by 5/2 so that the average of the newly created variable was zero. The coefficient of the variable we_t includes the effect of additional working days in the period t (month or quarter).

[1] Polynomial $\delta(L)$ includes unit roots of regular and seasonal differences, $\varphi(L)$ represents the stationary autoregressive component and $\theta(L)$ is an invertible moving average polynomial.

Sometimes it is necessary to include the effect of trading days in the model which can be defined by the following six regressors:

$$w_t^1 = (Mon_t - Sun_t),\ w_t^2 = (Thu_t - Sun_t),\ \ldots,\ w_t^6 = (Sat_t - Sun_t), \tag{3}$$

where $Mon_t, Thu_t, \ldots, Sun_t$ is the number of Mondays, Tuesdays, ..., Sundays during the period t (month, or quarter).
Eurostat and the European Central Bank also emphasize the importance of leap year correction in their recommendations. This type of periodicity can be modelled using the following zero mean variable:

$$ly_t = \begin{array}{ll} 0.75 & \text{if } t = \text{February of leap year} \\ -0.25 & \text{if } t = \text{February of non-leap year} \\ 0.75 & \text{if } t = \text{other month as February} \end{array} \tag{4}$$

3 Alternatives of the correction of the calendar effect

In this section we present five models which allow for different corrections of the calendar effects. The model called Alternative 0 is the basic alternative because it is the most similar to the model which some European NSI use for the correction of the calendar effect. Other Alternatives generalize it by adding new variables for other possible effects (Alternative 1 and 2) or simplify it (Alternative 3 and 4) pursuing goodness of fit of the model and prediction power.

Alternative 0

$$Y_t = \beta_1 + \beta_2 PublicHolidaysCent_t + \beta_3 we_t + \beta_4 ly_t + Error_t^{ARIMA} \tag{5}$$

The variable we_t is defined in (2), ly_t v (4) and the variable $PublicHolidaysCent_t$ is defined as

$$PublicHolidaysCent_t = PublicHolidays_t - DF\ .$$

The variable $PublicHolidays_t$ is the number of non-working days corresponding to Monday, Tuesday, ..., Friday in the month t and DF is the long run average of non-working days in the month excluding weekends. Empirically, DF is not known. In this application the average of the variable $PublicHolidays_t$ is used, which approaches the unknown value DF if the number of observations is large enough.

Alternative 1

$$Y_t = \beta_1 + \beta_2 PublicHolidaysCent_t + \sum_{i=1}^{6} \beta_{i+2} w_t^i + Error_t^{ARIMA} \tag{6}$$

The variable w_t^i is defined in (3). It means that this model analyzes the effect of trading days.

Alternative 2

$$Y_t = \beta_1 + \beta_2 PublicHolidaysCent_t + \beta_3 ly_t + \sum_{i=1}^{6} \beta_{i+3} w_t^i + Error_t^{ARIMA} \quad (7)$$

The only difference of this model and the model defined in (6) is that it includes the leap year correction with the variable ly_t defined in (4).

Alternative 3

$$Y_t = \beta_1 + \beta_2 WorkingDaysSinePH_t + Error_t^{ARIMA} \quad (8)$$

This model looks for parsimony (simplicity), because it uses only one explanatory variable $WorkingDaysSinePH$, which is defined as

$$WorkingDaysSinePH_t = WorkingDays_t - PublicHolidays_t - DL.$$

The variable $WorkingDays_t$ includes the number of working days and the variable $PublicHolidays$ indicates the number of non-working days corresponding to Monday, Tuesday, ..., Friday. Variable $WorkingDaysSinePH_t$ is centred by DL which stands for the average number of working days in period t. Similar to DF the value DL is unknown. The average value of $WorkingDays_t - PublicHolidays_t$ is used in this application. Variable $WorkingDaysSinePH$ represents the effect of working days, trading days, Easter effect and implicitly the effect of leap year, higher parsimony is thereby obtained.

Alternative 4

$$Y_t = \beta_1 + \beta_2 CalendarEffect_t + Error_t^{ARIMA} \quad (9)$$

This is also a parsimonious alternative, because variable $CalendarEffect_t$ is defined as

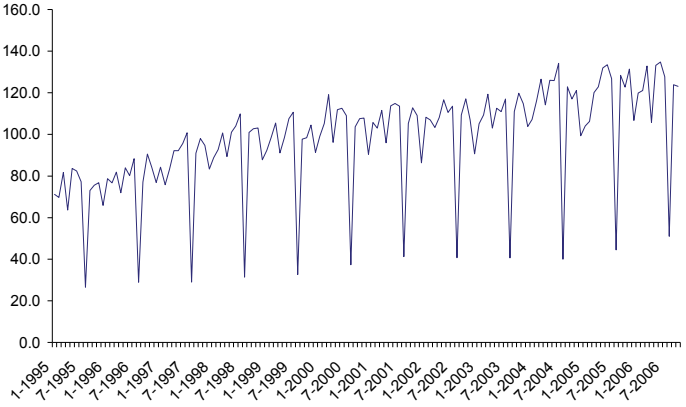
$$CalendarEffect_t = \left(Workingdays_t - PublicHolidays_t\right) - RDF\left(Weekends_t + PublicHolidays_t\right)$$

where the variable $Weekends_t$ is the number of Saturdays and Sundays during the period t. This regressor is similar to the variable we_t defined in (2) but it includes the effect of non-working days which are not weekends. The constant $5/2$ was used for centering of variable we_t. For variable $CalendarEffect$ the constant RDF is used, which is, in a long-term view, the ratio between working days and non-working days. In practice, this ratio is estimated by observed values, in our case 4.77/2.22. In this case, all the calendar effects are included in one variable $CalendarEffect$ and that is why it is not possible to identify effects of weekends, working days and leap year. Nevertheless, parsimony is achieved and all calendar effects can be modelled correctly.

4 Application: IPP of Álava

Models described in the previous section are estimated by the software package TSW using monthly data from 01/1995 to 10/2006 of the Industrial Production Index of the Spanish province of Álava. The original time series is presented in the Figure 1.

Figure 1: Industrial Production Index of the Spanish province Álava



The comparison of the estimated models is done by spectral analysis, recursive estimation, information criteria and t-statistics of the explanatory variables defined in (1). Monthly time series are characterized by pseudo-cyclic behaviour which is partly caused by the calendar effects. Cleveland a Devlin in [2] proposed different techniques of detection of these pseudo-cycles in monthly (quarterly) time series occasioned by the weekly cycle. While four cycles are expected to complete their period over a month, a component with fractional frequency (0.348) can still be presented. A second important frequency at 0.432 is also related with the calendar effect (see [2]). These frequencies are called calendar frequencies and can be detected in the periodogram if it shows peaks at both of them. In [9] a method of automatic detection of the working/trading days effect is analyzed according to some criterion based on the amplitude of the estimated spectrum at the calendar frequencies. This effect exists, if the spectrum crosses the limit of $6k$ with respect to the adjacent frequencies, where k is the difference between the maximum and minimum value of the spectrum divided by 52. This criterion is also used in the standard procedure X12-ARIMA.
As in [9] a consistent estimation of the spectral density is obtained by fitting an autoregressive model of high order that can capture the inertia of the analyzed time series. That is why an AR(50) model is used in the application presented bellow. The estimation of the spectral density function is carried out in a transformed series as the original series usually contains trend and seasonal components which may blur the information at the calendar frequencies (leakage). Therefore the use of the residuals is recommended in [2] and [8], which is the procedure used in this paper. We calculate and display the estimated spectrum of the residuals of the ARIMA decomposition performed by SEATS. The estimated spectrum is evaluated on 61 frequencies $j/120$, $j = 0, 1, \ldots, 60$, which allows to take into account the seasonal frequencies ($k/12, k = 1, 2, \ldots, 6$). The frequencies $j/120$ closer to the calendar frequencies 0.348, 0.432, and their adjacent frequencies are replaced by 0.348, 0.432 and $0.348 \pm 1/120$, $0.432 \pm 1/120$ respectively.
Figure 2 shows the estimated spectral densities of the random components and residuals obtained by SEATS for series not treated for calendar effect (Figures a)), and for series treated according to the five alternatives defined in (5) - (9)) above and presented in (Figures b) - f)). The vertical lines

represent the seasonal (dotted lines) and calendar frequencies (full lines). Calendar effects can be clearly identified in the non-treated series but it is reduced with all the proposed treatments.

Figure 2: Estimated spectral densities

Peridogram of random component Periodogram of residuals

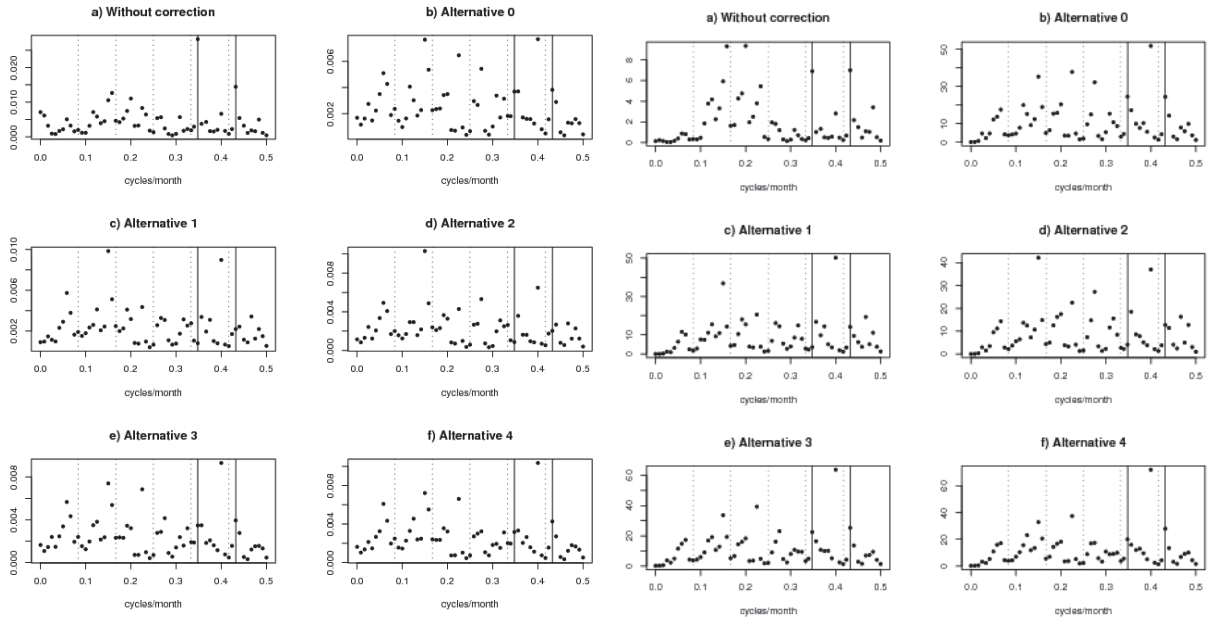


Table 1 shows the differences between the estimated spectrum value at the calendar frequencies and their neighbour frequencies minus $6k$, all expressed in k units ($\left[\hat{f}(s)-\hat{f}(s\pm 1/120)-6k\right]/k$, $s=0{,}348,\ \ 0{,}432$ for $k=\left\lfloor \max_l\left(\hat{f}(l)\right)-\min_l\left(\hat{f}(l)\right)\right\rfloor/52$), where l stands for the 61 frequencies considered in the analysis. Dividing by k standardizes the results and makes the comparison easier. If the differences with both adjacent frequencies have the same sign, we keep the results corresponding to the frequency with a spectral amplitude more similar to the one at the calendar frequency. If the signs are different, we show the results corresponding to the negative sign, since the existence of calendar effects requires both values to be positive. A positive value in the table therefore implies existence of calendar effect and an insufficient adjusting of the calendar effect.

Table 1 and Figure 2 show that Alternatives 1 and 2 capture the weekly cycle well but Alternatives 0, 3 and 4 do not adjust all this calendar effect. This is an expected result as the Alternative 1 and 2 include more variables which allow a higher degree of flexibility.

Recursive method for detection of the calendar effect was designed in [8]. It compares various alternatives using out of sample forecast errors. Let $Y_{t+h|t}$, for each t in $T_0 \le t \le T-h$, denote the forecast of Y_{t+h} obtained by estimating the ARIMA model and using this estimated model to forecast h steps from time t. The out of sample forecast error is $e_{t+h|t}=Y_{t+h}-Y_{t+h|t}$. The following sequence of accumulated residual sums of squares is therefore analyzed:

$$SS_{h,M}=\sum_{t=T_0}^{M} e_{t+h|t}^2 \quad M=T_0,T_1,\ldots,T-h\,. \tag{10}$$

Suppose there are two alternatives (Model 1 and Model 2) with forecast errors $e_{t+h|t}^{(1)}$ and $e_{t+h|t}^{(2)}$ and with accumulated sums $SS_{h,M}^{(1)}$ and $SS_{h,M}^{(2)}$. Then for the sake of comparison the normalized version of the following differences $SS_{h,M}^{(1)} - SS_{h,M}^{(2)}$ are plotted:

$$SS_{h,M}^{1,2} = \frac{SS_{h,M}^{(1)} - SS_{h,M}^{(2)}}{SS_{T-h}^{(2)} / \left(T - h - T_0\right)}, \quad \text{pro} \quad T_0 \le M \le T - h. \tag{11}$$

Since $SS_{h,M}^{i,i} = 0$ for $i = 1, 2$ the Model 1 is discarded in the case when the sequence $SS_{h,M}^{1,2}$ is increasing, as in this case its residual sums of squares are smaller.

According to the procedure outlined in [8] we analyze the five alternative models proposed. As a first step the value T_0 (number of observations in the first estimation) and h (the number of forecasts) is determined. In [9] an ideal value $T_0 = 61$ is recommended for monthly data of ten years period. This means that the first estimation uses monthly observations of the first five years. In this paper we use the same initial value $T_0 = 61$ because the number of observations of the series is similar.

Regarding the parameter h, values $h = 1$ and $h = 12$ are proposed in [9]. If $h = 1$, the values of the next month are forecasted, given observations up to the previous month. On the other hand, if $h = 12$ the values of the same month next year are forecasted. In this work, both horizons of prediction are considered but the conclusions in both cases are very similar and that is why only the results for $h = 1$ are presented in the figures.

The conclusions of recursive estimation are summarized in Table 1. The base model (Model 1) is at the top of the table (bold letters) and the rival models are in the first column of Table 1. Abbreviation *Sup* stands for superior, and in this case the base model is superior to the rival model, having lower accumulated sums (10). The opposite case is denoted as *NonSup* which stands for non superior. Alternative 4 shows the best accumulated sums of squares and Alternatives 1 and 2 have the worst performance.

It is important to note that the period (01/2000-06/2001) shows different results comparing to the remaining periods. This is probably due to the atypical calendar effect caused by the transition from the Spanish Peseta to the Euro currency in the period 1999-2002.

Table 1 also shows the information criteria BIC and AICC [1], [9], corresponding to estimations of the five analysed alternatives. The lowest value indicates the best model in terms of the best fit taking into account degrees of freedom. The smallest value of AICC and BIC corresponds to Alternative 3. This result is not surprising, since both indicators penalize the number of parameters of the estimated model. Alternatives 3 and 4 include all the possible calendar effects into one variable creating parsimonious models which are able to model the observed data adequately. However, the results of these indicators should be supplemented by the conclusions of the previous paragraphs, because the model selection may be dependent on the final objectives of the analysis and not only on the parsimony of the model.

Regarding the t-statistics, only variables describing the effects of working days (Monday - Friday) are not statistically significant at the 5% significance level. The effect of Saturday and the effect of the leap year are considered significant at the 10% level. Generally we can say that the alternatives that include calendar effects only into one (Alternative 3 and 4) or three (Alternative 0) variables

present these effects as statistically significant. Using multiple variables for capturing of trading days leads to insignificant variables.

Table 1. **Comparison of alternative models for time series of IPI Álava**

IPP ÁLAVA					
	Alternative				
	0	**1**	**2**	**3**	**4**
Spectral analysis					
0.348 cycle/month					
Random components	1,39	-4.69	-3,73	-1.04	-2,93
Residuals	-6,09	-7.37	-6,99	-6.08	-6.76
0,432 cycle/month					
Random components	4.18	-1.15	-4.32	3.56	4.37
Residuals	0.6	-3.33	-4.52	0.81	2.01
Recursive estimation					
Alternative 0	-	NonSup **	NonSup	Sup	Sup
Alternative 1	Sup **	-	NonSup	Sup **	Sup **
Alternative 2	Sup **	Sup	-	Sup **	Sup **
Alternative 3	NonSup	NonSup **	NonSup	-	Sup
Alternative 4	NonSup	NonSup **	NonSup	NonSup	-
Information criteria					
AICC	-813.14	-810.46	-810.76	**-816.29**	-813.91
BIC	-6.14	-6.06	-6.02	**-6.20**	-6.18
***t* -statistics**					
Public holidays	-5.90	-6.60	-6.51		
Working days	7.79				
Leap year	1.95		1.89		
Monday		-1.45	-1.16		
Tuesday		2.28	1.99		
Wednesday		1.19	1.43		
Thursday		0.96	0.68		
Friday		1.52	1.63		
Saturday		-1.87	-1.72		
Working days sine PH				10.55	
Calendar effect					10.33

** Difference between two alternatives is significant as of 06/2001.

5 Conclusions

The detection and correction of the calendar effects in the monthly and quarterly economic time series are necessary for correct interpretation of observed data as reported by Eurostat in its recommendations [3], [4], [5], [6]. However, there is several methods for the detection and modelling of these effects. In this paper we present five possible alternatives of modelling of calendar effects. Their comparison is made by the empirical application focused on the monthly

time series of IPI of Álava using four methodologies: spectral analysis, recursive estimation, information criteria and t-statistics.

We conclude that only Alternative 4 is uniformly superior to the other alternatives in terms of recursive estimation. On the other hand, alternatives that include the effect of working days (Alternative 1 and 2) capture the weekly cycle in a more comprehensive way than other alternatives, but they are overcome by Alternatives 3 and 4, which includes all the calendar effects into one variable, if forecast and information criteria are used. Regarding Alternative 0, spectral analysis shows that this alternative does not capture the weekly cycle completely as observed also in Alternative 3 and 4. Alternative 0 is the base model, which is most similar to the model that is likely to be used by NSI.

This analysis indicates that the detection and correction of the calendar effect should consider other factors which are not included in the models of the five analysed alternatives. Firstly, the effect of Easter should be studied in detail, since the subsumption of Easter to public holidays may be inadequate. It means that the effect of Easter may be assigned to the public holiday effect but its effect is probably more complicated. It should be taken into account that e.g. in some provinces of Spain Thursday and Friday before Easter Monday are public holidays and many people take advantage of it for organizing their spring holidays. Secondly, it is necessary to focus on the atypical observations which can have important impact on the processing of calendar effects.

Acknowledgement

The authors gratefully acknowledge the Basque Statistical Institute /Euskal Estatistika Erakundea (Eustat) for providing support and data needed. The authors are also grateful to the Department of Education of the Basque Government through grant IT-334-07 (UPV/EHU Econometrics Research Group) and from Spanish Ministerio de Educación and FEDER (SEJ2007-61362).

References

[1.] BÓGALO, J.: *Apuntes del curso: Identificación y estimación del modelo ARIMA en TSW*. La Escuela de Estadística de las Administraciones Públicas, Instituo Nacional de Estadística, Madrid, 2006.

[2.] CLEVELAND, W., DEVLIN, S.: *Calendar Effects in Monthly Time Series: Detection by Spectrum Analysis and Graphical Methods*. In Journal of the American Statistical Association, Vol. 75, pp. 487-496, 1980.

[3.] Di PALMA, F., MARINI, M.: *The Working/Trading Day Adjustment of Italian Quarterly National Accounts: Methodology and Presentation of the Main Result*. In Proceeding of Joint UNECE/Eurostat/OECD Meeting on National Accounts, Geneva, CES/AC.68/2004/12, 2004.

[4.] Eurostat: *Follow-up of the CMFB Task Force on Seasonal Adjustment of Quarterly National Accounts*. Eurostat Unit B2, Eurostat B1-B2/CN 514, 2002a.

[5.] Eurostat: *Methodology of Short-term Business Statistics- Interpretation and Guidelines.* Eurostat Commission, Theme 4, Industry Trade and Services, 2002b.

[6.] Eurostat: *Euro-Indicators Newsletter*. LN-122003-EN, 2003.

[7.] MARAVALL, A.: *Notes on Programs TRAMO and SEATS*. Part III, Signal Extraction in ARIMA Times Series, Banco de España, Madrid, 2003.

[8.] SOUKUP, R. J., FINDLEY, D. F.: *On the Spectrum Diagnostics Used by X12-ARIMA to Indicate the Presence of Trading Day Effects after Modeling or Adjustment*. In Proceeding of American Statistical Association, Business and Economic Statistics Section, Washington, 1999. Available at http://www.census.gov/ts/papers/rr9903s.pdf.
[9.] SOUKUP, R. J., FINDLEY, D. F.: *Detection and Modelling of Trading Day Effects*. In Proceeding of International Conference on Evolvable Systems (ICES), Edinburg, 2000.

Current address

Josu Arteche, PhD.
University of the Bask Country/Euskal Herriko Unibertsitatea
Department of Applied Economics III (Econometrics and Statistics)
Lehendakari Aguirre 83, E48015 BILBAO, Spain
Tel. +34946013852
e-mail: josu.arteche@ehu.es

Renata Majovská, PhD.
VSB - Technical University of Ostrava (Faculty of Economics)
Department of Mathematical Methods in Economics
Sokolska 33, 701 21 Ostrava, Czech Republic
Tel. +420597322506
e-mail: renata.majovska@vsb.cz

Petr Mariel, PhD.
University of the Bask Country/Euskal Herriko Unibertsitatea
Department of Applied Economics III (Econometrics and Statistics)
Lehendakari Aguirre 83, E48015 BILBAO, Spain
Tel. +34946013848
e-mail: petr.mariel@ehu.es

Susan Orbe, PhD.
University of the Bask Country/Euskal Herriko Unibertsitatea
Department of Applied Economics III (Econometrics and Statistics)
Lehendakari Aguirre 83, E48015 BILBAO, Spain
Tel. +34946013845
e-mail: susan.orbe@ehu.es

Journal of Applied Mathematics

volume IV (2011), number III

DEPENDENCE OF EXPENDITURES OF THE CZECH HOUSEHOLDS ON FINANCIAL POWER

BARTOŠOVÁ Jitka, (CZ), BÍNA Vladislav, (CZ)

Abstract. This contribution aims at the exploration of connection between the financial power of households and level and structure of their expenditures. During the research we stem from the assumption that total expenditures and their structure can in different income categories differ. For the spitting of households into several categories according to their financial power we use the poverty threshold and its multiples.

Key words. Equalized household incomes, poverty, similarity of structures, SRÚ 2008.

Mathematics Subject Classification: Primary 62P20; Secondary 62J02

1 Introduction

The transition process and successive process of economic globalization within the frame of EU lead into many economic, technical, political and legislative changes which significantly (both positively and negatively) influenced the structure of Czech and Slovak economy and thus the financial potential, poverty or prosperity of inhabitants in both countries (Bartošová, 2009). The most important factor which can lead to significant social problems is the endangerment of fraction of inhabitants with the poverty. According to the estimates of the World Bank in 2007 one quarter of inhabitants of developing countries were considered as poor.

Poverty and limited financial potential is generally one of the basic factors influencing the consumption of households. According to the fact that households lying beneath the poverty threshold have limited disposable financial resources should they consumption behavior differ from the behavior of households relatively wealthier (see Čermáková, 2001). It means that the structure of expenditures should differ according to the level of financial potential which can be considered as relatively very low, low, medium or high. Especially in the expenditure structure of the poorest and wealthiest group significant differences can be expected.

This contribution is a subsequent article in a series of analyses of financial potential, social situation and poverty in the Czech Republic and in Slovakia which were carried out in last years in the framework of projects GAČR 402/09/0515, VEGA 1/4586/07 and VEGA 1/0370/08. Specifically, in articles concerning the comparison of income differentiation of inhabitants in Czech Republic and Slovakia (Bartošová, Stankovičová, 2009), the determination of subjective poverty measure in Czech Republic and Slovakia (Stankovičová, 2009, Labudová and Sipková, 2008), modeling the influence of different factors on the risk-of-poverty rate for Czech and Slovak households (Pastorek, Stankovičová, 2009), measuring the risk, depth and harshness of poverty in regions of Slovakia (Bartošová and Forbelská, 1010, Stankovičová, 2010, Sipková and Sipko, 2010, Želinský, 2010) etc.

2 Methodology

2.1 The data base

Information about Czech and Slovak households are acquired from the sample survey SRÚ (Household Budget Survey) which is organized annually by the Czech Statistic Office. The basic notion of this survey is so called household unit. This notion is defined as a voluntary declaration of the persons living in chosen flat who live and manage their household (pay for food, living, etc.).

The income of household adjusted on one consuming unit (SJ) will serve as an indicator of poverty. The size of household (number of consuming units) is according to the definition of EU given by EJ = 1 + 0,3 younger children (0 – 13 years) + 0,5 other persons in household (without the head).

The sample files arising from SRÚ survey cannot be considered as representative. For recalculation of sample data to the whole population frequency variable PKOEF is used. It is constructed using minimization of difference between estimated and recomputed sample characteristics. For the construction of model the newest available data are used, the data from SRÚ survey 2008. This data file contains information about incomes and expenditures of Czech households.

2.2 Poverty measurements in EU

The poverty is considered as serious problem not only in countries of the third world but also in developed states - like EU. Especially in the present time the economic development is endangered by crisis there arises a question of fight against the poverty. For the necessary monitoring and comparison of poverty in different countries the quantitative formulation of its measure is inevitable (see Morduch, 2005). The quantification of poverty can be approached from two different perspectives. The first one is consideration of minimal level of incomes the second assess the minimal level of expenditures to provide the basic needs of inhabitants (more details e.g. Želinský, 2010).

For comparison of the poverty in developed countries (and thus also in EU countries) the most commonly used indicator is the risk-of-poverty rate which is defined as a percentage of persons with equivalent disposable income under the poverty threshold (see Ravallion, 1998). The poverty threshold is determined in each country separately and is defined by OECD and EU as 60 % of national median equalized household income. This relative poverty measure ranks among the

additive Foster-Greer-Thorbecke metrics (see e.g. Želinský, 2009) and will be used in this contribution.

2.3 Coefficient of structure dissimilarity

For exploration and quantitative expression of structural changes in the consumption of households which occur under the poverty threshold the measures of similarity (respectively dissimilarity) of structures is used (see e.g. Kahounová, 1994). Two structures with relatively expressed components are considered as same if

$$p_{1k} = p_{2k}\,,\ k = 1,2,...,m$$

where p_{1k} is the proportion of kth component within the total of first structure, p_{2k} is the proportion of kth component within the total of second structure and m is the number of structure components.

For characterization of similarity (resp. dissimilarity) of two structures $\vec{p}_1 = (p_{11}, p_{12},..., p_{1m})$ and $\vec{p}_2 = (p_{21}, p_{22},..., p_{2m})$ different coefficients can be used. Usually, they are constructed on the basis of measuring the distance of two points in m-dimensional Euclidean space. For evaluation of the similarity it is convenient to used normalized measures with values in interval $\langle 0,1 \rangle$. In this contribution two different measures are used – Gatev coefficient of structure dissimilarity and cosine coefficient of structure similarity.

Gatev coefficient of dissimilarity of structures measures absolute and relative changes of structures in their mutual conjunction. It is an integral coefficient of structural changes given by formula

$$k_1(\vec{p}_1, \vec{p}_2) = \sqrt{1 - \frac{2\sum_{k=1}^{m} p_{1k} p_{2k}}{\sum_{k=1}^{m} (p_{1k}^2 + p_{2k}^2)}}$$

The values of this measure lay in the interval where the lower bound represents complete identity and the upper bound dissimilarity.

Cosine coefficient of structure similarity uses for the measurement cosine of the angle of two structures $\vec{p}_1$ and $\vec{p}_2$ and is given by formula

$$k_2(\vec{p}_1, \vec{p}_2) = \frac{\sum_{k=1}^{m} p_{1k} p_{2k}}{\sqrt{\sum_{k=1}^{m} p_{1k}^2 \sum_{k=1}^{m} p_{2k}^2}}$$

in case of identity it takes the value of 1, in case of complete dissimilarity it takes value of 0.

3 Results

3.1 Classification of households according to their financial power

According to the information of EUROSTAT the population of the Czech Republic and Slovakia belongs to the least endangered by the monetary poverty in EU. For example in 2007 EUROSTAT alleged that the risk of poverty endangers some 16 % of EU population. But the Czech Republic was at that time evaluated as a country with the lowest risk-of-poverty rate (9.8 %) and also Slovakia (with 10.5 % of relatively poor) ranked among the least endangered countries of EU.

But such relative measure more than the real poverty exhibits the measure of income differentiation and at present time – when EU countries are faced to the economic crisis – this situation can dramatically alter. The lowest endangering of monetary poverty indicates that incomes of the Czech households are still not markedly differentiated (in comparison with other countries); namely in the area of low incomes, i.e. in the left part of the income distribution. It means that the group of households with "low" financial potential (households under the poverty threshold) is in the Czech Republic very close to the group of households with slightly higher incomes, which can be considered as "medium-low". Such households are also partially endangered by monetary poverty since in time of economic crisis they can easily move under the poverty threshold. Financial potential of other, the most numerous group of households from the centre of income distribution can be regarded as "medium-high" and finally, the financial potential of households from the right-hand side of the income distribution can be considered as "high".

As far as we know, there does not exist, a criterion for division of households into groups according to the financial potential. Before categorization of data into the above mentioned four groups it is necessary to choose suitable criterion for the division. For the definition of category boundaries various quantile or momentum characteristics can be used. According to the character of income distribution (non-uniformity of density, skewness and outliers on the right-hand side) it is advisable to choose robust characteristics. It can be:

- basic quantile characteristics (lover quartile, median and upper quartile),
- other well known and frequently used quantile characteristics (20 % quantile, median and 80 % quantile).
- poverty threshold according to the EU definition, i.e. 60 % of median (poverty threshold, 1.5 multiple of poverty threshold and 2.5 multiple of poverty threshold) etc.

In this contribution we will use the third possibility – categorization using the poverty threshold and its multiples. We consider this choice as the best since it stems from the definition of quantity which is in EU recognize as a criterion for the determination of monetary poverty and which in some sense detects the "insufficiency" of financial power of individual. The corresponding boundaries are shown in Table 1.

Table 1 contains also information concerning the count of households in each above defined category of financial power and their percentage in the sample of all households. The values stem from data file of Household Budget Survey (SRÚ) in 2008 thus it may differ from officially reported from sample survey EU – SILC.

Table 1. Division of the Czech households according to their financial power (SRÚ 2008).

Financial Power		Lower boundary	Upper boundary	Number of households	Percentage
1	**Low**	minimum $x_{\min}$	poverty threshold $0.6\widetilde{x}$	202	6.88%
2	**Medium-low**	poverty threshold $0.6\widetilde{x}$	1.5 poverty threshold $0.9\widetilde{x}$	756	25.77%
3	**Medium-high**	1.5 poverty threshold $0.9\widetilde{x}$	2.5 poverty threshold $1.5\widetilde{x}$	1446	49.28%
4	**High**	2.5 poverty threshold $1.5\widetilde{x}$	maximum $x_{\max}$	530	18.06%
Total				2934	100%

It can be seen that under the official poverty threshold (in the area of low financial power) lay only 202 (i.e. 6.88 %) from total 2934 surveyed households. But other 756 (nearly a quarter of surveyed households) have after recalculation on consuming unit their incomes under median; they do not exceed 90 % of median (medium-low financial power). The most abundant category (nearly one half of surveyed households) is formed by households with medium-high financial power (equalized incomes under 1.5 multiple of median. Even the last category of households with relatively high financial power on consuming unit is not small – it consist of nearly one fifth (18.06%) of surveyed households.

3.2 Differentiation in the Czech household's total expenditures

Now we will explore how the financial potential affects the structure of household expenditures. The differences in consumption behavior should theoretically grow simultaneously with the growth of the difference in financial potential of households. Especially in the expenditure structure of the poorest and wealthiest households significant differences can be expected.

Table 2. Chosen characteristics of net monthly incomes of the Czech households Recomputed on the consumption unit (SRÚ 2008).

Financial power								
Type of characteristic	**low**	**medium low**	**medium high**	**high**	**low**	**medium low**	**medium high**	**high**
Net monthly incomes (CZK per consuming unit)					**Monthly expenditures / Monthly incomes**			
Minimum	2698	8839	13260	22100	38.93%	32.86%	32.15%	8.90%
Median	**8395**	**17480**	**28180**	**42050**	**97.48%**	**91.10%**	**82.49%**	**71.09%**
Mean	9705	17930	28710	46730	107.50%	91.70%	84.42%	72.73%
Maximum	26560	40630	55670	310600	**162.30%**	232.90%	209.10%	160.70%

Table 2 contains chosen characteristics of location of net monthly incomes of households recalculated on consuming unit in particular income categories. On the left hand side the net incomes in CZK are shown; on the right hand side percentage shares of monthly expenditures from the total net monthly incomes are computed. It can be seen that the proportions significantly differ (from 8.9% to 232.9%). It is also obvious that with the growth of their financial power this

proportion decreases. For instance, median of proportion of finance used for consumption decreases from 97.5 % (in category low) to 71 % (in category high). In case of mean this decrease is even more striking – from 107.5 % to 72.7 %. And let us mention the fact that maximal proportions of expenditures among net incomes in all categories considerably exceed 100 %. It can be caused by nonrecurring expenditures of households.

For testing of differences in expenditures relative to the net incomes in particular categories Kruskal – Wallis rank sum test was used. First of all the difference in relative expenditures between at least one pair of categories was tested on 5 % significance level. And then we performed individual paired comparisons using Bonferoni correction. Tests showed that among all above defined categories there exist a significant difference in proportion of net incomes given monthly on consumption.

3.3 Differentiation in structure of the Czech household's expenditures

Tables 3 and 4 contain information concerning the frequencies of particular types of expenditures on consumption relatively to total household incomes.

Table 3. Chosen characteristics of monthly expenditures of the Czech households recomputed on the consumption unit (SRÚ 2008) for decreasing proportion of expenditures.

Financial power								
Type of characteristic	**low**	**medium low**	**medium high**	**High**	**low**	**medium low**	**medium high**	**high**
Food and beverages (CZK per consuming unit)					**Food and beverages / Total expenditures**			
Minimum	731	688	564	934	9.34%	5.81%	3.09%	3.27%
Median	**2285**	**3727**	**4582**	**4830**	**26.54%**	**24.12%**	**19.71%**	**15.79%**
Mean	2707	3965	4694	5043	27.34%	24.88%	20.18%	16.40%
Maximum	8090	11280	13440	16910	59.48%	57.87%	56.67%	37.01%
Housing and energies (CZK per consuming unit)					**Housing and energies / Total expenditures**			
Minimum	0	0	0	0	0%	0%	0%	0%
Median	**2810**	**3610**	**4273**	**4720**	**31.21%**	**23.85%**	**19.08%**	**16.15%**
Mean	2898	3822	4548	5189	31.34%	25.40%	20.16%	17.27%
Maximum	7894	18970	24140	26500	66.95%	63.79%	55.08%	54.37%
Health (CZK per consuming unit)					**Health / Total expenditures**			
Minimum	0	0	0	0	0%	0%	0%	0%
Median	**195**	**384**	**444**	**564**	**2.07%**	**2.56%**	**1.95%**	**1.93%**
Mean	289	503	589	762	3.12%	3.38%	2.52%	2.41%
Maximum	1736	5698	12630	10000	21.98%	45.99%	42.43%	20.95%
Communication (CZK per consuming unit)					**Communication / Total expenditures**			
Minimum	0	0	0	0	0%	0%	0%	0%
Median	**431**	**692**	**1042**	**1210**	**4.55%**	**4.59%**	**4.43%**	**4.10%**
Mean	549	836	1145	1362	5.25%	5.04%	4.88%	4.42%
Maximum	2395	4116	6145	5889	15.54%	20.91%	25.99%	22.80%

The Czech Statistical Office uses for classification of consumption expenditures international classification COICOP (Classification of individual consumption by purpose) according which the total consumption is divided into 12 groups:

- Food and beverages
- Housing and energies

- Health
- Communication
- Household equipment
- Recreation and culture
- Transport
- Alcohol and tobacco
- Clothing and footwear
- Education
- Restaurants and hotels
- Others

On the left hand side of tables 3 and 4 monthly expenditures per consuming unit can be seen; on the right hand side there appears shares of particular type of expenditures among the total expenditures of household. In table 3 there are collected types of expenditures where the share on total expenditures decreases. On the other hand table 4 contains expenditures for which the share on total expenditures increases. First group can be regarded as inferior goods whereas the second group of goods preferred by high-income households.

Similarly like in the case of total expenditures also in this case we can observe significant differences among relative expenditures of households with various financial powers. For testing of differences in relative representation of particular categories it is again shown that there exist significant differences among all categories.

Table 4. Chosen characteristics of monthly expenditures of the Czech households recomputed on the consumption unit (SRÚ 2008) for increasing proportion of expenditures.

Financial power								
Type of characteristics	**low**	**medium low**	**medium high**	**High**	**low**	**medium low**	**medium high**	**high**
Household equipment (CZK per consuming unit)					**Household equipment / Total expenditures**			
Minimum	0	17	0	62	0%	0.23%	0%	0.26%
Median	**266**	**630**	**1122**	**1658**	**2.96%**	**4.15%**	**5.07%**	**5.59%**
Mean	448	967	1647	2486	3.97%	5.43%	6.62%	7.54%
Maximum	4226	12710	23320	19720	31.35%	41.56%	47.10%	49.82%
Recreation and culture (CZK per consuming unit)					**Recreation and culture / Total expenditures**			
Minimum	0	15	86	262	0%	0.23%	0.56%	1.30%
Median	**582**	**1162**	**2101**	**3072**	**6.44%**	**8.03%**	**9.49%**	**10.38%**
Mean	803	1528	2606	3742	7.58%	8.90%	10.58%	11.59%
Maximum	8006	9384	13400	16510	27.54%	33.54%	49.44%	40.38%
Transport (CZK per consuming unit)					**Transport / Total expenditures**			
Minimum	0	0	0	5	0%	0%	0%	0.05%
Median	**203**	**760**	**1744**	**2646**	**2.33%**	**5.04%**	**7.83%**	**9.30%**
Mean	465	1256	2707	5023	3.88%	6.31%	9.64%	13.38%
Maximum	3834	21850	57240	47360	18.64%	47.95%	64.45%	72.48%
Alcohol and tobacco (CZK per consuming unit)					**Alcohol and tobacco / Total expenditures**			
Minimum	0	0	0	0	0%	0%	0%	0%
Median	**110**	**258**	**396**	**465**	**0,0120%**	**0,0168%**	**0,0175%**	**0,0152%**
Mean	311	488	688	766	0,0289%	0,0302%	0,0305%	0,0257%
Maximum	4465	4617	6015	5115	0,0289%	0,0302%	0,0305%	0,0257%
Clothing and footwear (CZK per consuming unit)					**Clothing and footwear / Total expenditures**			
Minimum	0	0	0	32	0%	0%	0%	0,0017%
Median	**263**	**561**	**1127**	**1636**	**0,0280%**	**0,0370%**	**0,0492%**	**0,0544%**
Mean	383	746	1319	1897	0,0348%	0,0422%	0,0532%	0,0594%
Maximum	2752	6268	6613	10330	0,1736%	0,1685%	0,1948%	0,2161%
Education (CZK per consuming unit)					**Education / Total expenditures**			
Minimum	0	0	0	0	0%	0%	0%	0%
Median	**0**	**0**	**0**	**0**	**0%**	**0%**	**0%**	**0%**
Mean	39	94	166	231	0,0029%	0,0041%	0,0059%	0,0059%
Maximum	1885	2967	5385	10000	0,1130%	0,1539%	0,1859%	0,1936%
Restaurants and hotels (CZK per consuming unit)					**Restaurants and hotels / Total expenditures**			
Minimum	0	0	0	0	0%	0%	0%	0%
Median	**186**	**470**	**1021**	**1491**	**1.94%**	**3.17%**	**4.63%**	**4.90%**
Mean	435	782	1353	1787	4.10%	4.31%	5.38%	5.56%
Maximum	4432	8483	11587	7885	53.84%	35.13%	29.66%	30.91%
Others (CZK per consuming unit)					**Others / Total expenditures**			
Minimum	34	12	120	262	0.60%	0.13%	0.91%	1.20%
Median	**484**	**1180**	**2372**	**3486**	**5.86%**	**7.64%**	**10.49%**	**11.73%**
Mean	752	1512	2655	3876	6.75%	8.70%	11.07%	12.34%
Maximum	4230	7488	12240	12990	22.62%	34.51%	46.96%	36.39%

3.3 Structure similarity in case of the Czech household's expenditures

Thus it was shown the proportion of particular expenditure types differs in different income categories. But it does not necessarily mean that the structure of expenditures in these categories differs. For measuring of (dis)similarity of structures we use the following two coefficients.

Table 5. Gatev coefficient of structure dissimilarity in case of expenditures relative to the total expenditures (SRÚ 2008).

Financial power	low	medium low	medium high	high
low	0			
medium low	0.1260044	0		
medium high	0.2784147	0.1605326	0	
high	**0.3908020**	0.2816644	0.1288356	0

Table 6. Cosine coefficient of structure similarity in case of expenditures relative to the total expenditures (SRÚ 2008).

Financial power	low	medium low	medium high	high
low	1			
medium low	0.9895620	1		
medium high	0.9422485	0.979256	1	
high	**0.8727787**	0.9296593	0.9841162	1

The results of structure (dis)similarity investigation are shown in tables 5 and 6. According to the assumptions both coefficients confirmed the highest differences between both "extreme" categories (low and high).

4 Conclusions

The proposed contribution is focused on the determination of differences of expenditures in case of households with different levels of financial power. First of all households were divided into four categories according to their financial power (low, medium-low, medium-high and high). The breaks between particular categories were derived from officially provided poverty threshold.

From the analysis of total expenditures of the Czech households in 2007 (SRÚ 2008), it seems that the differentiation of household incomes given their consumption is very high. It appears that maximal values of this ratio several fold surpass 100 %. Simultaneously it is shown that this proportion with growing financial power decreases and the differences between particular income groups are statistically significant.

In the following part the expenditure structure was investigated and its dependence on financial power. For this task the categorization into 12 groups according to COICOOP methodology was used. It appears that relative shares of expenditure types differ significantly according to the category of financial power. According to this behaviour the expenditures were divided into two groups. Among the expenditures more preferred by low-income households were the expenditures on food and beverages, housing and energies, health and communication. Their proportion was highest in the case of poorest households. The share of other expenditure groups, i.e. expenditures on household equipment, recreation and culture, transport, alcohol and tobacco, clothing and footwear, education, restaurants and hotels and others was greatest in case of the wealthiest households. But the question whether the total structure of incomes in these groups differs significantly must be answered negatively. Values of the both similarity coefficients shows

relatively high similarity of expenditure structure in all categories. But in spite of this fact, some dissimilarity was confirmed in case of the two "extreme" categories (low and high financial power).

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. *Aplimat – Journal of Applied Mathematics, Vol. **2**, Nr. **3***, Slovak Technical University, Bratislava, pp. 31-36.

[2.] BARTOŠOVÁ, J., FORBELSKÁ, M.: Porovnání regionální monetární chudoby v Čechách a na Slovensku. In Pauhofová, Hudec, Želinkský (eds.): *Sociálny kapitál, ľudský kapitál a chudoba v regiónech Slovenska (Zborník statí)*: Sociálny kapitál, ľudský kapitál a chudoba v regiónech Slovenska, 13 October 2010, Herlany. Košice: Technical University of Košice, 2010, pp. 76-84.

[3.] BARTOŠOVÁ, J., STANKOVIČOVÁ, I.: Diferenciace příjmů a chudoba českých a slovenských domácností. In Loster, Řezanková, Pavelka, Soukup (eds.): *MSED – Sborník příspěvků*: Mezinárodní statisticko-ekonomické dny na VŠE,17-18 září 2009, Praha. Praha: VŠE, 2009 (CD).

[4.] ČERMAKOVÁ, J.: The influence of income differentiation on the structure of household expenditures. *Finance a úvěr, vol. **51**, issue **1***, 2001, pp. 33-45.

[5.] KAHOUNOVÁ, J.: *Měření podobnosti struktur*. Praha: VŠE, 1994.

[6.] LABUDOVÁ, V., SIPKOVÁ, Ľ.: Facts of poverty in Slovakia. In: *Statistics in management of social and economic development*: 14th Ukrainian-Polish-Slovak scientific conference, September 24-28, 2007, Odessa, Ukraine. PALMIRA, 2008, pp. 40-48.

[7.] MORDUCH, J.: Poverty Measures. In: *Handbook on Poverty Statistics: Concepts, Methods and Policy Use*. New York: United Nations, Department of Economic and Social Affairs, 2005.

[8.] RAVALLION, M.: *Poverty Lines in Theory and Practice*. [LSMS Working Paper 133]. Washington, D. C.: The World Bank, 1998.

[9.] PASTOREK, L., STANKOVIČOVÁ, I.: Analýza činiteľov monetárnej chudoby v českých a slovenských domácnostiach v roku 2006. In: *Finanční potenciál domácností 2009 (Sborník dokumentů k pracovnímu semináři pořádanému v rámci řešení projektu GAČR 402/09/0515*. Finanční potenciál domácností 2009, 11. prosince 2009, J. Hradec. J. Hradec: FM VŠE, 2009 (CD).

[10.] SIPKOVÁ, Ľ., SIPKO, J.: Úroveň miezd v krajoch Slovenskej republiky. In Pauhofová, Hudec, Želinkský (eds.): *Sociálny kapitál, ľudský kapitál a chudoba v regiónech Slovenska (Zborník statí)*: Sociálny kapitál, ľudský kapitál a chudoba v regiónech Slovenska, 13 October 2010, Herlany. Košice: Technical University of Košice, 2010, pp. 51-66.

[11.] STANKOVIČOVÁ, I.: Regionálne aspekty monetárnej chudoby na Slovensku. In Pauhofová, Hudec, Želinkský (eds.): *Sociálny kapitál, ľudský kapitál a chudoba v regiónech Slovenska*

(Zborník statí): Sociálny kapitál, ľudský kapitál a chudoba v regiónech Slovenska, 13 October 2010, Herlany. Košice: Technical University of Košice, 2010, pp. 67-75.

[12.] STANKOVIČOVÁ, I.: Analýza činiteľov monetárnej chudoby v domácnostiach Českej republiky. *Forum Statisticum Slovacum, roč.* ***5****, č.* ***7/2009*** . Slovenská štatistická a demografická spoločnosť, Bratislava, pp. 151-156.

[13.] ŽELINSKÝ, T.: Analýza chudoby na Slovensku založená na koncepte relatívnej deprivácie. *Politická ekonomie, vol,* ***58****, issue* ***4***, 2010, pp. 542-565.

[14.] ŽELINSKÝ, T.: Pohľad na regióny Slovenska cez prizmu chudoby. In Pauhofová, Hudec, Želinkský (eds.): *Sociálny kapitál, ľudský kapitál a chudoba v regiónech Slovenska (Zborník statí)*: Sociálny kapitál, ľudský kapitál a chudoba v regiónech Slovenska, 13 October 2010, Herlany. Košice: Technical University of Košice, 2010, pp. 37-50.

[15.] ŽELINSKÝ, T.: Odhad vybraných ukazovateľov chudoby a ich štandardných chýb na regionálnej úrovni SR v prostredí R. In *Forum Statisticum Slovacum* ***7/2009***. SŠDS, Bratislava, pp. 209-214.

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

Journal of Applied Mathematics

volume IV (2011), number III

DIFFERENTIATION AND DYNAMICS OF HOUSEHOLD INCOMES IN THE CZECH EU-SILC SURVEY IN THE YEARS 2005 - 2008

BARTOŠOVÁ Jitka, (CZ), FORBELSKÁ Marie, (CZ)

Abstract. Finite mixtures of regression models are frequently used to model unobserved heterogeneity. Mixed models are widely applied to longitudinal data, modelling multiple observations from a single subject collected over time. We apply multivariate mixture model to the Czech longitudinal survey of household income in the European Union Statistics on Income and Living Conditions in 2005 – 2008. By means of the Finite mixtures of regression models, households are categorized according to four classes of income dynamics.

Keywords. Equivalised household income, finite mixture model, clustering, linear mixed models, poverty rate.

Mathematics Subject Classification: Primary 62H30, Secondary 30C40.

1 Introduction

Income and its progression over the years, is of interest in all countries. Studying how the average income evolves over time is not sufficient, because it does not inform about the extremes, and may hide growing poverty and differentiation. Studying the distribution of income in a sequence of years separately does not suffice either, because such an analysis does not inform about the stability of income, about the extent to which the income of some households grows faster (or changes differently) than the income of others.

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 equivalised household income into homogeneous subpopulations using the *mclust library of R* (see [3], [6]). Subsequently, regression analysis is performed using linear mixed models.

The article deals with cluster analysis of household income dynamics based on the results of statistical survey EU SILC between 2005 and 2008. We apply finite mixture models to compute the number of components of distribution of equalised income in Czech households and to characterize income stability and nobility in the Czech Republic between 2003 and 2007. We also study modeling of income dynamics in the respective components of the mixture. The article builds on works by Paap and van Dijk (1998), Pittau (2005) a Pittau and Zelli (2006).

1.1 EU-SILC survey during 2005 – 2008 years

The European Union Statistics on Income and Living Conditions (EU-SILC) is an instrument aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. EU-SILC is the main source for the compilation of comparable indicators on social cohesion used for policy monitoring at EU level in the framework of the Open Method of Coordination. This instrument is anchored in the European Statistical System (ESS).

The EU-SILC was launched under a gentleman's agreement with six EU-15 countries plus Norway in 2003 and re-launched under a Regulation with twelve EU-15 countries (Belgium, Denmark, Greece, Spain, France, Ireland, Italy, Luxembourg, Austria, Portugal, Finland and Sweden) and in Estonia, Norway and Iceland in 2004. In 2005 the rest of the EU-25 countries (Czech Republic, Slovakia etc.) joined the EU-SILC. Bulgaria, Romania, Turkey and Switzerland have launched SILC in 2006.

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 entrance to the European Union, Microcensus was replaced by annual survey of income and living conditions of households EU-SILC. The European Union Statistics on Income and Living Conditions (EU-SILC) is an instrument aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. This instrument is anchored in the European Statistical System (ESS).

1.2 Equivalised household income

The *equivalised household income* is used to allow comparisons between households of different sizes and composition. The equivalised household income is obtained by dividing the available household income by the number of consumption equivalents in the household. It is assumed that, as the size of the household increases and depending on the age of the children, cost savings are achieved in the household through joint budgeting (economies of scale). For weighting purposes, the EU scale (modified OECD scale) is used to calculate a household's resource requirements. An adult living on his or her own is taken as the reference point (*consumption equivalent*), with an allocated weighting of 1. For each additional adult, the assumed resource requirement increases by 0.5 consumption equivalents. Each child under the age of 14 is weighted with a consumption equivalent of 0.3. So a household comprising a father, mother and child would have a calculated consumption equivalent of 1.8 compared with a single-person household.

1.3 Poverty rate

The *poverty rates* discussed here are defined as the percentage of those having less than 60% of the median income. It means, the poverty rates is percentage of persons in the total population with an equivalised disposable income below the "national poverty line" (i.e. below 60% of the national median equivalised disposable income). Total population is all persons living in private household on the national territory. Total disposable income of a household is calculated by adding together the personal income received by all of the household members, plus income received at household level. Disposable household income includes all income from work, private income from investment en property, transfers between households and all social transfers received in cash including old-age pensions (see remarks for more detailed definition).

Table 1: Annual national poverty levels

Year	**2005**	**2006**	**2007**	**2008**
Poverty threshold	80986	85714	92212	101016
Risk of poverty rate				
Longitudinal data (income per households)	9,38%	9,31%	8,88%	9,27%
All data (income per households)	10,31%	9,80%	10,25%	10,63%
Longitudinal data (income per individuals)	9,10%	9,6%	9,8%	10,4%

2 Model Specification

The estimation of income distributions is important for assessing income inequality and poverty and for making comparisons of inequality and poverty over time. Distributions have been estimated both parametrically and non-parametrically.

The normal distribution, with a particular symmetric shape of its density, is the mainstay of statistical modelling. Possibly after some adjustment for covariates, normality is an appropriate assumption for modelling many phenomena. In a variety of settings, income has a skewed distribution; income cannot be negative, a small fraction of the units has very small (or zero) income, a large fraction has income in a relatively narrow range, and a few units have high income spread across a wide range. Lognormal distribution is better suited for modelling income (in the particular setting) than normal distribution, but further improvement on the fit would be highly desirable. We seek improvement by mixture modelling. A sample (or a population) of units is said to be a mixture if it comprises several subsamples (groups), each with a distinct distribution of the studied variable.

2.1 The Explanatory Mixture Model

Next we assume that the equivalised household incomes can be broken down into K homogeneous subpopulations (strata) with proportions $\pi_1,\ldots,\pi_K$ and that each household income Y follows normal distribution. We let y_i denote the value of Y corresponding to the ith entity ($i = 1,\ldots, N$). With the mixture approach to clustering, $y_1,\ldots,y_n$ are assumed to be an observed random sample from mixture of a finite number of groups in some unknown proportions $\pi_1,\ldots,\pi_K$. The mixture density of y_i is expressed as

$$f(y_i;\boldsymbol{\Psi}) = \sum_{j=1}^{K} \pi_j f_j(y_i;\boldsymbol{\theta}_j) \tag{2.1}$$

where the mixing proportions $\pi_1,\ldots,\pi_K$ sum to one and the group-conditional density $f_j(y_i;\boldsymbol{\theta}_j)$ is specified up to a vector $\boldsymbol{\theta}_j$ of unknown parameters ($j= 1,\ldots,K$). The vector of all the unknown parameters is given by $\boldsymbol{\Psi} = (\pi_1,\ldots,\pi_{K-1},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)$. Using an estimate of $\boldsymbol{\Psi}$, this approach gives a probabilistic clustering of the data into k clusters in terms of estimates of the posterior probabilities of component membership,

$$\omega_j(y_i) = \frac{\pi_j f_j(y_i;\boldsymbol{\theta}_j)}{f(y_i;\boldsymbol{\Psi})}, \tag{2.2}$$

where $\omega_j(y_i)$ is the posterior probability that y_i (really the entity with observation y_i) belongs to the j-th component of the mixture ($i= 1,\ldots,n$, $j= 1,\ldots,K$). In the Bayesian framework, we use the rule which assigns observation x_i to the class for which y_i has the highest posterior probability. The parameter vector $\boldsymbol{\Psi}$ 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]).

We use the *mclust* package (Fraley and Raftery, 2006) of the R environment for fitting mixtures model. Figure 1 and Table 2 show the resulting optimal component partitions given by EM algorithm with BIC criterion. The Figure 1 also shows that the first component from 2006 to 2008 in 2005 split into two sub-components, so the next will be considered merged.

Table 2: Estimated parameters for the optimal component mixtures of *equivalised income* over years 2005 – 2008

Year	Comp.	Estimated parameters of mixtures				Estimated parameters of classified households			
		prop	mean	sigma	var. coeff.	prop	mean	sigma	var. coeff.
2005	1	0.218	90187	28096	31.15%	0.192	73728	17434	23.65%
2005	2	0.248	114211	19431	17.01%	0.342	114283	11115	9.73%
2005	3	0.384	161557	42841	26.52%	0.378	175815	29871	16.99%
2005	4	0.141	246162	97621	39.66%	0.080	316674	64074	20.23%
2005	**5**	**0.010**	**707948**	**498193**	**70.37%**	**0.007**	**849986**	**489290**	**57.56%**
2006	1	0.376	110919	25935	23.38%	0.480	105383	19910	18.89%
2006	2	0.502	158402	51381	32.44%	0.447	176518	43518	24.65%
2006	3	0.113	269374	107049	39.74%	0.067	345870	74185	21.45%
2006	**4**	**0.009**	**760232**	**535303**	**70.41%**	**0.006**	**941468**	**538509**	**57.20%**
2007	1	0.288	119070	23967	20.13%	0.375	113295	15504	13.68%
2007	2	0.481	155809	50168	32.20%	0.481	164823	46265	28.07%
2007	3	0.215	242493	94951	39.16%	0.133	308895	63590	20.59%
2007	**4**	**0.016**	**613300**	**409740**	**66.81%**	**0.011**	**771951**	**407788**	**52.83%**
2008	1	0.267	126521	25554	20.20%	0.355	118488	16858	14.23%
2008	2	0.438	165796	48445	29.22%	0.458	175165	39299	22.44%
2008	3	0.277	237614	96542	40.63%	0.175	295190	84013	28.46%
2008	**4**	**0.018**	**630300**	**380840**	**60.42%**	**0.012**	**792083**	**356501**	**45.01%**

Estimated parameters of mixtures and classified households show that the first three components contain almost all equivalised income. Their variance is about half as low as the variance of the fourth component that models extremes of income distribution. As a consequence, it shows high variance and significant shift of values to the right. The fourth component only comprises between 0,6 and 1,8 % of Czech households. If mostly comprises households with high and extreme income. The average of income is thus about three times as high as in the third component.

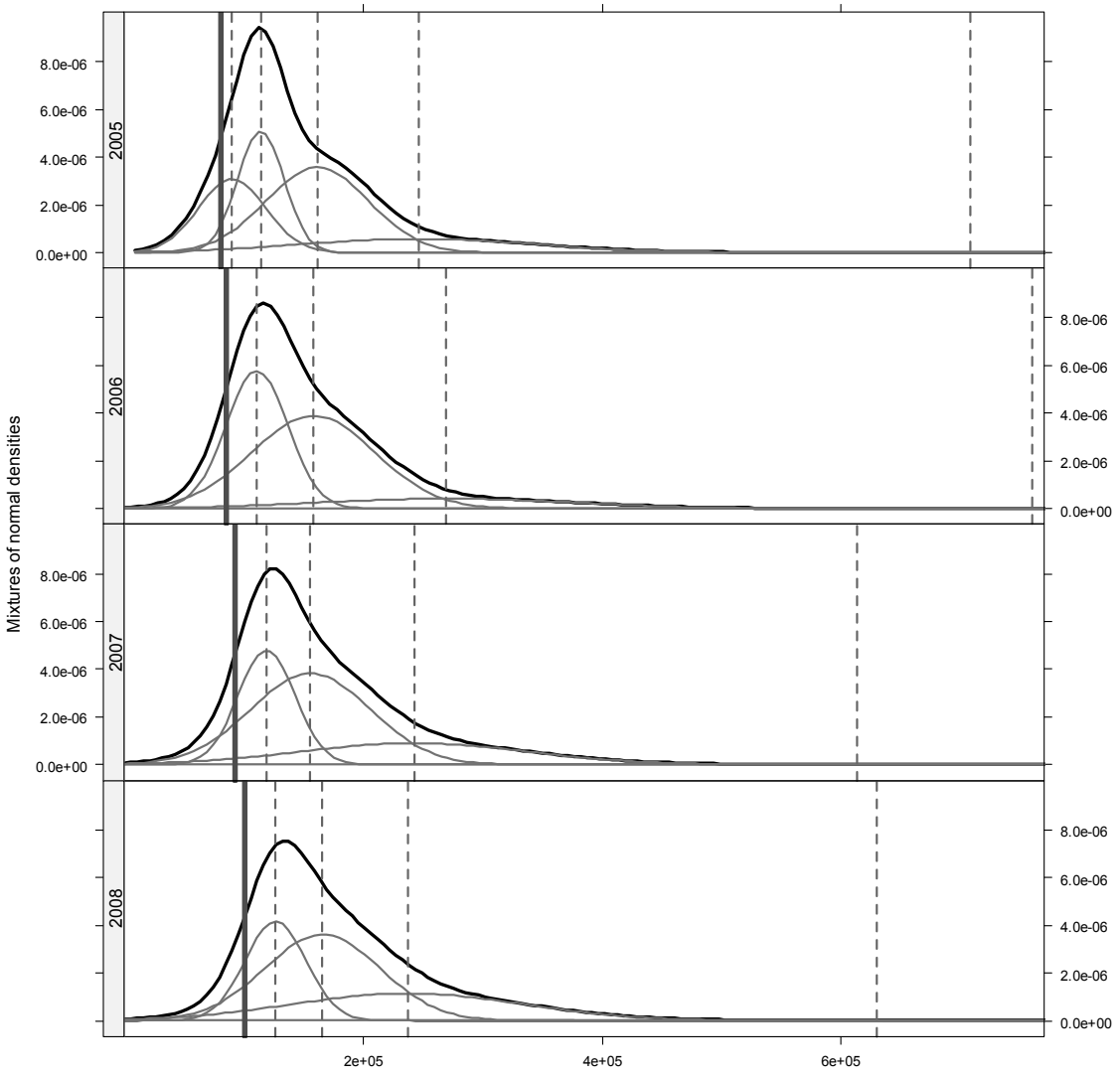


Figure 1: *BIC*-optimal component mixtures of *equivalised income* in the years 2007 – 2008. (Bold solid line indicates the level of poverty (see Table 1), dashed lines indicate the positions of the mean values of individual components). Data source SILC 2005 – 2008.

It shows that Bayesian Information Criterion *BIC* divides households into four income categories using the EM algorithm. The categories are: Households with low income, average income, higher income and a small category of households with high income. This corresponds to the real image of partition of the population into the low-income class that is at risk of poverty, lower-middle class that may drop under the poverty line as a consequence of the crisis and lastly the middle class that

may remain relatively stable or even move up to the high-income-class. Risk-of-poverty-rate in individual components can be found in Table 3.

The situation is also pictured in Figure 1, where aside from the income distribution and its components, we also exhibit the poverty line. We can see that the income distribution in 2005 is steeper from the left than in upcoming years. This indicates a shallow depth of poverty and also explains why only households from the first component (17.55%) fell below the poverty line in 2005. From Figure 1 we can see how poverty in the Czech republic gradually raises, i.e., how an increasing proportion of households at risk of monetary poverty drift away from poverty to the left – in direction of lower values. At the same time, we can observe partial selection within the same component, which is a prerequisite for creating convergence classes. For example, Table 2 shows that in the first two components, the ratio of estimated parameters sigma and mean (variation coefficient of estimate) decreased in 2008 compared to 2006 (in the first component by 3.38 percentage points in the second by 3.22 pp)., while the third component increased by. 1.11 pp Also, the relative variability in household income included in the first and second components between years 2006 and 2008 decreased (in the first component from 18.89% to 14.23% and in the second from 24.65% to 22.44%). In contrast, the relative variability of income of households belonging to the third component increased from 21.45% to 28.46%. However, the biggest drop in variability appeared in the fourth component, where variation coefficient of parameter estimates for the period 2006 – 2008 decreased from 70.37% to 60.42%, i.e. nearly 10 pp and relative variability of income in this component decreased from 57.56% to 45.01%, i.e. by 11.55 pp.

Table 3: Percentage of Czech households at risk of monetary poverty in respective components. Data source EU SILC 2005 – 2008.

Component	Year	Poverty No	Poverty Yes
1	2005	82,45%	**17,55%**
	2006	83,73%	**16,27%**
	2007	89,82%	**10,18%**
	2008	84,01%	**15,99%**
2	2005	100	0
	2006	96,21	**3,79**
	2007	89,51	**10,49**
	2008	94,07	**5,93**
3	2005	100	0
	2006	100	0
	2007	100	0
	2008	94,83	**5,17**
4	2005	100	0
	2006	100	0
	2007	100	0
	2008	100	0

The percentage of the components in the category of Czech households living below the poverty line between 2005, and 2008, is given in Table 4. The table demonstrates in another way the fact that in 2005, all poor households are located in the first component. The above described developments of income in the Czech Republic resulted in the years 2006 – 2008 in occurrence of poverty in other (middle-income) components.

Table 4: Percentages of the components below poverty line in the Czech Republic. Data source EU SILC 2005 – 2008.

Year	2005		2006		2007		2008	
Poverty	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
Comp.1	48,65	100	42,96	81,27	36,77	42,76	32,55	60,68
Comp.2	42,29	0	48,81	18,73	47,59	57,24	47,45	29,29
Comp.3	8,12	0	7,76	0	14,5	0	18,79	10,04
Comp.4	0,95	0	0,47	0	1,13	0	1,2	0

2.2 Model Specification over Short Time Periods

Let $\{Y_{ijt}\}_{t=1,\ldots,T}$ be a panel of multiple time series observed for N units ($t=1,\ldots,T, j=1,\ldots,K, i=1,\ldots,n_j$, $n_1+\cdots+n_K=N$). Denote also by $Y_{ij}=(Y_{ij1},\ldots,Y_{ijT})'$ (ij)-th univariate time series with the joint density $f(y;\boldsymbol{\theta}_{ij})$, where $\boldsymbol{\theta}_{ij}$ is unknown parameters that need to be estimated from the data. If T were large, the parameters $\boldsymbol{\theta}_{ij}$ could be estimated for each time series $Y_{ij}=(Y_{ij1},\ldots,Y_{ijT})'$ individually. However, if T is relatively small one might use information from the other time series in the panel to estimate unknown parameters.

2.2.1 Simple Linear Mixed Model (LMM)

For our longitudinal data we assume very simple linear mixed model described by the structure

$$Y_{ij}=(1,t)\beta_{ij}^{*}+\varepsilon_{ij} \tag{2.3}$$

where $\beta_{ij}^{*}=(a_{ij}^{*},b_{ij}^{*})'=(a_j+u_{ij},b_j+v_{ij})'$, 1 is vector of ones, $t=(1,\ldots,T)'$, $j=1,\ldots,K$, $i=1,\ldots,n_j,.\, n_1+\cdots+n_K=N$. In this model we call $\beta_j=(a_j,b_j)'$ the *fixed effects* (fixed intercept and fixed slope) and $z_{ij}=(u_{ij},v_{ij})'$ the *random effects* (random intercept and random slope). We assume that random vectors z_{ij} and ε_{ij} are independent and identically distributed. Parameters of the mixed model can be estimated using Maximum Likelihood Estimation (MLE) or Restricted Maximum Likelihood Estimation (RMLE), while the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) can be used as measures of "goodness of fit" for particular models, where smaller values for both are considered more preferable. We use the *lme4* package (Bates and Maechler, 2010) of the R environment for fitting and examining linear mixed-effects models.

Table 5: Estimated parameters of the LMM model of *equivalised income* over years 2005 – 2008

Parameter	Component 1	Component 2	Component 3	Component 4
a (intercept)	119601.752	172356.282	281959.48	682573.97
b (year)	7683.591	1737.615	-10716.08	-53996.76

REML-estimates of random parameters related to their variability and correlation are summarized in Table 6.

Table 6: REML-estimates of variation characteristics of random effects.
(Data-resource EU SILC, own calculations)

Parameters of random effects		
$\sigma_1 = 1548.6$	$\sigma_2 = 606.2$	$\rho = 0.601$
Parameter of variability of error component		
$\sigma_e = 34330.9$		

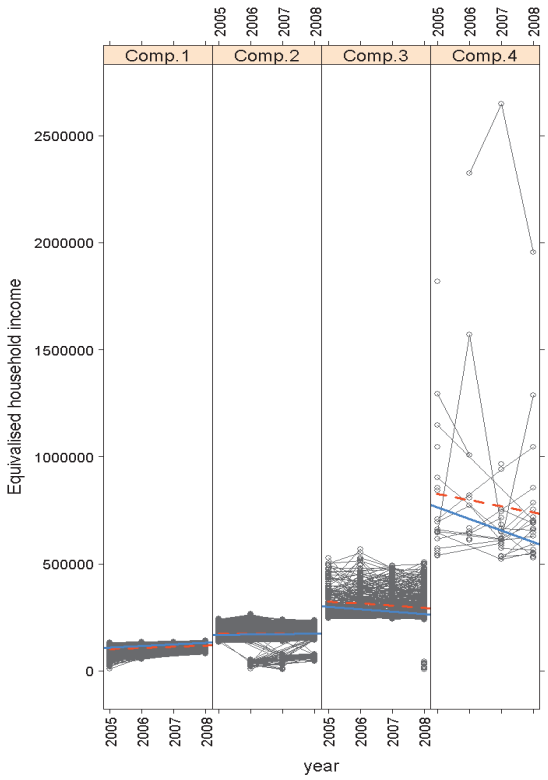


Figure 2: Resulting LMM model of *equivalised income* over years 2007 – 2008
for BIC-optimal components **(**solid line is a straight line with fixed parameters and describes the components of the average trajectory, the dashed line is the ordinary OLS regression line).
Data source SILC 2005 – 2008.

Figure 2 pictures all trajectories of equivalent household income throughout the researched period of four years-separately for each component. In addition, there are also the regression lines with fixed coefficients, a_j, and b_j, i.e. lines that characterize the average behaviour of each component. Another line of interest, the regression line estimated from all observations of a given component using the classical method of least squares (OLS Fitted curve), is plotted in the graph.

The choosen method models dynamics of components in each year separately; hence equivalent income of some households is not included in the same component for the whole four years. The households are devided into components regardless which component they belonged to in the previous or following year. Therefore the trajectory of their four-year-development can be

disconnected (see Figure 2). This inaccuracy can be rectified by applying Regression Mixtures, where we partition the trajectories of equivalent income of households into components in an iterative way, so we avoid development discontinuity.

For every (*ij*)-th household inside of the *j*-th component, the LMM model predicts its own values of parameters of lines $a_{ij}^{*} = a_j + u_{ij}$, and $b_{ij}^{*} = b_j + v_{ij}$. Their values are plotted in the graph (see Figure 2) and around every pair of parameters for each component, there is an ellipse of concentration together with the position of fixed parameters (a_j, b_j) (coloured full squares) and sampling means of parameters of each component $(\bar{x}_j, \bar{y}_j)$ (red point). The images show how coefficients of the LMM model estimated by two-step-method form disjoint clusters.

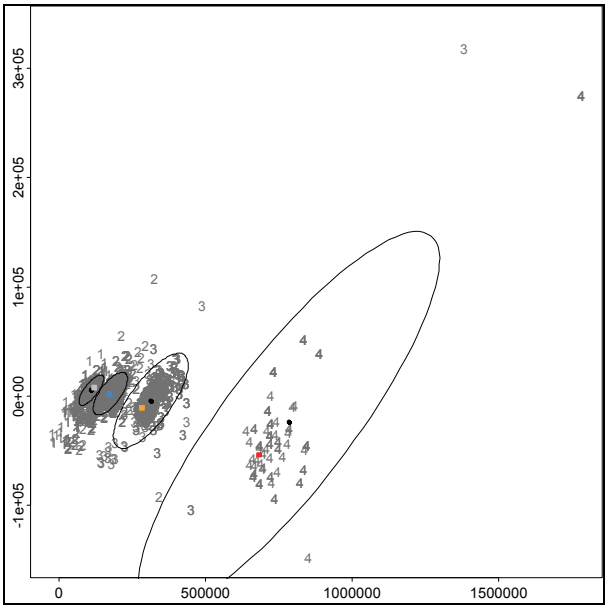


Figure 3: The position of parameters $(a_{ij}^{*} = a_j + u_{ij}, b_{ij}^{*} = b_j + v_{ij})$ expressed by the number of components for *(ij)-th* household. Full squares: values of fixed parameters (a_j, b_j), black circles: sample mean coefficients of components (data - the source of EU-SILC, the calculation).

4 Conclusions

Present paper demonstrates that using Finite Mixture Models and Linear Mixture Models, we can devide households into several classes that correspond to natural partition of the population into low-income class, lower middle class, higher middle class and high-income class. Further, we can analyse in detail the structure of individual components of the mixture from both social and demographic point of view. The position, variability and shape of income distribution in the components can assess stability, mobility and differentiation, respectively polarization of income, identify mutual connections and make predictions. Linear Mixture Models allow us sensitively model dynamics of income development in individual components and observe the trajectories of individual development of financial power of households.

Finite mixtures are thus a very suitable stochastic tool to make so-called classification without a teacher, that we sometimes also call stochastic cluster analysis. Or, if we want to be more specific, we call mixture-model-based clustering.

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] BATES, D., MAECHLER, M.: *lme4*: *Linear mixed-effects models using S4 classes. R package version* 0.999375-37. http://CRAN.R-project.org/package=lme4, 2010
[3] 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.
[4] FRALEY, C., RAFTERY, A. E.: Model-based Clustering, Discriminant Analysis and Density EstimationJournal of the American Statistical Association97:611-631; 2002.
[5] FRALEY, C., RAFTERY, A. E.: *MCLUST Version 3 for R: Normal Mixture Modelling and Model-based Clustering*. Technical Report No. 504, Department of Statistics, University of Washington, 2006 (revised 2009).
[6] HOROVÁ, I., ZELINKA, J. *Contribution to the bandwidth choice for kernel density estimates*. In Computational Statistics, Springer, 22, 1, pp. 31-47, 2007.
[7] McLACHLAN, G. J. , Peel, D.: *Finite mixture models*. New York: Wiley & Sons, 2000.
[8] 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
[9] SCHWARTZ, G.: *Estimating the Dimension of a Model*. In The Annals of Statistics, 6 (2), pp. 461-464, 1978.

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: forbel@math.muni.cz

Journal of Applied Mathematics

volume IV (2011), number III

FORECASTING VOLATILITY WITH WAVELETS: METHODOLOGY

BAŠTA Milan, (CZ)

Abstract. The dynamics of volatility of financial markets shows different behavior at different time scales. We propose to use the Haar Maximal Overlap Discrete Wavelet Transform (MODWT) to forecast volatility. We also propose to analyze the logarithm of volatility to avoid positivity constraints and to use the logarithm of the Garman-Klass estimator to proxy log volatility. Interscale dynamics is shortly studied and discussed. Ideas for further research are outlined.

Key words and phrases. Time series, wavelets, volatility, forecasting, heterogeneous agents

Mathematics Subject Classification. Primary 62M20; Secondary 42C40, 91B28, 91B69.

1 Introduction and motivation

It may be assumed that financial markets consist of several groups of heterogeneous agents. For example, the US stock market consists of market makers, intraday traders, arbitrageurs, traders within a few days, hedge funds, portfolio managers, "buy and hold" traders, investment funds, investment banks etc. Each group of traders uses different tools to analyze the market, has an access to a different sort of information, has a different level of risk aversion, different institutional and personal trading constraints, strategies and preferences. These differences among groups lead to different time scales each group operates on (see e.g. [16]). For example market makers operate on the shortest investment horizons of several seconds, whereas "buy and hold" traders may hold their financial assets for several days. Such a heterogeneity of the market implies a different reaction to the same news. For example, a high-frequency trader using the technical analysis to buy and sell will promptly react to changing patterns in the chart of the price but will be immune to the fundamentals of the company, the shares of which he

owns. On the other hand, a "buy and hold" trader might be concerned with these fundamentals and be completely inattentive to the indicators of the technical analysis.

The reaction to news and information is closely related to the volatility of the market. Thus, if the reaction to information differs across the groups of traders, then it might be possible that the dynamics of volatility is different across different time scales. There have been several studies of stock as well as FX markets that support the above outlined hypothesis. For example, the results of [1] and [23] suggest the existence of the causal information cascade in volatility. More specifically, shocks in volatility at larger scales propagate further to shorter scales. On the other hand, shocks on shorter scales do not influence the larger scales much. Such a result was also confirmed by [12]. Various models have emerged to explain this kind of dynamics, see e.g. [12] or [15] for some of these. In this paper we argue that wavelets are an intriguing tool suitable for studying volatility of heterogeneous markets.

2 Volatility, its traditional models and proxies

Let close_t be the time series of the closing stock price. The time series of returns is defined as

$$R_t \equiv \frac{\text{close}_t}{\text{close}_{t-1}} - 1. \tag{1}$$

In many practical situations it might be more convenient to work with logarithmic returns (hereafter log returns) defined as

$$r_t \equiv \ln(R_t + 1). \tag{2}$$

Let h_t be the conditional variance, hereafter called volatility, of log returns defined as[1]

$$h_t \equiv E_{t-1}\left\{r_t^2\right\}, \tag{3}$$

where $E_{t-1}\{.\}$ is the expectation operator given the set of information available at time $t-1$.

2.1 Traditional models of volatility

A well-known way of modeling time-varying volatility of financial markets are ARCH(q) models by [7]

$$h_t = \omega + \sum_{i=1}^{q} \alpha_i r_{t-i}^2, \tag{4}$$

where ω and α_i are parameters, which are constrained to $\omega > 0$ and $\alpha_i \geq 0$ to ensure that volatility is positive. Bollerslev [4] proposed the GARCH(p,q) models of volatility

$$h_t = \omega + \sum_{i=1}^{q} \alpha_i r_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i}, \tag{5}$$

[1]For financial markets it is usually assumed that $E_{t-1}\{r_t\} = 0$.

where $\omega > 0$, $\alpha_i \geq 0$ and $\beta_i \geq 0$ are parameters. The empirical analysis of the time series r_t^2 (e.g. [11]), $|r_t|$ (e.g. [10]) or $\log r_t^2$ (e.g. [5]) suggests that the processes of volatility have long-memory, which cannot be captured by traditional ARCH(q) or GARCH(p,q) of Eq. 4 and 5. To capture long-memory e.g. the FIGARCH model by [2] may be used.

2.2 Proxies to volatility

It might be difficult to assess the predictive power of the forecasts, because volatility is a latent variable and thus is not directly observable. If high-frequency data are not available a common proxy to daily volatility are the squared log returns. To illustrate, let us assume

$$r_t = \sqrt{h_t} u_t, \tag{6}$$

where $u_t \sim N(0,1)$ are i.i.d. random variables with zero mean, unit variance and Gaussian cumulative distribution function. Then

$$r_t^2 = h_t u_t^2, \tag{7}$$

$$E[r_t^2] = h_t. \tag{8}$$

Thus squared log returns are an unbiased estimator of volatility. However, at the same time this estimator is very noisy (see e.g. [20]).

For these reasons it might be more desirable to obtain less noisy estimates of volatility. If high frequency data are not available but if we have information on the high, low, closing and opening price for each day we can use the Garman-Klass estimator of the intraday volatility (i.e. the volatility from open to close), which is defined as (see [8])

$$K_t \equiv 0.511(m_t - l_t)^2 - 0.019\left[c_t(m_t + l_t) - 2m_t l_t\right] - 0.383c_t^2, \tag{9}$$

where

$$m_t \equiv \ln(\max_t) - \ln(\text{open}_t), \quad l_t \equiv \ln(\min_t) - \ln(\text{open}_t), \quad c_t \equiv \ln(\text{close}_t) - \ln(\text{open}_t), \tag{10}$$

where $\max_t$, $\min_t$, close_t and open_t is the high, low, closing and opening price of day t.

3 MODWT

As argued in the section 1 volatility exhibits scale-dependent dynamics. It might be thus interesting to study whether wavelets, which are a suitable tool for analyzing dynamics on different time scales, might be used for the forecast of volatility. In this section the Maximum Overlap Discrete Wavelet Transform (MODWT) will be summarized, which will be used for the decomposition of the volatility with respect to the time scale and its multiscale forecast afterward. Our short introduction to MODWT (i.e. section MODWT) is based on the book [19], which encompasses further details and proofs of the statements given below.

3.1 MODWT filters

A linear filter a_t is a sequence of weights, i.e.

$$a_t \equiv \{\ldots a_{-2}, a_{-1}, a_0, a_1, a_2 \ldots\}. \tag{11}$$

The linear filtration of a time series x_t (or a stochastic process $\{x_t\}$) is defined as

$$a_t * x_t \equiv \sum_m a_m x_{t-m}, \tag{12}$$

where $*$ stands for the operation of convolution. An important characteristics of the linear filter is its frequency response defined as the Fourier transform of a_t, i.e.

$$A(f) \equiv \sum_t a_t \exp(-i2\pi ft), \quad -\infty < f < \infty, \tag{13}$$

where f is the frequency.

The MODWT may be thought of as the linear filtration of the time series or the stochastic process with a special set of linear filters. This special set of linear filters may be created if two elementary filters are given[2] – the wavelet filter $\tilde{h}_{1,t}$ and the scaling filter $\tilde{g}_{1,t}$. It holds that $\tilde{h}_{1,t} \equiv 0$ for $t < 0$ and for $t \geq \mathcal{L}_1$. The number $\mathcal{L}_1$ is called the length of the filter. The filters $\tilde{h}_{1,t}$ and $\tilde{g}_{1,t}$ are interconnected via the so called quadrature mirror relationship

$$\tilde{g}_{1,t} \equiv (-1)^{l+1} \tilde{h}_{1,L-1-t}, \tag{14}$$

and fulfill certain specific conditions, e.g.

$$\sum_t \tilde{h}_{1,t} = 0, \quad \sum_t \tilde{h}_{1,t}^2 = \frac{1}{2}, \quad \sum_t \tilde{h}_{1,t}\tilde{h}_{1,t+2n} = 0 \text{ for } n \neq 0, \tag{15}$$

$$\sum_t \tilde{g}_{1,t} = 1, \quad \sum_t \tilde{g}_{1,t}^2 = \frac{1}{2}, \quad \sum_t \tilde{g}_{1,t}\tilde{g}_{1,t+2n} = 0 \text{ for } n \neq 0 \tag{16}$$

and

$$|\tilde{H}_1(f)| \approx \begin{cases} 1, & \frac{1}{4} < |f| \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}, \tag{17}$$

$$|\tilde{G}_1(f)| \approx \begin{cases} 1, & 0 \leq |f| \leq \frac{1}{4} \\ 0, & \text{otherwise} \end{cases}, \tag{18}$$

where $\tilde{H}_1(f)$ is the frequency response of the linear filter $\tilde{h}_{1,t}$ and $\tilde{G}_1(f)$ is the frequency response of the linear filter $\tilde{g}_{1,t}$. Thus, the filter $\tilde{h}_{1,t}$ is approximately a high-pass filter that passes frequencies in the region $\frac{1}{4} < |f| \leq \frac{1}{2}$ and the filter $\tilde{g}_{1,t}$ is approximately a low-pass filter that passes frequencies in the region $0 < |f| \leq \frac{1}{4}$.

From the filters $\tilde{h}_{1,t}$ and $\tilde{g}_{1,t}$ two sets of filters may be created via the so called pyramid algorithm (the algorithm not specified here):

[2]To keep the notation the same as in [19] tildes above the letters are given throughout the text.

1. the set of wavelet filters: $\tilde{h}_{1,t}$, $\tilde{h}_{2,t}$, $\tilde{h}_{3,t}\ldots$

2. the set of scaling filters: $\tilde{g}_{1,t}$, $\tilde{g}_{2,t}$, $\tilde{g}_{3,t}\ldots$

Let $\mathcal{L}_j$ be the length of the filter $\tilde{h}_{j,t}$ and $\tilde{g}_{j,t}$, $j = 1, 2, \ldots$. Thus, it holds that $\tilde{h}_{j,t} = 0$ for $t < 0$ and $t \geq \mathcal{L}_j$. Similarly, $\tilde{g}_{j,t} = 0$ for $t < 0$ and $t \geq \mathcal{L}_j$. Let $\tilde{H}_j(f)$ be the frequency response of the linear filter $\tilde{h}_{j,t}$ and $\tilde{G}_j(f)$ the frequency response of the linear filter $\tilde{g}_{j,t}$. It can be shown that it holds

$$|\tilde{H}_j(f)| \approx \begin{cases} 1, & \frac{1}{2^{j+1}} < |f| \leq \frac{1}{2^j} \\ 0, & \text{otherwise} \end{cases}, \tag{19}$$

$$|\tilde{G}_j(f)| \approx \begin{cases} 1, & 0 < |f| \leq \frac{1}{2^{j+1}} \\ 0, & \text{otherwise} \end{cases}. \tag{20}$$

Thus, the filter $\tilde{h}_{j,l}$ is a band-pass filter for the range of frequencies $\frac{1}{2^{j+1}} < |f| \leq \frac{1}{2^j}$ and $\tilde{g}_{j,l}$ is a low-pass filter that passes frequencies in the range $0 < |f| \leq \frac{1}{2^{j+1}}$.

One example are the Haar filters. It can be shown that for the Haar wavelet filters it holds

$$\tilde{h}_{j,t} \equiv \begin{cases} \frac{1}{2^j}, & \text{for } t = 0, \ldots, 2^{j-1} - 1 \\ -\frac{1}{2^j}, & \text{for } t = 2^{j-1}, \ldots, 2^j - 1 \\ 0, & \text{otherwise} \end{cases}, \tag{21}$$

and for the Haar scaling filters it holds

$$\tilde{g}_{j,t} \equiv \begin{cases} \frac{1}{2^j}, & \text{for } t = 0, \ldots, 2^j - 1 \\ 0, & \text{otherwise} \end{cases}. \tag{22}$$

The lengths of the Haar filters $\tilde{h}_{j,t}$ and $\tilde{g}_{j,t}$ are equal to $\mathcal{L}_j = 2^j$.

3.2 MODWT coefficients

If we apply the wavelet filter $\tilde{h}_{j,t}$ onto the time series $x_t : \ t = 0, \ldots, N-1$ we get the sequence

$$\tilde{w}_{j,t} \equiv \tilde{h}_{j,t} * x_t, \tag{23}$$

called the jth level MODWT wavelet coefficients. Similarly, if we apply the scaling filter $\tilde{g}_{j,t}$ onto the time series x_t we get

$$\tilde{v}_{j,t} \equiv \tilde{g}_{j,t} * x_t, \tag{24}$$

called the jth level MODWT scaling coefficients. It can be shown that the wavelet coefficients $\tilde{w}_{j,t}$ are (for a wide class of filters) associated with the dynamics on the scale of 2^{j-1}. On the other hand, the scaling coefficients $\tilde{v}_{j,t}$ are associated with the dynamics on the scale of 2^j.

To calculate $\tilde{w}_{j,t}$ (of Eq. 23) for all values of $t = 0, \ldots, N-1$ (so that the sequence $\tilde{w}_{j,t}$ has the same length as is the length of x_t) values of x_t preceding the value x_0 are needed (because $\tilde{h}_{j,t}$ is a causal filter). If these values are not available then circularity of x_t is usually assumed, i.e. $x_{-1} \equiv x_{N-1}$, $x_{-2} \equiv x_{N-2}$ etc. The same chain of thoughts applies to $\tilde{v}_{j,t}$.

4 Multiscale forecasting of volatility with Haar wavelets

In our multiscale forecasting procedure we may exploit the fact that an additive decomposition of x_t is possible via *Haar* MODWT coefficients, i.e. (see e.g. [19], exercise [5.10] at p. 205)

$$x_t = \tilde{v}_{J,t} + \sum_{j=1}^{J} \tilde{w}_{j,t}, \tag{25}$$

where $\tilde{v}_{j,t}$ and $\tilde{w}_{j,t}$ are Haar scaling and wavelet coefficients of the time series x_t and $J \geq 1$ is an integer. The decomposition of Eq. 25 holds only for Haar wavelets, which will thus be assumed to be used further on. Coefficients unaffected by the circularity assumption may be easily used for forecasting. The h-step ahead forecast of x_t, denoted as $x_t[h]$, is thus accomplished as

$$x_t[h] = \tilde{v}_{J,t}[h] + \sum_{j=1}^{J} \tilde{w}_{j,t}[h], \tag{26}$$

where $\tilde{v}_{J,t}[h]$ and $\tilde{w}_{j,t}[h]$ are the h-step ahead forecasts of $\tilde{v}_{J,t}$ and $\tilde{w}_{j,t}$.

4.1 Avoiding positivity constraints

Forecasting based on Eq. 26 can generally lead to forecasts of volatility that are negative. However, volatility must be always positive. To avoid these problems we propose to work with the logarithm of volatility (hereafter log volatility) defined as

$$H_t \equiv \ln h_t. \tag{27}$$

If Eq. 6 holds and if u_t (of Eq. 6) is Gaussian then (see also e.g. [5])

$$\ln(r_t^2) = H_t + \ln(u_t^2). \tag{28}$$

$$E\left\{\ln(u_t^2)\right\} = -1.27, \quad \text{var}\left\{\ln(u_t^2)\right\} = \frac{\pi^2}{2} \doteq 4.93. \tag{29}$$

Logarithm of squared log returns is thus a very noisy estimator of log volatility. Moreover, the probability distribution of the noise $\ln(u_t^2)$ is highly asymmetric.

Analogously to subsection 2.2 we may rather use the logarithm of the Garman-Klass estimator defined in Eq. 9 rather than the logarithm of squared log returns to proxy log volatility. Let us define

$$L_t \equiv \ln(K_t). \tag{30}$$

If the process of stock prices is random walk then

$$L_t = H_t + \ln(\zeta_t), \tag{31}$$

where $\ln(\zeta_t)$ are i.i.d. variables independent of H_t with approximately Gaussian distribution and [14]

$$E\left\{\ln(\zeta_t)\right\} \doteq -0.13, \tag{32}$$

$$\text{var}\left\{\ln(\zeta_t)\right\} \doteq 0.26. \tag{33}$$

4.2 Dynamics of individual scales

In Eq. 26 we should thus substitute x_t for L_t and at the same time assume that $\tilde{w}_{j,t}$ and $\tilde{v}_{j,t}$ are the Haar wavelet and scaling coefficients of L_t. To accomplish the multiscale forecast each individual scale has to be forecast (i.e. wavelet and scaling coefficients of individual levels have to be forecast). We may either forecast each scale only from the past dynamics of the scale itself (i.e. independently from the past dynamics of other scales) or we may accomplish the forecast of each scale taking into account not only the dynamics of the scale itself but also the dynamics of other scales. While the former approach is an easier one the latter would presumably result in more accurate forecasts because the dynamics of individual scales are not independent (see e.g. [6]).

If we resort to forecasting each individual scale independently from other scales still many forecasting approaches may be used ranging from linear (e.g. ARMA models) to non-linear forecasting (e.g. neural networks), from parametric to non-parametric etc. Different authors (while not dealing with forecasts of volatility) have proposed different models for the dynamics of (wavelet and scaling) coefficients in general. To give a few examples, Yousefi et al. [22] fit a spline to the scaling coefficients and a sine to the wavelet coefficients. Renaud et al. [21] propose to use only a "special" subset of coefficients of each scale to forecast the next coefficient. The dynamics of this subset is modeled as an autoregressive process.

4.3 ARMA models for individual scales

To illustrate the pitfalls of ARMA modeling of coefficients let MODWT up to level J be applied on the historical time series of L_t. Let $\tilde{w}_{j,t}$, $j = 1, \ldots, J$ be the jth level wavelet coefficients and $\tilde{v}_{J,t}$ the Jth level scaling coefficients. If L_t is stationary then $\tilde{w}_{j,t}$, $j = 1, \ldots, J$ and $\tilde{v}_{J,t}$ are also stationary (see e.g. [3]). Moreover, the wavelet coefficients have been obtained by applying a band-pass filter (passing approximately the band of frequencies $\frac{1}{2^{j+1}} < |f| \leq \frac{1}{2^j}$) onto the time series L_t. Similarly, the scaling coefficients have been obtained by applying a low-pass filter (passing approximately the band of frequencies $0 \leq |f| \leq \frac{1}{2^{j+1}}$) onto the time series L_t. Thus the dynamics of the coefficients is different from the dynamics of the traditional full-band time series.

To truly explore the dynamics of MODWT coefficients we took 28 components[3] of the Dow Jones Index (DJI)[4] in the period from January 1, 1995 till August 31, 2010. In Fig. 1 the average autocorrelation sequence (ACF) and the average partial autocorrelation sequence (PACF) of the MODWT coefficients is plotted for individual scales. The deviations of ACF and PACF of individual components of DJI from the average shape are very small (and thus are not plotted in the figure). We see that wavelet coefficients seem to be stationary and have short memory only. Moreover, coefficients $\tilde{w}_{j,t}$ and $\tilde{w}_{j,t+\tau}$ for $\tau \geq 2^j$ are not correlated. This is

[3]Alcoa, American Express, Boeing, Bank of America, Caterpillar, Chevron, EI DuPont de Nemours, Walt Disney, General Electric, Cisco, Home Depot, Hewlett-Packard, IBM, Intel, Johnson&Johnson, JP Morgan Chase, Coca-Cola, McDonald's, 3M, Merck, Microsoft, Pfizer, Procter & Gamble, AT&T, United Technologies, Verizon Communications, Wal-Mart Stores, Exxon Mobil.

[4]Data from http://finance.yahoo.com/

in agreement with the fact that MODWT wavelet coefficients of long-memory processes (which might be the case of volatility) downsampled by a factor of 2^j are approximately uncorrelated (see e.g. [19]). It might be thus possible to model the jth level MODWT wavelet coefficients as an MA($2^j - 1$) process. On the other hand the scaling coefficients have long memory and could be modeled for example as an ARFIMA process.

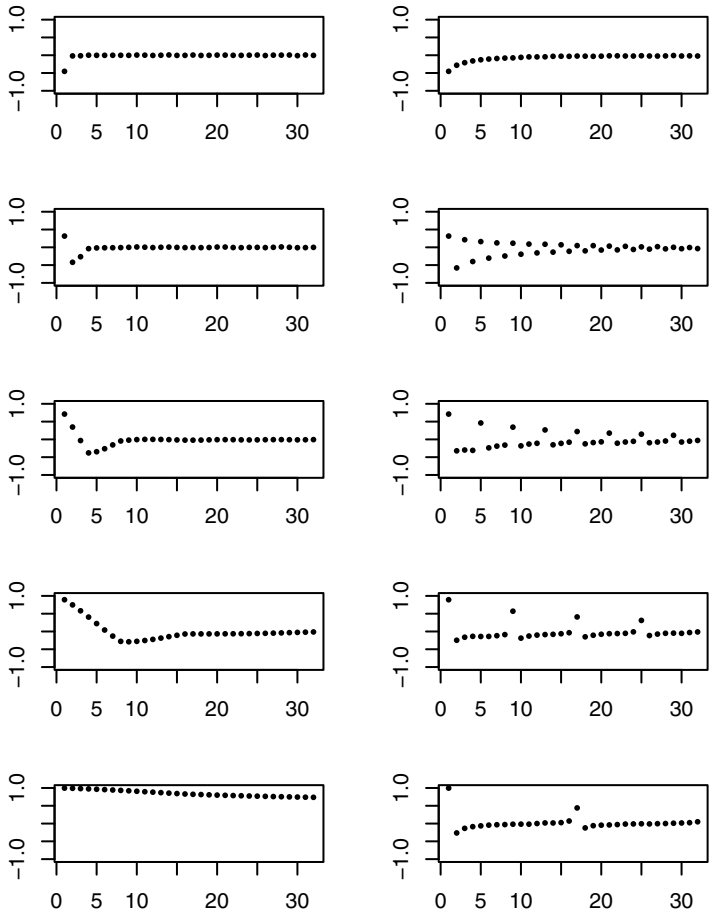


Figure 1: The average ACF (first column) and the average PACF (second column) for the MODWT coefficients of the log Garman-Klass estimate of 28 components of DJI for levels $j = 1, \ldots, 4$ of wavelet coefficients (row 1 to 4) and the 4th level of scaling coefficients (row 5).

4.4 Evaluation of the forecast

When the model for each individual scale is identified (and estimated) the forecast of the each individual scale can be carried out. Then, the h-step ahead forecast of L_t is given as

$$L_t[h] = \tilde{v}_{J,t}[h] + \sum_{j=1}^{J} \tilde{w}_{j,t}[h], \tag{34}$$

Traditional forecast measures such as MAE, MAPE, MSE, Theil's U-statistic etc. may be used to assess the forecast accuracy. In the evaluation of the forecast accuracy L_t may be used to proxy the true H_t (i.e. the forecast error is given as $L_{t+h} - L_t[h]$). Different classes of models may be studied to assess which one has the greatest prediction power:

- Multiscale models based on the mathematical identity of Eq. 34.
- Multiscale models such as Market-component ARCH (see [12]), HARCH (see [15]) etc. rewritten for the *logarithmic* volatility.
- Classical models of log volatility such as the log-GARCH(p,q) model of [9], [18], [13], the EGARCH model of [17], long memory stochastic volatility model of [5] etc.

5 Conclusion and discussion

In this paper we have argued that forecasting volatility based on the multiscale approach might be suitable and might capture the observed multiscale properties of volatility, whereas the traditional models of volatility are not capable of capturing these properties (see e.g. [12]). We have proposed a multiscale approach to modeling volatility that is based on the mathematical identity of Eq. 25. To avoid positivity constraints we propose to work with the logarithm of volatility and to use the logarithm of the Garman-Klass estimator as its proxy. Still many questions remain open: "Should we forecast each scale independently or all scales simultaneously?", "What model for the scale dynamics should be used?" etc. We have also demonstrated (on simple ARMA models) that series of wavelet and scaling coefficients should be treated in a little bit different way than traditional full-band time series. Further comprehensive analysis and research (which is being pursued by the author) is required to quantitatively assess whether multiscale models of volatility provide any substantial benefit for forecasts of volatility.

Acknowledgement

This paper was supported by the means of the institutional support of the long-term conceptual advancement of science and research at the Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic.

References

[1] ARNEODO A., MUZY J.-F., SORNETTE D.: *Causal cascade in the stock market from the Infrared to the Ultraviolet*, In European Physical Journal, Vol. B 2, pp. 277-282, 1998.

[2] BAILLIE, R. T., BOLLERSLEV, T., MIKKELSEN, H.-O.: *Fractionally integrated generalized autoregressive conditional heteroskedasticity*. In Journal of Econometrics, Vol. 74, pp. 3-30, 1996.

[3] BAŠTA, M.: *Waveletová transformace a její aplikace při analýze ekonomických a finančních časových řad*. Ph.D. dissertation thesis. Faculty of Informatics and Statistics, University of Economics, Prague, 2010.

[4] BOLLERSLEV, T.: *Generalized autoregressive conditional heteroskedasticity*. In Journal of Econometrics, 31, pp. 307-327, 1986.

[5] BREIDT, F. J., CRATO, N., De LIMA, P.: *The detection and estimation of long memory in stochastic volatility*. In Journal of Econometrics, Vol. 83, pp. 325-348, 1998.

[6] CROUSE, M. S., NOWAK, R. D., BARANIUK, R. G.: *Wavelet-based statistical signal processing using hidden Markov models*. In IEEE Transactions on Signal Processing, Vol. 46, pp. 886-902, 1998.

[7] ENGLE, R. F.: *Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation*. In Econometrica, Vol. 50, pp. 987 - 1007, 1982.

[8] GARMAN, M. B., KLASS, M. J.: *On the estimation of security price volatilities from historical data*. In The Journal of Business, Vol. 53, pp. 67-78, 1980.

[9] GEWEKE, J.: *Modeling the persistence of conditional variances: A comment* In Econometric Review, Vol. 5, pp. 57-61, 1986.

[10] GRANGER, C. W. J., DING, Z.: *Varieties of long memory models*. In Journal of Econometrics, Vol. 73, pp. 61-77, 1996.

[11] LOBATO, I. N., SAVIN, N. E.: *Real and spurious long-memory properties of stock-market data*. In Journal of Business & Economic Statistics, Vol. 16, pp. 261-268, 1998.

[12] LYNCH P. E., ZUMBACH G. O.: *Market heterogeneities and the causal structure of volatility*. In Quantitative Finance, Vol. 3, pp. 320-331, 2003.

[13] MILHOJ, A.: *A multiplicative parametrization of ARCH models*, working paper, Department of Statistics, University of Copenhagen, 1987.

[14] MOLNÁR, P.: *Properties of range-based volatility estimators*, working paper, Norwegian School of Economics and Business Administration, 2010.

[15] MÜLLER, U., DACOROGNA, M., DAVÉ, R., OLSEN, R., PICTET, O., von WEIZSÄCKER, J.: *Volatilities of different time resolutions, analyzing the dynamics of market components*. In Journal of Empirical Finance, Vol. 4, pp. 213-239, 1997.

[16] MÜLLER, U., DACOROGNA, M., DAVÉ, R., PICTET, O., OLSEN, R., WARD, J.: *Fractals and intrinsic time: A challenge to econometricians*. In Proc. of the 39th Intern. Conf. of the Applied Econometrics Assoc. on Real Time Econometrics, Luxembourg, 1993.

[17] NELSON, D. B.: *Conditional heteroskedasticity in asset returns: A new approach*. In Econometrica, Vol. 59, pp. 347-370, 1991.

[18] PANTULA, S. G.: *Modeling the persistence of conditional variances: A comment*. In Econometric Review, Vol. 5, pp. 71-74, 1986.

[19] PERCIVAL, D. B., WALDEN, A. T.: *Wavelet Methods for Time Series Analysis*. Cambridge University Press. ISBN 9780521640687, 2002 (reprint).

[20] POON, S.-H., GRANGER, C.: *Forecasting volatility in financial markets: A review*. In Journal of Economic Literature, Vol. 41, pp. 478-539, 2003.

[21] RENAUD O., STARCK J. L., MURTAGH F.: *Wavelet-based combined signal filtering and prediction*. In IEEE Transactions on Systems, Man, and Cybernetics, B - Cybernetics, Vol. 35, pp. 1241-1251, 2005.

[22] YOUSEFI, S., WEINREICH, I., REINARZ, D.: *Wavelet-based prediction of oil prices*. In Chaos, Solitons & Fractals, Vol. 25, pp. 265-275, 2005.

[23] ZUMBACH G., LYNCH, P.: *Heterogeneous volatility cascade in financial markets*. In Physica A: Statistical Mechanics and its Applications, Vol. 298, pp. 521-529, 2001.

Current address

Mgr. Milan Bašta, Ph.D.
Faculty of Informatics and Statistics, University of Economics,
sq. W. Churchill 4, Prague, Czech Republic, 130 67,
email: milan.basta@vse.cz

Journal of Applied Mathematics

volume IV (2011), number III

USE OF THE L-MOMENT METHOD IN MODELING THE WAGE DISTRIBUTION

BÍLKOVÁ Diana (CZ)

Abstract. L-moments are based on the linear combinations of order statistics. The question of L-moments presents a general theory covering the summarization and description of sample data sets, the summarization and description of theoretical distributions, but also the estimation of parameters of probability distributions and hypothesis testing for parameters of probability distributions. L-moments can be defined for any random variable in the case that it's mean exists. Within the scope of modelling of wage distributions we currently use the method of conventional moments, the quantile method or the maximum likelihood method. The theory of L-moments parallels the other theories and the main advantage of the method of L-moments over these methods is that L-moments suffer less from impact of sampling variability. L-moments are more robust and they provide more secure results in the case of small samples.
The three-parametric lognormal distribution is one from the most frequent used distributions within the frame of modelling wage and income distribution. In the case of wage distributions we usually work with very large data sets and in such cases the method of L-moments provides say about alike accurate results as for example the method of moment or quantile method. The question of fitness of concrete parametric distribution for model of wage distribution tends to rejection of tested hypothesis about supposed form distribution practically always in the cases of such large samples.
In this connection we can see close relationship between sample size and the value of criterion χ^2, too. The forecasts of wage distributions were constructed based on the observance of previous development. Within the frame of the financilal crisis were set free the employees with very low wages above all. The effect of this truth to forecasts of wage distribution will be exactly known in the autumn of this year.

Key words. L-moments, linear combination of order statistics, wage distributions, lognormal distribution

Mathematics Subject Classification: Primary 62P07, 62E27; Secondary 97M06

1 Introduction

In the capitalistic economy, the attention of economists is attracted by methods of forecasting the population consumption and correspondingly the demand for goods and services. It however needs to be noted that the forecast of the population demand for goods and services is not the only goal of the population income analysis. It is possible to use the knowledge of the population income distributions for instance in analysis of the standard of living or in its interregional or international comparison. It would on the other hand not be fully correct to state that the wage distribution or differentiation (reflected by some volatility measure) analysis are itself sufficient for the standard of living analysis. Detailed analysis of the wage component of the standard of living requires the knowledge of the full wage distribution in the particular time of interest, especially the knowledge of the number of units with the income below certain threshold.
In the field of statistics, it is possible to encounter another indirect usage of the knowledge of the wage distribution. Namely, it is the improvement of the statistical sampling methodology when researching a variable which is highly correlated with the wages. For example we can name the expenditures of an individual or household, equipment of the households, time usage, buying behavior, or variables representing opinions in the case of sociological researches.
Common statistical methodology for description of the statistical samples is based on using conventional moments or cummulants. Also when fitting with an appropriate parametrical distribution for the given data sample, the moment method is often used. This method is based on setting the conventional sample moments equal to the corresponding moments of the theoretical distribution. Nevertheless the moment method is not always appropriate, especially in the case of small samples. It can be found in the statistical literature [2] that the moment method of the parameter estimates is often less accurate than other methods such as the maximum likelihood method.
An alternative approach is based on using different characteristics which are called the L-moments. The L-moments are an analogy to the conventional moments, but are based on linear combinations of the rank statistics, i.e. the L-statistics. Using the L-moments s theoretically more appropriate than the conventional moments because the L-moments characterize wider range of distributions. When estimating from a sample, L-moments are more robust to the existence of the outliers in the data. The experience shows that in comparison with the conventional moments are the L-moments more difficult to distort and in finite samples converges faster to the asymptotical normal distribution. The parameter estimates using the L-moments are especially in the case of small samples are often even more precise than estimates calculated using the maximum likelihood method.
This text concerns with the application of the L-moments in the case of larger samples and with the comparison of the precision of the method of the L-moments with the precision of other methods (moment and quantile method) of the parameter estimates in the case of larger samples. Based on these analysis using three parametric lognormal distribution, predictions were calculated for the period of the financial crisis, which occurred in the fall 2008, but assuming that the recent trend will continue. The wage distributions are analyzed by sectors of economy. The data used were collected and published by the Czech statistical office. Namely the shares of the employees in ranges of the monthly gross wages by industry in the period since 2002 to 2008 and the corresponding sample sizes. There were 84 datasets of the monthly gross wages in the form of the interval frequency distributions with opened lowest and highest interval. Data were analyzed with MS Excel and specialized statistical software SAS and Statgraphics.

2 Methodology

2.1 L-moments of the probability distributions

We will assume that X is a real random variable with the distribution function $F(x)$ and quantile function $x(F)$ and $X_{1:n} \le X_{2:n} \le \ldots \le X_{n:n}$ are the rank statistics of the random sample of the size n selected from the distribution X. Then the r-th L-moment of the random variable X is defined as

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k:r}, \quad r = 1, 2, 3, \ldots. \tag{1}$$

The letter 'L' in the name 'L-moments' is to stress the fact that r-th L-moment λ_r is a linear function of the expected rank statistics. Natural estimate of the L-moment λ_r based on the observed sample is furthermore a linear combination of the ordered values, i.e. the so called L-statistics. The expected value of the rank statistic is of the form

$$EX_{j:r} = \frac{r!}{(j-1)! \cdot (r-j)!} \cdot \int x\,[F(x)]^{j-1} \cdot [1 - F(x)]^{r-j}\, \mathrm{d}F(x). \tag{2}$$

If we plough the equation (2) in the equation (1), we get after some operations

$$\lambda_r = \int_0^1 x(F)\, P^*_{r-1}(F)\, \mathrm{d}F, \quad r = 1, 2, 3, \ldots, \tag{3}$$

where

$$P^*_r(F) = \sum_{k=0}^{r} p^*_{r,k}\, F^k, \tag{4}$$

and

$$p^*_{r,k} = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k}. \tag{5}$$

The symbol Where $P^*_r(F)$ represents r-th shifted Legender's polynom představuje which is related to the usual Legender's polynoms. Shifted Legender's polynoms are orthogonal on the interval (0,1) with a constant weight function. First four Legender's polynoms are of the form

$$\lambda_1 = EX = \int_0^1 x(F)\, \mathrm{d}F, \tag{6}$$

$$\lambda_2 = \frac{1}{2} E(X_{2:2} - X_{1:2}) = \int_0^1 x(F) \cdot (2F - 1)\, \mathrm{d}F, \tag{7}$$

$$\lambda_3 = \frac{1}{3} E(X_{3:3} - 2\, X_{2:3} + X_{1:3}) = \int_0^1 x(F) \cdot (6F^2 - 6F + 1)\, \mathrm{d}F, \tag{8}$$

$$\lambda_4 = \frac{1}{4} E(X_{4:4} - 3\ X_{3:4} + 3\ X_{2:4} - X_{1:4}) = \int_0^1 x(F) \cdot (20F^3 - 30F^2 + 12F\ - 1)\, \mathrm{d}F. \tag{9}$$

Details about the L-moments can be found in [3] or [4]. The coefficients of the L-moments are defined as

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \quad r = 3, 4, 5, ... \tag{10}$$

L-moments $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_r$ and coefficients L-moments $\tau_1, \tau_2, \tau_3, \ldots, \tau_r$ can be used as the characteristics of the distribution. L-moments are in a way similar to the conventional central moments and coefficients of L-moments are similar to the moment ratios. Especially L-moments λ_1 and λ_2 and coefficients of the L-moments τ_3 and τ_4 are considered to be characteristics of the location, variability and skewness.
Using the equations (6) to (8) and the equation (10), we get the first three L-moments of the three parametric lognormal distribution LN(μ, σ^2, ξ), which is described e.g. in [1]. The following relations are valid for these L-moments

$$\lambda_1 = \xi + \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{11}$$

$$\lambda_2 = \exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \mathrm{erf}\left(\frac{\sigma}{2}\right), \tag{12}$$

$$\tau_3 = \frac{6\pi^{-1/2}}{\mathrm{erf}\left(\frac{\sigma}{2}\right)} \cdot \int_0^{\sigma/2} \mathrm{erf}\left(\frac{x}{\sqrt{3}}\right) \cdot \exp(-x^2)\, \mathrm{d}x, \tag{13}$$

where erf(z) is the so called error function defined as

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \cdot \int_0^z e^{-t^2}\, \mathrm{d}t. \tag{14}$$

2.2 Sample L-moments

We will assume that $x_1, x_2, \ldots, x_n$ is a random sample and $x_{1:n} \le x_{2:n} \le \ldots \le x_{n:n}$ is the ordered sample. The r-th sample L-moment is defined as

$$l_r = \binom{n}{r}^{-1} \cdot \sum_{1 \le i_1 \le i_2 \le \ldots \le i_r \le n} \sum \sum \ldots \sum r^{-1} \cdot \sum_{k=0}^{r-1} (-1)^k \cdot \binom{r-1}{k} \cdot x_{i_{r-k:n}}, \quad r = 1, 2, ..., n. \tag{15}$$

We can write specifically for the first four sample L-moments

$$l_1 = n^{-1} \cdot \sum_i x_i, \tag{16}$$

$$l_2 = \frac{1}{2}\cdot\binom{n}{2}^{-1}\cdot\sum_{i>j}\sum (x_{i:n} - x_{j:n}), \tag{17}$$

$$l_3 = \frac{1}{3}\cdot\binom{n}{3}^{-1}\cdot\sum_{i>j>k}\sum\sum (x_{i:n} - 2x_{j:n} + x_{k:n}), \tag{18}$$

$$l_4 = \frac{1}{4}\cdot\binom{n}{4}^{-1}\cdot\sum_{i>j>k>l}\sum\sum\sum (x_{i:n} - 3x_{j:n} + 3x_{k:n} - x_{l:n}). \tag{19}$$

Sample L-moments can be used similarly as the conventional sample L-moments because they characterize basic properties of the sample distribution and estimates the corresponding properties of the distribution from which were the data sampled. They might be also used to estimate the the parameters of this distribution. In these cases, L-moments are of then used instead of the conventional moments because as linear functions of the data are less sensitive on the sample variability or the error size in the case of the presence of the extreme values in the data than the conventional moments. Therefore it is assumed that the L-moments provide more precise and robust estimates of the characteristics or parameters of the population probability distribution.

2.3 Parameter estimates

Let us denote the distribution function of the standard normal distribution as Φ, then Φ^{-1} represents the quantile function of the standard normal distribution. The following relation holds for the distribution function of the three parametric lognormal distribution LN(μ, σ^2, ξ)

$$F = \Phi\left[\frac{\ln(x-\xi)-\mu}{\sigma}\right]. \tag{20}$$

The coefficients of the L-moments (10) are then commonly estimated using the following estimates

$$t_r = \frac{l_r}{l_2}, \quad r = 3, 4, 5, ... \tag{21}$$

The estimates of the three parametric lognormal distribution can then be calculated as

$$z = \sqrt{\frac{8}{3}}\cdot\Phi^{-1}\left(\frac{1+t_3}{2}\right), \tag{22}$$

$$\hat{\sigma} \approx 0{,}999\,281 z - 0{,}006\,118\,z^3 + 0{,}000\,127\,z^5, \tag{23}$$

$$\hat{\mu} = \ln\left[\frac{l_2}{\operatorname{erf}\left(\frac{\sigma}{2}\right)}\right] - \frac{\hat{\sigma}^2}{2}, \tag{24}$$

$$\xi = l_1 - \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right). \tag{25}$$

2.4 Appropriateness of the model

When judging the appropriateness of the constructed model, it is necessary to take an advantage of some criterion which could be for example the sum of the absolute differences of the observed and theoretical frequencies for all intervals

$$S = \sum_{i=1}^{k} \left| n_i - n \cdot \pi_i \right| \tag{26}$$

or the popular criterion χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot \pi_i)^2}{n \cdot \pi_i}, \tag{27}$$

where n_i are the observed frequencies in the particular intervals, π_i are the theoretical probabilities n the particular intervals and $n.\pi_i$ are theoretical frequencies in the intervals $i = 1, 2, \ldots, k$.
The question of the appropriateness of the given curve as a model was described for example in [1]. The problem is that for large samples, which are common in the case of wage distributions, is for a given significance level is the power of the test so high that the test uncovers even the smallest differences of the observed distribution from the theoretical distribution. The test results in almost every case it the rejection of the tested hypothesis about the tested distribution. From practical point of view however, negligible differences are not important and a approximate correspondence of the model with the reality. In these cases, we only 'borrow' the model distribution. The criterion χ^2 is used only for indication and the most important is the logical analysis and experience.

3 Outputs

The first sample L-moments for the wage distributions in the period 2002 – 2008 by industry segments, parameter estimates of the three parametric lognormal distribution using the method of the L-moments for each of these distributions, prediction of the first three L-moments for the years 2009 and 2010 based on the assumption of the continuing trend in the wage distributions and parameter estimates of the three parametric log normal distribution based on these predictions, were computed. Calculated values for the education, health and financial intermediation are presented in Tables 1 – 3.

Tab 1: Sample L-moments and estimated parameters of the three-parametric lognormal distribution using the L-moment method for the „Education“

Rok	Sample L-moments			Estimated parameters		
	l_1	l_2	l_3	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\varsigma}$
2002	16 386,38	3 439,67	622,48	9,643 557	0,139 152	– 146,938 7
2003	18 058,37	3 612,52	553,68	9,877 577	0,099 364	– 2 422,762 8
2004	18 572,98	4 005,83	165,42	11,336 010	0,007 139	– 65 511,643 2
2005	19 986,45	4 452,03	330,59	10,847 275	0,023 124	– 32 005,152 6
2006	21 287,43	4 506,65	560,78	10,323 292	0,065 244	– 10 155,013 7
2007	22 807,50	5 282,73	795,83	10,276 473	0,095 959	– 7 661,113 6
2008	23 572,45	5 439,92	981,61	10,105 232	0,138 335	– 26 50,734 2
2009	**24 920,40**	**5 978,94**	**1 498,13**	**9,808 665**	**0,270 734**	**4 092,847 8**
2010	**26 126,50**	**6 479,92**	**2 052,26**	**9,574 152**	**0,440 836**	**8 190,418 3**

Source: own research

Tab 2: Sample L-moments and estimated parameters of the three-parametric lognormal distribution using the L-moment method for the „Health“

Rok	Sample L-moments			Estimated parameters		
	l_1	l_2	l_3	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\varsigma}$
2002	16 596,18	3 820,22	1 158,81	9,107 016	0,402 672	5 566,413 5
2003	17 919,29	4 024,26	1 144,10	9,248 044	0,351 712	5 538,512 0
2004	18 493,39	4 322,67	1 586,68	8,944 760	0,602 922	8 128,071 3
2005	19 465,65	4 566,45	1 617,91	9,055 873	0,559 007	8 133,748 1
2006	20 995,56	4 948,78	1 659,84	9,220 306	0,497 544	8 042,630 4
2007	22 223,22	5 590,14	2 181,30	9,100 374	0,687 796	9 587,627 8
2008	23 417,13	5 932,97	2 385,35	9,108 038	0,733 925	10 387,246 5
2009	**24 780,30**	**6 569,63**	**2 672,97**	**9,188 873**	**0,753 225**	**10516,425 0**
2010	**26 228,40**	**7 217,97**	**3 029,03**	**9,226 932**	**0,805 984**	**11015,122 1**

Source: own research

Tab 3: Sample L-moments and estimated parameters of the three-parametric lognormal distribution using the L-moment method for the ”Financial intermediation“

Rok	Sample L-moments			Estimated parameters		
	l_1	l_2	l_3	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\varsigma}$
2002	27 071,61	6 269,51	1 274,60	10,109 901	0,176 348	220,161 2
2003	28 866,04	6 070,04	1 018,58	10,296 447	0,119 376	– 2 583,336 5
2004	31 304,60	7 538,02	1 926,61	10,015 190	0,282 043	5 554,051 9
2005	32 510,71	7 542,84	1 762,21	10,128 246	0,234 407	4 356,585 2
2006	33 722,95	7 734,91	1 643,30	10,268 057	0,192 940	2 008,516 8
2007	39 245,79	11 529,78	4 098,97	9,976 728	0,563 082	10 728,402 3
2008	40 796,58	11 651,34	3 823,41	10,109 117	0,475 072	9 643,984 5
2009	**43 455,80**	**14 173,20**	**4 580,78**	**10,327 378**	**0,460 073**	**4 994,253 2**
2010	**46 538,70**	**16 605,50**	**5 659,51**	**10,406 706**	**0,514 733**	**3 748,150 4**

Source: own research

The estimated value of the parameter ξ is in many cases negative. That means that the lognormal curve starts in negative values. Because however, in the lower tail, the curve is very close to the horizontal axes, the negative values if the parameter ξ do not have to distort the correspondence of

the model with the reality. In such cases, the parameter ξ should however not have any real interpretation.
The values of the test criterion (27) result in all cases of the wage distributions to the rejection of the tested hypothesis of the assumed three parametric lognormal distribution, as can be in the case of the wage and income distribution expected. It was obvious during the calculation of the sum of the absolute differences (26) and the test criterion (27) that most of the inaccuracies are in both tails of the distributions. If we excluded the lowest and highest (opened) intervals, the method of L-moments would result in much more precise results. The preciseness of the method of L-moments was here compared to the method of moments and to the quantile method. It can be stated that in the case of such large samples, the method of L-moments is as accurate as the method of moments and the quantile method.
The predictions of the wage distributions for the years 2009 and 2010 assuming continuing trend can be found in the table 4 and 5.

4 Conclusion

One of the most often distributions applied for the modeling of the wage and income distributions is the lognormal distribution. Most often the three parametric lognormal distribution. To estimate the parameters of this distribution, several different methods can be used. E.g. the method of moments, the quantile method, Kemsley's or Cohen's method etc. One of the possibilities of how to estimate parameters of this distribution is using the method of the L-moments.
The method of L-moments commonly leads, in the case of small samples, in much more accurate results then other methods (including the maximum likelihood method). It was shown that in the case of large smples, the method of L-moments leads to similarsly accurate results as the method of moments or the quantile method.
When solving the question which method of the parameter estimates of the three parametric lognormal distribution is the most accurate, the dependence of the criterion χ^2 on the sample size was apparent. As is common for the samples of such size, all tests resulted in rejection of the null hypothesis about the assumed distribution (see above). The predictions of the wage distributions for the particular segments of the industry for the years 2009 and 2010 were constructed assuming that the trend will continue. The question however remains what will be the impact of financial crisis on the distribution of the wages (both the location and variability). During the financial crisis, many employees with low wages lost their jobs. This fact can have impact on the location of the wage distribution and for sure will have an impact on the wage differentiation.

References

[1.] BÍLKOVÁ, D.: *Modelování mzdových rozdělení v České republice v letech 2004 a 2005 s využitím logaritmicko-normálních křivek a křivek Pearsonova a Johnsonova systému*. In Statistika, Vol. 88, No. 2, pp. 149 − 166, 2008.
[2.] HÁTLE, J., HUSTOPECKÝ, J., NOVÁK.: *Modelování a krátkodobá předpověď příjmových rozdělení. Výzkumná práce č. 66.* Výzkumný ústav sociálně ekonomických informací a Vysoká škola ekonomická, Prague, 1975.
[3.] HOSKING, J. R. M.: *L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics*. In Journal of the Royal Statistical Society (Series B), Vol. 52, No. 1, pp. 105 – 124, 1990.

[4.] HOSKING, J. R. M., WALES, J. R.: *Regional frequency analysis: An Approach Based on L-moments.* Cambridge University Press, New York, 1997.

Current address

Diana Bílková, Ing. Dr.
Prague University od Economics
Faculty of Informatics and Statistics
Department of Statistics and probability
Sq. W. Churchill 4
130 67 Prague 3
phone: +420 224 095 484
e-mail: bilkova@vse.cz

Table 4: Estimated Percentage of employees by the band of gross monthly wages by sectors of economy in 2009 (in %)

Wage bounds	Odvětví											
	Agricult.	Industry	Building industry	Trade, repairs	Accomm., catering	Traffin, storage	Financial Intermed.	Real est., renting	Public Administ.	Education	Health service	Other
0 – 4 000	0,00	0,00	0,00	0,00	3,52	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4 001 – 8 000	0,00	0,00	0,00	0,00	11,84	0,00	0,00	0,00	0,00	0,00	0,00	0,00
8 001 – 12 000	0,00	0,00	0,00	0,00	17,28	0,00	1,46	13,29	0,00	5,31	0,00	0,00
12 001 – 16 000	21,28	16,46	13,66	0,00	17,3	0,00	5,11	17,69	7,17	15,3	23,74	26,96
16 001 – 20 000	24,85	24,97	18,52	24,22	14,4	25,14	8,11	14,41	17,43	19,06	23,33	22,12
20 001 – 24 000	18,37	17,84	16,71	14,62	10,87	20,12	9,47	10,98	19,46	17,05	15,85	14,26
24 001 – 28 000	11,91	11,84	13,06	9,36	7,78	13,64	9,59	8,33	16,31	13,15	10,41	9,32
28 001 – 32 000	7,44	7,91	9,65	6,36	5,41	9,17	8,99	6,38	12,12	9,43	6,94	6,31
32 001 – 36 000	4,63	5,41	6,98	4,51	3,7	6,27	8,08	4,95	8,53	6,53	4,74	4,41
36 001 – 40 000	2,90	3,79	5,01	3,32	2,51	4,39	7,08	3,89	5,85	4,45	3,32	3,17
40 001 – 44 000	1,84	2,71	3,61	2,51	1,70	3,15	6,11	3,10	3,98	3,01	2,37	2,34
44 001 – 48 000	1,19	1,98	2,61	1,94	1,15	2,3	5,23	2,50	2,71	2,04	1,73	1,76
48 001 – 52 000	0,78	1,48	1,90	1,53	0,78	1,71	4,45	2,04	1,85	1,38	1,29	1,35
52 001 – 56 000	0,52	1,12	1,39	1,23	0,53	1,29	3,77	1,68	1,27	0,94	0,97	1,05
56 001 – 60 000	0,35	0,86	1,03	1,00	0,37	0,99	3,20	1,39	0,88	0,65	0,74	0,83
60 001 – 64 000	0,24	0,67	0,77	0,82	0,25	0,77	2,71	1,16	0,61	0,45	0,57	0,66
64 001 – 68 000	0,17	0,52	0,58	0,68	0,18	0,61	2,30	0,98	0,43	0,31	0,45	0,54
68 001 – 72 000	0,12	0,42	0,44	0,57	0,12	0,48	1,95	0,83	0,31	0,22	0,36	0,44
72 001 – 76 000	0,08	0,33	0,34	0,49	0,09	0,39	1,66	0,71	0,22	0,16	0,29	0,36
76 001 – 80 000	0,06	0,27	0,26	0,41	0,06	0,31	1,41	0,61	0,16	0,11	0,23	0,30
80 001 – 84 000	0,04	0,22	0,2	0,36	0,04	0,25	1,21	0,53	0,11	0,08	0,19	0,25
84 001 – 88 000	0,03	0,18	0,16	0,31	0,03	0,21	1,03	0,46	0,08	0,06	0,15	0,21
88 001 – 92 000	0,02	0,15	0,12	0,27	0,02	0,17	0,89	0,4	0,06	0,04	0,13	0,18
92 001 – 96 000	0,02	0,12	0,10	0,23	0,02	0,14	0,76	0,35	0,05	0,03	0,10	0,15
96 001 – 100 000	0,01	0,10	0,08	0,20	0,01	0,12	0,66	0,30	0,03	0,02	0,09	0,13
100 001 – 104 000	0,01	0,09	0,06	0,18	0,01	0,10	0,57	0,27	0,03	0,02	0,07	0,11
104 001 – 108 000	0,01	0,07	0,05	0,16	0,01	0,09	0,49	0,24	0,02	0,01	0,06	0,10
108 001 – 112 000	0,01	0,06	0,04	0,14	0,00	0,07	0,43	0,21	0,01	0,01	0,05	0,08
112 001 – 116 000	0.00	0,05	0,03	0,13	0,00	0,06	0,37	0,19	0,01	0,01	0,04	0,07
116 001– ∞	3,12	0,38	2,64	24,45	0,02	8,06	2,91	2,13	0,31	0,17	1,79	2,54
Total	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Average wage	**22 614**	**26 423**	**27 366**	**28 434**	**18 524**	**28 200**	**43 456**	**30 313**	**28 082**	**24 920**	**24 780**	**25 393**
Standard deviation	**9 744**	**15 829**	**14 286**	**26 950**	**11 782**	**16 392**	**29 397**	**29 451**	**11 783**	**11 614**	**15 121**	**18 690**
Coeff. of variation	**0,43**	**0,60**	**0,52**	**0,95**	**0,64**	**0,58**	**0,68**	**0,97**	**0,42**	**0,47**	**0,61**	**0,74**

Source: own research

Table 5: Estimated Percentage of employees by the band of gross monthly wages by sectors of economy in 2010 (in %)

	Sectors of Economy											
Wage bounds	Agricult.	Industry	Building industry	Trade, repairs	Accomm., catering	Traffin, storage	Financial Intermed.	Real est., renting	Public Administ.	Education	Health service	Other
0 – 4 000	0,00	0,00	0,00	0,00	4,72	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4 001 – 8 000	0,00	0,00	0,00	0,00	9,56	0,00	0,00	0,00	0,00	0,00	0,00	0,00
8 001 – 12 000	0,00	0,00	2,46	0,00	13,66	0,00	0,00	14,03	0,00	0,00	0,00	0,00
12 001 – 16 000	14,82	0,00	11,32	0,00	14,82	0,00	0,00	17,17	5,33	15,6	20,9	0,00
16 001 – 20 000	22,43	26,05	16,19	25,51	13,68	23,45	7,78	13,65	18,55	20,44	23,16	23,74
20 001 – 24 000	19,47	19,4	15,74	16,33	11,41	20,95	8,61	10,37	20,23	17,34	16,21	15,41
24 001 – 28 000	14,03	12,87	13,14	10,65	8,92	14,53	8,56	7,92	16,17	12,85	10,88	10,14
28 001 – 32 000	9,42	8,63	10,24	7,31	6,69	9,91	8,04	6,13	11,7	9,10	7,40	6,94
32 001 – 36 000	6,17	5,95	7,74	5,23	4,87	6,88	7,29	4,82	8,18	6,36	5,15	4,92
36 001 – 40 000	4,02	4,21	5,76	3,87	3,48	4,89	6,49	3,84	5,67	4,44	3,67	3,60
40 001 – 44 000	2,63	3,06	4,27	2,94	2,46	3,55	5,70	3,10	3,95	3,12	2,67	2,70
44 001 – 48 000	1,74	2,27	3,17	2,29	1,72	2,63	4,97	2,54	2,77	2,22	1,98	2,07
48 001 – 52 000	1,16	1,71	2,36	1,81	1,20	1,99	4,32	2,10	1,96	1,59	1,49	1,62
52 001 – 56 000	0,78	1,32	1,76	1,46	0,84	1,52	3,74	1,75	1,40	1,15	1,14	1,28
56 001 – 60 000	0,53	1,03	1,33	1,19	0,58	1,19	3,24	1,47	1,01	0,84	0,89	1,03
60 001 – 64 000	0,37	0,81	1,00	0,98	0,41	0,93	2,80	1,25	0,74	0,62	0,70	0,84
64 001 – 68 000	0,26	0,65	0,76	0,82	0,28	0,74	2,43	1,07	0,55	0,47	0,55	0,69
68 001 – 72 000	0,18	0,52	0,59	0,69	0,2	0,6	2,10	0,91	0,41	0,35	0,44	0,57
72 001 – 76 000	0,13	0,42	0,45	0,59	0,14	0,49	1,83	0,79	0,31	0,27	0,36	0,48
76 001 – 80 000	0,09	0,35	0,35	0,50	0,10	0,40	1,59	0,69	0,23	0,20	0,29	0,40
80 001 – 84 000	0,07	0,29	0,27	0,43	0,07	0,33	1,38	0,60	0,18	0,16	0,24	0,34
84 001 – 88 000	0,05	0,24	0,21	0,37	0,05	0,27	1,21	0,53	0,14	0,12	0,2	0,29
88 001 – 92 000	0,04	0,20	0,17	0,33	0,04	0,23	1,06	0,46	0,11	0,10	0,17	0,25
92 001 – 96 000	0,03	0,17	0,13	0,28	0,03	0,19	0,93	0,41	0,08	0,08	0,14	0,22
96 001 – 100 000	0,02	0,14	0,11	0,25	0,02	0,16	0,81	0,36	0,07	0,06	0,12	0,19
100 001 – 104 000	0,02	0,12	0,09	0,22	0,01	0,14	0,72	0,32	0,05	0,05	0,10	0,16
104 001 – 108 000	0,01	0,11	0,07	0,20	0,01	0,12	0,63	0,29	0,04	0,04	0,08	0,14
108 001 – 112 000	0,01	0,09	0,06	0,17	0,01	0,10	0,56	0,26	0,03	0,03	0,07	0,13
112 001 – 116 000	0,01	0,08	0,05	0,16	0,01	0,09	0,49	0,23	0,03	0,02	0,06	0,11
116 001– ∞	1,51	9,31	0,21	15,42	0,01	3,72	12,72	2,94	0,11	2,38	0,94	21,74
Total	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Average wage	**24 804**	**28 475**	**29 182**	**31 024**	**20 566**	**30 184**	**46 539**	**32 381**	**28 800**	**26 127**	**26 228**	**28 211**
Standard deviation	**10 694**	**18 247**	**15 553**	**29 493**	**13 255**	**18 630**	**35 109**	**35 504**	**13 096**	**13 350**	**16 933**	**23 280**
Coeff. of variation	**0,43**	**0,64**	**0,53**	**0,95**	**0,64**	**0,62**	**0,75**	**1,10**	**0,45**	**0,51**	**0,65**	**0,83**

Source: own research

Journal of Applied Mathematics

volume IV (2011), number III

ROBUST FILTERING OF TIME SERIES

BLATNÁ Dagmar, (CZ)

Abstract. Several applications of robust filtering of time series is demonstrated. These methods are increasingly important due to their stability to sudden impulses and outliers. Median smoothers together with moving median absolute deviation were directly used as a part of price indicators in technical analysis. Both Moving Median and Robust Bollinger Bands proved to be useful in forecasting of principal turning point in WIG stock index. Second, Artificial Neural Networks (ANN) represents another tool with 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. Once again, these properties were demonstrated on 10-week forecasts of WIG stock index. The results obtained clearly show, ANN with one or two hidden layers are capable to predict correct direction.

Key words. Robust filtering, technical analysis, artificial neural networks, forecasting in time series

Mathematics Subject Classification: Primary 60A05, 08A72; Secondary 28E10.

1 Introduction to robust filtering

Noise removal and signal extraction are important aspects in time series analysis. Using linear smoothers, e.g. moving averages, the results obtained are influenced by sudden impulses, outliers and noise coming from heavy tailed distributions with high kurtosis. For these reasons, nonlinear smoothers are of increasing importance.

Median smoother related to a value x_t is defined as

$$M_n(t) = med\left(x_{t-n}, x_{t-n+1}, ..., x_t, ..., x_{t+n-1}, x_{t+n}\right) \tag{0.1}$$

with running window of length $2n+1$. Tukey suggested to combine running median smoothers and linear smoothers to improve the properties of running medians [1]. Thus, smoothers named *3RSS, 5RSS, 3RSSH, 5RSSH* and further have been created. The notation used has the following meaning. *R* denotes repeated resmoothing, i.e. one smoother is applied to the results of the same smoother used previously. *S* denotes splitting, where the data sequence is divided into two separate parts, each end is smoothed separately and the parts are joined together. *H* denotes so-called hanning, i.e. the application of running weighted average provided that the weights sum must be one.
LULU smoothers are local nonlinear filters consisting of the suboperators *L* (Low) and *U* (Upper). They are defined as [2]

$$\begin{aligned} L_n(t) &= \max\left[\min\left(x_{t-n},...,x_t\right),...,\min\left(x_t,...,x_{t+n}\right)\right] \qquad n=1,2,... \\ U_n(t) &= \min\left[\max\left(x_{t-n},...,x_t\right),...,\max\left(x_t,...,x_{t+n}\right)\right] \end{aligned} \tag{0.2}$$

Then, we can combine these basic smoothers to define class of *LULU* smoothers. For example, the smoothers C_n (Ceiling) and F_n (Flooring) are recursively defined as

$$\begin{aligned} C_n &= L_n U_n C_{n-1} \qquad & C_1 &= L_1 U_1 \\ F_n &= U_n L_n F_{n-1} \qquad & F_1 &= U_1 L_1 \end{aligned} \tag{0.3}$$

The result of the L_1U_1 application is demonstrated on Fig.1. Clearly, both upper and lower outliers are filtered out.

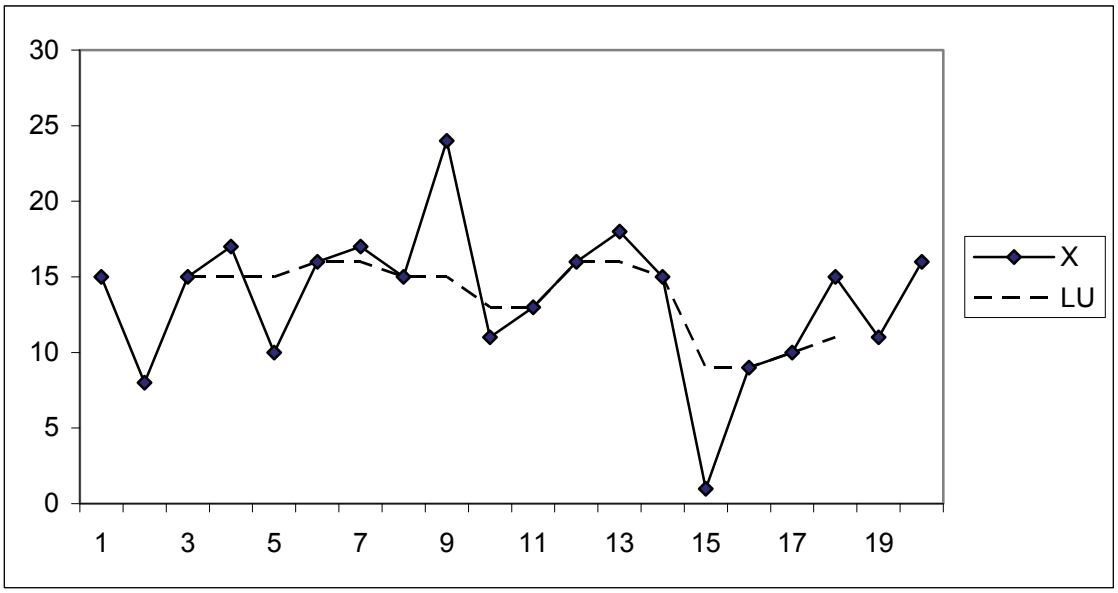


Fig. 1 The time series X filtered by $\mathbf{L_1U_1}$ combination

Similarly, we can introduce *Moving Median Absolute Deviation* as an important robust moving characteristics of scale:

$$\omega_n(t) = med\left[\left|x_{t-n} - M_n(t)\right|,...,\left|x_t - M_n(t)\right|,...,\left|x_{t+n} - M_n(t)\right|\right] \tag{0.4}$$

Both this characteristic and running medians will be used in the following chapter.

2 Price indicators in technical analysis

Technical Analysis is a method of the estimation of stock prices, based on the study of the behaviour of individual stocks and global market [3]. Probable future price development is predicted on the basis of past prices and/or trade volumes.
Technical indicator is a function assuming for each trade day t certain value dependent on past information. We employ here selected price indicators, where only past stock prices are used for generating of trade signals. Because we want to predict trend decrease, only SELL signals will be investigated. Our aim will be the robustification of individual indicators, which may result from the use of moving medians and moving median absolute deviation.

Moving Median (MM) of length n is computed here using the last n price values

$$MM_t = med\left(X_t, X_{t-1}, ..., X_{t-n+1}\right) \tag{2.1}$$

In the basic form, trade signals are given by intersection of the price and moving median and SELL signals are generated on condition

$$\left(X_{t-1} \geq MM_{t-1}\right) \wedge \left(X_t < MM_t\right) \Rightarrow SELL \tag{2.2}$$

However, this version produces rather high number of trade signals. Therefore, more preferable is the use of a combination of two moving averages of different lengths: short-period n_1 and long-period n_2. Then, the formula (2.2) will be modified to

$$\left\{MM_{t-1}\left(n_1\right) \geq MM_{t-1}\left(n_2\right)\right\} \wedge \left\{MM_t\left(n_1\right) < MM_t\left(n_2\right)\right\} \Rightarrow SELL \tag{2.3}$$

The natural choice in our analysis is the selection $n_1 = 5$ (weekly moving medians) and $n_2 = 21$ (monthly moving medians).

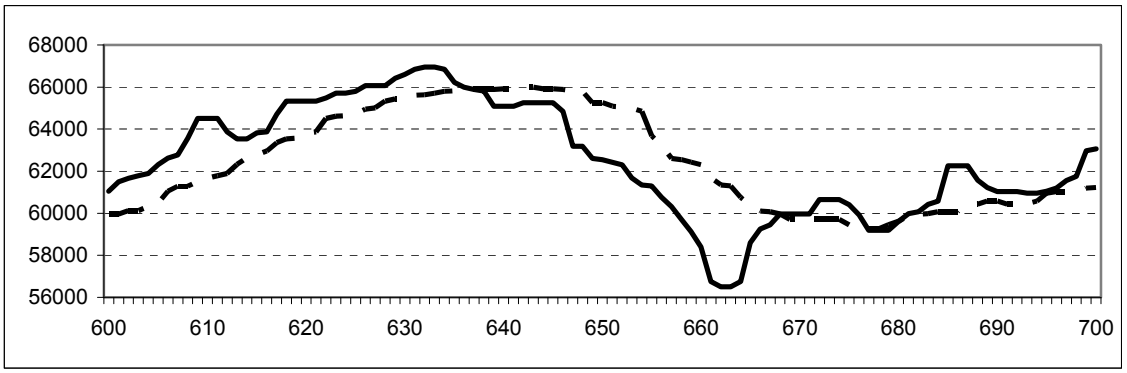


Fig. 2 MM Indicator: intersections of 5-day and 21-day moving medians

Thus, according to our trade system, the first SELL signal occurs at 638, i.e. 7 days after the absolute maximum at 631.
Several types of technical indicators are based on the construction of bands (envelopes) around moving averages. The bandwidth can be either constant or variable in dependence on price volatility. We construct here **Robust Bollinger Bands (RBB)**

$$RBB_t^U = MM_t(n) + c\omega_t(n), \qquad RBB_t^L = MM_t(n) - c\omega_t(n) \tag{2.4}$$

and use bandwidths proportional to median absolute deviation $\omega_t(n)$ with the values $n = 21, c = 1$. The *SELL* signal is generated under condition

$$\left(X_{t-1} > RBB_{t-1}^U\right) \ \wedge \ \left(X_t \le RBB_t^U\right) \ \Rightarrow \ SELL \tag{2.5}$$

This trade system produces SELL signals at time points 611, 613, 627 and 633, i.e. 20, 18 and 4 days before reaching the absolute maximum at 631 (see Fig.3).

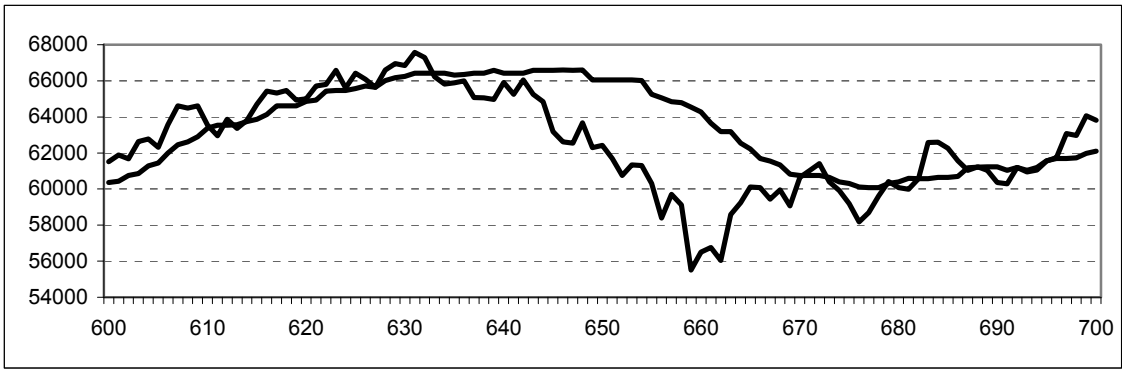


Fig. 3 Robust Upper Bollinger Band with parameters n = 21, c = 1

3 Artificial 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 [4],[5]. Recently, they have been increasingly applied in financial time series analysis as well. 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 [4]

$$\begin{aligned} n_{k,t} &= w_{k,0} + \sum_{i=1}^{I} w_{k,i} x_{i,t} \\ N_{k,t} &= 1/\left[1+\exp\left(-n_{k,t}\right)\right] \\ Y_t &= \gamma_0 + \sum_{k=1}^{K} \gamma_k N_{k,t} + \sum_{i=1}^{I} \beta_i x_{i,t} \end{aligned} \tag{3.1}$$

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.
To demonstrate the capability of ANN from regression point of view, the simulation was performed using the following model

$$X_t = 3\sin\left(t\right) + u_t I\left(Y_t = 0\right) + 5u_t I\left(Y_t = 1\right) \tag{3.2}$$

where the first term on right hand side represents the signal (pure sine wave with amplitude 3) and u_t is $N\left(0,1\right)$ i.i.d., whereas Y_t is Bernoulli i.i.d. with parameter $p = 0.1$. Thus, normal random noise was randomly contaminated. The picture of the whole situation is following

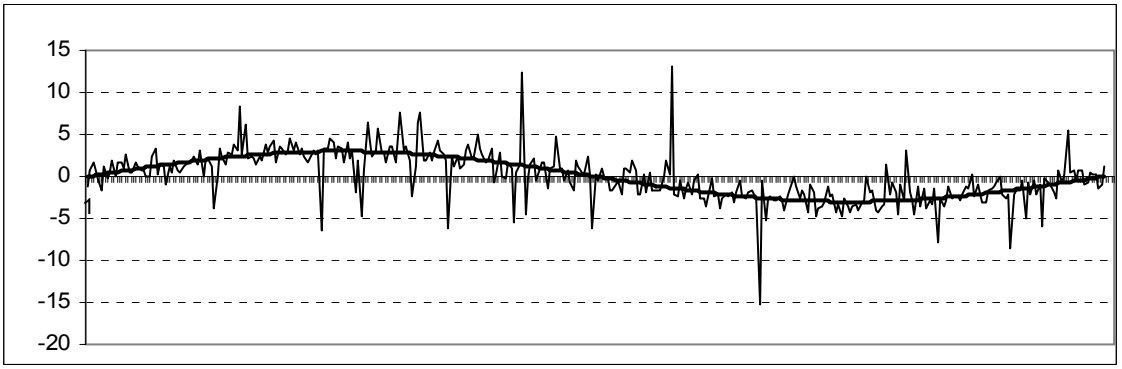


Fig. 4 Sinusoidal signal with contaminated random noise

Using optimal configuration of ANN with two hidden layers and 8 and 6 neurons in them, we obtained predicted values P_t and residuals. Then, we are able to compute individual deviations of predicted values from the signal and mean absolute deviation

$$d_t = 3\sin t - P_t \qquad MAD = \frac{1}{n}\sum_{t=1}^{n}\left|d_t\right| \tag{3.3}$$

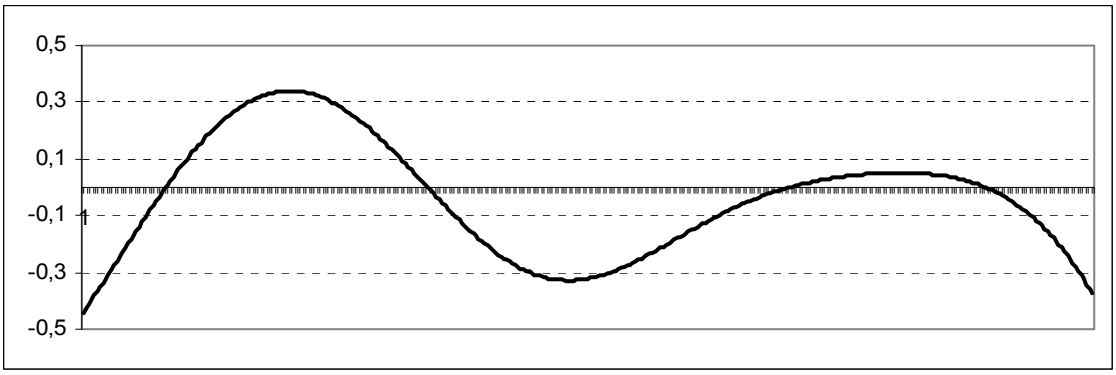


Fig. 5 Time series of individual deviations of predicted values from the signal MAD = 0.165, n = 360

Further, this tool has been applied to weekly closing values of WIG index in Poland during the period 2005-2009, i.e. 252 weekly observations. Obviously, there is sharp peak at the time 127 corresponding with absolute maximum. Thus, ANN was trained at the interval 1-127 with the aim to create 10-week forecast 127-138.

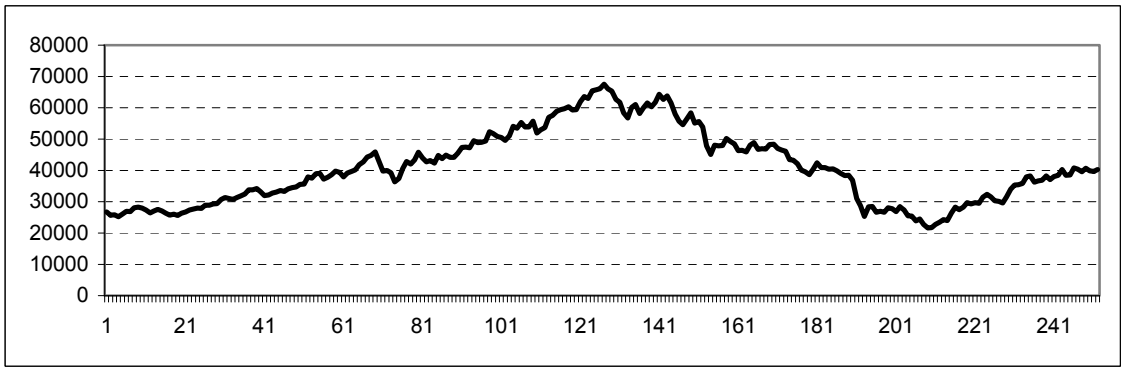


Fig. 6 Time series of weekly closing values of WIG index (Poland)

The results obtained are summarized in Tab.1 and Fig.7. As for ANN type, either 5 or 10 lagged closing values were used for either 1 or 2 hidden layers. For example, 5(4) denotes the model with 5 lagged closing values and 1 hidden layer containing 4 neurons, whereas 10(2,4) denotes the model with 10 lagged closing values and 2 hidden layers containing 2 and 4 neurons. Further, X are true closing values and X_p forecasts. The forecast accuracy has been measured by mean absolute deviation, computed as

$$MAD = \frac{1}{10}\sum\left|X - X_p\right| \tag{3.4}$$

Tab. 1 The results of ANN forecasting: weekly closing values of WIG index (Poland)

ANN type		5(2,3)	5(4)	10(2,4)	5(2,3)	5(4)	10(2,4)
Number	X	Xp	Xp	Xp	AD	AD	AD
128	65989,73	62646,67	63034,94	64142,05	3343,06	2954,79	1847,68
129	65257,29	61615,29	60533,67	63212,76	3642,00	4723,62	2044,53
130	62609,47	60846,77	59109,48	61741,32	1762,70	3499,99	868,15
131	61684,42	60507,00	58850,11	60828,24	1177,42	2834,31	856,18
132	58395,54	60524,90	60287,03	60408,29	2129,36	1891,49	2012,75
133	56759,16	60268,80	60828,28	59766,55	3509,64	4069,12	3007,39
134	60073,46	60144,76	60655,35	60008,91	71,30	581,89	64,55
135	61010,54	60035,70	59957,44	60076,89	974,84	1053,10	933,65
136	58174,13	59962,91	59060,75	59813,00	1788,78	886,62	1638,87
137	59984,91	59931,79	58744,60	59666,58	53,12	1240,31	318,33
		MAD	MAD	MAD			
		1845,221	2373,522	1359,209			

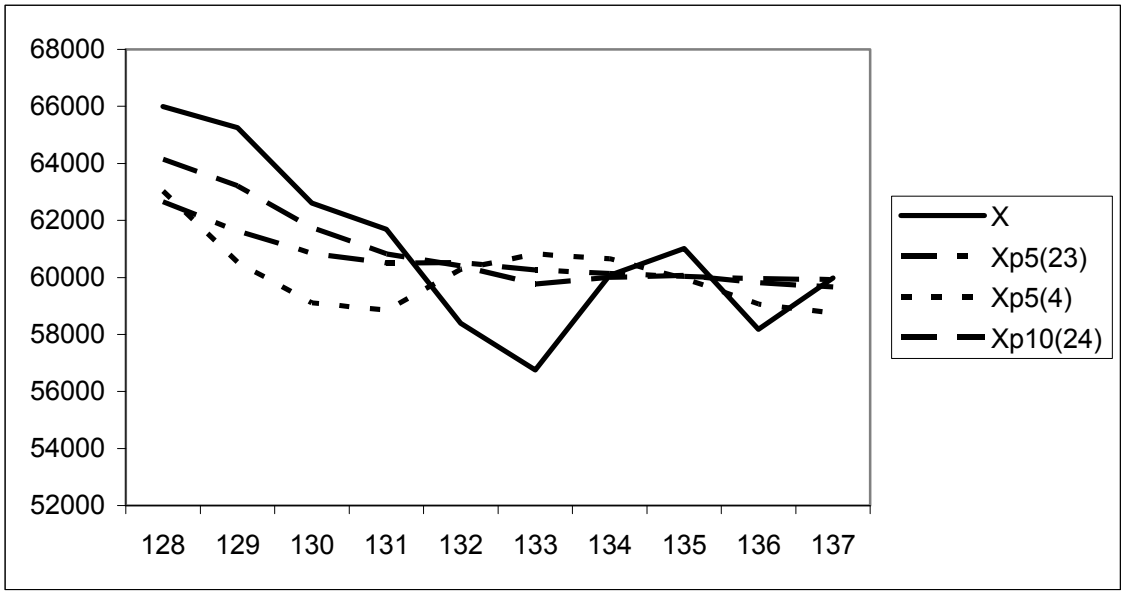


Fig.7 The results of ANN forecasting: weekly closing values of WIG index (Poland)

Acknowledgement

Financial support of this research by grant from institutional support of longtime conceptual scientific and research development at FIS VŠE is gratefully acknowledged.

References

[1] TUKEY, J.W.: *Exploratory Data Analysis.*. Addison-Wesley, Massachussetts, 1971.
[2] CONRADIE, W.J., et al.: *Performance of Nonlinear Smoothers in Signal Recovery*. Appl.Stochastic Models Bus.Ind. Vol.25,pp. 425-444, 2009.
[3] VÍŠKOVÁ, H.: *Technická analýza akcií.* HZ, Praha, 1997.
[4] McNELIS, P.D.: *Neural Networks in Finance*. Elsevier Academic Press, Amsterdam, 2005.
[5] FRANSES, P.H., van DIJK, D.: *Non-Linear Time Series Models in Empirical Finance.*
[6] Cambridge University Press, Cambridge, 2000.
[7] TSAY, R.S.: *Analysis of Financial Time Series.* Wiley, New York, 2002.

Current address:

doc.Ing.Dagmar Blatná, CSc.,
University of Economics Prague,
W.Churchill Sq.4, 13067 Prague, CZ
e-mail: blatna@vse.cz

Journal of Applied Mathematics

volume IV (2011), number III

PRELIMINARY RESULTS ON ASYMMETRIC BAXTER-KING FILTER

BUSS Ginters, (LV)

Abstract. The paper proposes an extension of the symmetric Baxter-King band pass filter to an asymmetric Baxter-King filter. It turns out the optimal correction scheme of the ideal filter weights is the same as in the symmetric version, i.e, cut the ideal filter at the appropriate length and add a constant to all filter weights to ensure zero weight on zero frequency. Since the symmetric Baxter-King filter is unable to extract the band of frequencies at the very ends of the series, the extension to an asymmetric filter is useful whenever the real time estimation is needed. The paper assesses the filter's properties in extracting business cycle frequencies, in comparison to the symmetric Baxter-King filter and symmetric and asymmetric Christiano-Fitzgerald filter, by using Monte Carlo simulation. The results show that the asymmetric Baxter-King filter is superior to the asymmetric Christiano-Fitzgerald filter for the whole sample space, including the very ends of a sample, thus indicating that the asymmetric Baxter-King filter should be preferred over the asymmetric Christiano-Fitzgerald filter in real time signal extraction exercises.

Key words and phrases. real time estimation, Christiano-Fitzgerald filter, Monte Carlo simulation, band pass filter, asymmetric filter.

Mathematics Subject Classification. Primary 60G35; Secondary 62M20.

1 Introduction

This paper considers a simple extension of the symmetric Baxter-King band pass filter (Baxter and King, 1999) to an asymmetric version of it. Such modification, to the best of my knowledge, has not yet been discussed in the literature. Symmetric filters are not applicable at the very ends of an input signal without the extension of the ends with forecasts. Thus, asymmetric band pass filters are necessary to extract the desired band of frequencies at the ends of an input signal, if forecasting is not used for extending the ends of the input signal.

The closest band pass filter to the Baxter-King filter is Christiano-Fitzgerald band pass filter (Christiano and Fitzgerald, 2003) which, in general, is asymmetric, and whose default specification is optimized for an input signal following a random walk (RW) process, but it allows the input signal to follow other data generating processes (DGP). However, Christiano and Fitzgerald (2003) argue that their default specification of the filter is a good approximation to many DGPs observed in macroeconomic time series and, thus, macroeconomists may opt for it. Although Christiano and Fitzgerald (2003) compares their filter to the symmetric Baxter-King filter, they do not elaborate on an asymmetric version of the Baxter-King filter.

Thus, this paper tries to fill the gap in the literature by formally developing an asymmetric version of the Baxter-King filter and assessing its properties in extracting business cycle frequencies, in comparison to the symmetric Baxter-King filter and symmetric and asymmetric default specification of Christiano-Fitzgerald filter, by using Monte Carlo simulation. The results show that the asymmetric Baxter-King filter is superior to the asymmetric Christiano-Fitzgerald filter at the very ends of a sample, thus indicating that the asymmetric Baxter-King filter should be preferred over the asymmetric Christiano-Fitzgerald filter in real time signal extraction exercises.

The paper is organized as follows. Section 2 develops the filter and Section 3 assesses the performance of the filter by means of Monte Carlo simulation.

2 The Asymmetric Baxter-King Filter

Consider the following orthogonal decomposition of the zero-mean covariance stationary stochastic process, x_t:

$$x_t = y_t + \tilde{x}_t. \tag{1}$$

The process, y_t, has power only in frequencies belonging to the interval $\{[a_1, a_2] \cup [-a_2, -a_1]\} \subset (-\pi, \pi)$, where $0 < a_1 < a_2 < \pi$. The process, $\tilde{x}_t$, has power only in the complement of this interval in $(-\pi, \pi)$. By the spectral representation theorem,

$$y_t = b(L)x_t, \tag{2}$$

where the ideal band pass filter, $b(L)$, is

$$b(L) = \sum_{h=-\infty}^{\infty} b_h L^h, \quad L^h x_t = x_{t-h}, \tag{3}$$

where

$$\begin{aligned} b_h &= \frac{\sin(ha_2) - \sin(ha_1)}{\pi h}, \quad h = \pm 1, \pm 2, \ldots \\ b_0 &= \frac{a_2 - a_1}{\pi}, \quad a_1 = \frac{2\pi}{p_u}, \quad a_2 = \frac{2\pi}{p_l}, \end{aligned} \tag{4}$$

and $p_u, p_l \in (2, \infty)$ define the upper and lower bounds of the wave length of interest. With b_h's specified as in (4), the frequency response function of the ideal filter at frequency ω is

$$\begin{aligned} \beta(\omega) &= 1 \quad \text{for } \omega \in [a_1, a_2] \cup [-a_2, -a_1] \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{5}$$

Baxter and King (1999) have proposed to obtain a symmetric, fixed length approximation to the ideal filter, (3) and (4), by minimizing

$$\begin{aligned} Q &= \int_{-\pi}^{\pi} \delta(\omega)\delta(-\omega)d\omega \\ &\text{s.t.} \\ \hat{\beta}(0) &= \sum_{k=-K}^{K} \hat{b}_k = 0 \\ \hat{b}_k &= \hat{b}_{-k}, \end{aligned} \tag{6}$$

where $\delta(\omega) = \beta(\omega) - \hat{\beta}(\omega)$ is the discrepancy between the exact and the approximate filters at frequency ω, and the constraint $\hat{\beta}(0) = 0$ is to ensure zero weight on the trend frequency, in line with the assumption $a_1 > 0$. The solution to (6) is a truncation of the ideal filter symmetrically at length K, and addition of a constant $(-\sum_{k=-K}^{K} b_k)/(2K+1)$ to all filter weights to ensure $\hat{\beta}(0) = 0$. Baxter and King (1999) suggest the value of K to be about 3 years, i.e, K=12 for quarterly data, and K=36 for monthly data. The symmetry of the filter together with the condition $\hat{b}_k = \hat{b}_{-k}$ implies that the filter renders stationary time series that is integrated of oder 2 (I(2)) or less. Thus, the symmetric BK filter has trend-reduction property and, therefore, it can be applied to nonstationary, up to I(2) series.

Since the symmetric BK filter can not be used to extract the desired frequencies at the very end of the input series (for the first and the last K observations), a natural extension of the Baxter and King (1999) filter is to allow the approximate filter to be asymmetric, to be able to use the filter in real time. In order to optimally approximate an ideal symmetric linear filter in a Baxter-King sense, the problem is to minimize

$$\begin{aligned} Q &= \int_{-\pi}^{\pi} \delta(\omega)\delta(-\omega)d\omega \\ &\text{s.t.} \\ \hat{\beta}(0) &= \sum_{h=-p}^{f} \hat{b}_h = 0. \end{aligned} \tag{7}$$

The condition $\hat{\beta}(0)$ ensures zero weight on zero frequency, thus this asymmetric filter also has a trend-reduction property, however, it alone, without symmetry, is not sufficient to render I(2) process stationary. Thus, the ability of the asymmetric BK filter of real time signal extraction comes at a cost of losing the power to eliminate two unit roots from the input series.

To solve (7), form the Lagrangian

$$\mathcal{L} = Q - \lambda\hat{\beta}(0) \tag{8}$$

with first order conditions (FOCs):

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{b}_h} &= \frac{\partial Q}{\partial \hat{b}_h} - \lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= -\hat{\beta}(0) = 0. \end{aligned} \tag{9}$$

Since

$$\frac{\partial}{\partial \hat{b}_h}[\delta(\omega)\delta(-\omega)] = \frac{\partial \delta(\omega)}{\partial \hat{b}_h}\delta(-\omega) + \delta(\omega)\frac{\partial \delta(-\omega)}{\partial \hat{b}_h}, \tag{10}$$

and since the frequency response function of the approximating filter is $\hat{\beta}(\omega) = \sum_{h=-p}^{f} \hat{b}_h e^{-i\omega h}$, it follows that

$$\frac{\partial \delta(\omega)}{\partial \hat{b}_h} = -e^{-i\omega h}. \tag{11}$$

(11) implies

$$\frac{\partial Q}{\partial \hat{b}_h} = -\int_{-\pi}^{\pi} [e^{-i\omega h}\delta(-\omega) + \delta(\omega)e^{i\omega h}]d\omega. \tag{12}$$

By the property $\int_{-\pi}^{\pi}[f(\omega) + f(-\omega)]d\omega = 2\int_{-\pi}^{\pi} f(\omega)d\omega$ (since $\int_{-\pi}^{\pi} f(\omega)d\omega = \int_{0}^{\pi} f(\omega)d\omega + \int_{-\pi}^{0} f(\omega)d\omega = \int_{0}^{\pi}[f(\omega) + f(-\omega)]d\omega$ is real, then $\int_{-\pi}^{\pi} f(\omega)d\omega = \int_{-\pi}^{\pi} f(-\omega)d\omega$, and the property follows), (12) becomes

$$\frac{\partial Q}{\partial \hat{b}_h} = -2\int_{-\pi}^{\pi} \delta(\omega)e^{i\omega h}d\omega. \tag{13}$$

By the property

$$\int_{-\pi}^{\pi} e^{i\omega n}e^{-i\omega m}d\omega = \int_{-\pi}^{\pi} e^{-i\omega(m-n)}d\omega = 0 \quad \text{for } n \neq m$$
$$= 2\pi \quad \text{for } n = m, \tag{14}$$

obtain

$$\int_{-\pi}^{\pi} \delta(\omega)e^{i\omega h}d\omega = \int_{-\pi}^{\pi} \left[\sum_{k=-\infty}^{\infty} b_k e^{-i\omega k} - \sum_{j=-p}^{f} \hat{b}_j e^{-i\omega j}\right] e^{i\omega h}d\omega = 2\pi[b_h - \hat{b}_h]. \tag{15}$$

Given (15), the FOCs are

$$-4\pi[b_h - \hat{b}_h] - \lambda = 0. \tag{16}$$

If there is no constraint on $\hat{\beta}(0)$, the optimal approximate (in Baxter-King sense) filter is simply derived by truncation of the ideal filter's weights. If there is a constraint on $\hat{\beta}(0)$, then λ must be chosen so that the constraint is satisfied. For this purpose, rewrite (16) as

$$\hat{b}_h = b_h + \theta,$$

where $\theta = \lambda/(4\pi)$. In order to have $\hat{\beta}(0) = \sum_{h=-p}^{f} \hat{b}_h = 0$, the required adjustment is

$$\theta = \frac{-\sum_{h=-p}^{f} b_h}{p + f + 1}, \tag{17}$$

which yields the same optimal weight adjustment scheme as in the symmetric Baxter-King filter case.

The next section describes the results from Monte Carlo simulation to assess the performance of the proposed filter.

3 Comparing Filters By Means Of Monte Carlo Simulation

This section assesses the performance of the proposed filter to extract business cycle frequencies (corresponding to wave length between 1.5 and 8 years) in comparison to i) the original symmetric fixed-length BK filter with K=12 (see (6)), as well as ii) fixed-length symmetric CF filter with K=12 for RW processes, and iii) default asymmetric specification of CF filter for RW processes (Christiano and Fitzgerald, 2003). Thus, the asymmetric CF filter assumes that the first difference of the input signal, i.e, $x_t - x_{t-1}$, is zero-mean covariance stationary process. The symmetric CF filter allows for the input signal to follow RW with drift.

Consider the following data generating process (DGP):

$$y_t = \mu_t + c_t, \tag{18}$$

where

$$\mu_t = \mu_{t-1} + \epsilon_t \tag{19}$$

$$c_t = \phi_1 c_{t-1} + \phi_2 c_{t-2} + \eta_t \tag{20}$$

$$\epsilon_t \sim nid(0, \sigma_\epsilon^2),\ \eta_t \sim nid(0, \sigma_\eta^2). \tag{21}$$

Equation (18) defines the input signal, y_t, as the sum of a permanent component (stochastic trend), μ_t, and a cyclical component, c_t. The trend, μ_t, in this case is specified as a random walk process. The dynamics of the cyclical component, c_t, is specified as a second order autoregressive (AR(2)) process so that the peak of the spectrum of c_t could be at zero frequency or at business cycle frequencies. Disturbances, ϵ_t and η_t, are assumed to be uncorrelated.

The spectrum of an AR(2) process is

$$f_c(\omega) = \frac{\sigma_\eta^2}{1 + \phi_1^2 + \phi_2^2 - 2\phi_1(1-\phi_2)\cos\omega - 2\phi_2\cos(2\omega)} \tag{22}$$

with a peak at frequency other than zero for

$$\phi_2 < 0 \,\text{and}\, \left|\frac{\phi_1(1-\phi_2)}{4\phi_2}\right| < 1, \tag{23}$$

with the corresponding frequency $\omega = \cos^{-1}[-\phi_1(1-\phi_2)/(4\phi_2)]$ (Box, Jenkins and Reinsel, 1994; Priestley, 1981).

Data are generated from (18) with $\phi_1 = 1.2$ and different values for ϕ_2 to control the location of the peak in the spectrum of the cyclical component. I also vary the ratio of standard deviations of the disturbances, $\sigma_\epsilon/\sigma_\eta$, to change the relative importance of components of y_t. Such DGP can create series with spectral characteristics typical to macroeconomic variables, such as gross domestic product and inflation (Watson, 1986; Guay and St-Amant, 2005). The idea of such simulation is taken from Ahamada and Jolivaldt (2010) who, in turn, take it from Guay and St-Amant (2005).

Particularly, 10000 samples of length 401 are created, with the first 200 observations of each sample dropped off as burn-in. The vector $[\phi_1, \phi_2]$ is set to five different values, as shown in Table 1. The value of $\sigma_\epsilon/\sigma_\eta$ is set to change from 0 to 9.9 with step size 0.15 (Watson (1986)

ϕ_1	ϕ_2	Fundamental period of the cycle (yrs)
1.2	-0.25	$\approx \infty$
1.2	-0.35	$\gg 8$
1.2	-0.44	8.2
1.2	-0.5	3.5
1.2	-0.8	1.9

Table 1: Five different values of $[\phi_1, \phi_2]$ for the DGP.

estimated this ratio for the U.S. GNP to be 0.75).

The performance of filters is assessed by comparing the correlation of the true cyclical component with the estimated cyclical component, and by comparing the true AR(2) regression coefficients for the cycle with the fitted AR(2) regression coefficients.

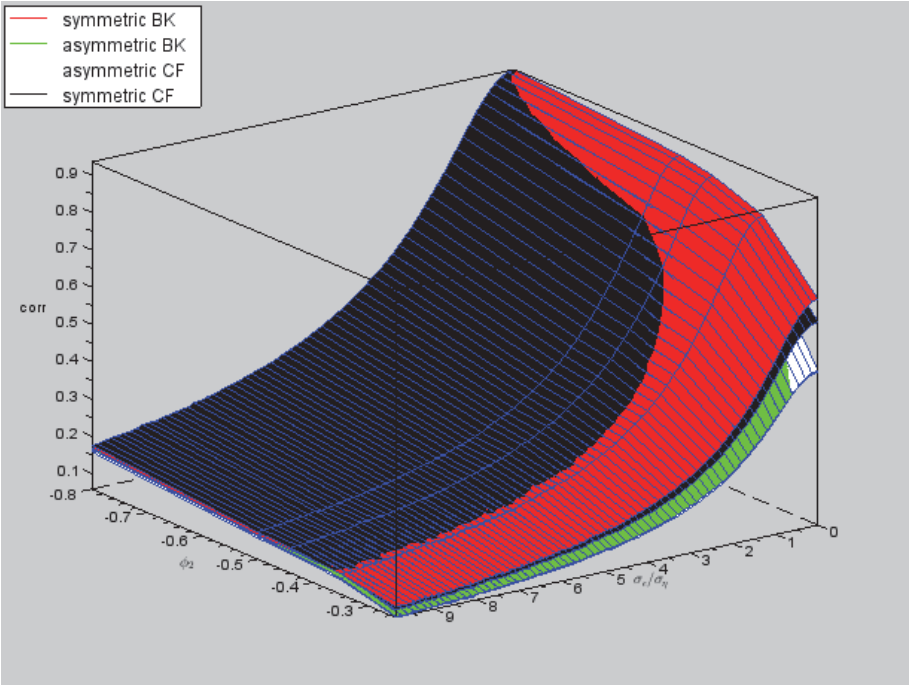


Figure 1: Average correlation between the true and estimated cyclical components for given $[\phi_1, \phi_2]$ and $\sigma_\epsilon/\sigma_\eta$ values. The correlation is estimated for the whole sample interval except for the first and the last K=12 observations.

Figure 1 shows an average correlation between the true and estimated cyclical components for given $[\phi_1, \phi_2]$ and $\sigma_\epsilon/\sigma_\eta$ values. The correlation is estimated for the whole sample interval except for the first and the last K=12 observations, since fixed-length symmetric filters do not produce the estimated cycle for those observations. These K=12 observations are deleted from the output of the asymmetric filters for a fair comparison between symmetric and asymmetric filters. Figure 1 shows a similar behavior between the filters - their performance decreases with $\sigma_\epsilon/\sigma_\eta$, which is an expected result. When $\sigma_\epsilon/\sigma_\eta = 0$, the input signal is the true cycle, so the output signal (estimated cycle) correlates highly with the input. As $\sigma_\epsilon/\sigma_\eta$ increases, the

influence of the permanent component in the input increases, thus making harder for filters to extract the cycle, thus the estimated correlation between the true and estimated cycles, $\hat{\rho}(c, \hat{c})$, decreases.

Figure 1 also shows that the performance of all filters decreases with an increasing ϕ_2. The value of $\phi_2 = -0.8$ together with $\phi_1 = 1.2$ corresponds to the length of the cycle 1.9 years, which is close to the usually defined minimum length of a business cycle, 1.5 years. The value of $\phi_2 = -0.44$ together with $\phi_1 = 1.2$ produces the cycle of length approximately 8.2 years, which is close to the usually defined maximum length of a business cycle, 8 years. With higher then $\phi_2 = -0.44$ values, the length of the true cycle rapidly increases. Although with $\phi_2 = -0.25$ the cycle still is considered stationary ($\phi_1 + \phi_2 < 1$), it is a close approximation to a nonstationary process in a finite sample (Campbell and Perron, 1991). Thus, Figure 1 shows expected deterioration in performance of BK filters as ϕ_2 increases. The similar deterioration in performance of the CF filters with an increasing length of the cycle was less expected. Another unexpected result is the inferior performance of asymmetric filters to their shorter symmetric counterparts. The results also show that the symmetric BK filter is superior to the symmetric CF filter for $0 \leq \sigma_\epsilon/\sigma_\eta < 0.5$ if cycle length is longer than 2 years. For most of the rest of the region, particularly - cycle length less than 8 years, given $\sigma_\epsilon/\sigma_\eta \geq 1$ - the symmetric CF filter is slightly superior to the symmetric BK filter. For the remainder, i.e., $0.5 \leq \sigma_\epsilon/\sigma_\eta < 1$, the symmetric CF filter shows superiority when cycle is relatively short (up to about 3.5 years), and the symmetric BK filter shows superiority when the cycle is longer than approximately 3.5 years.

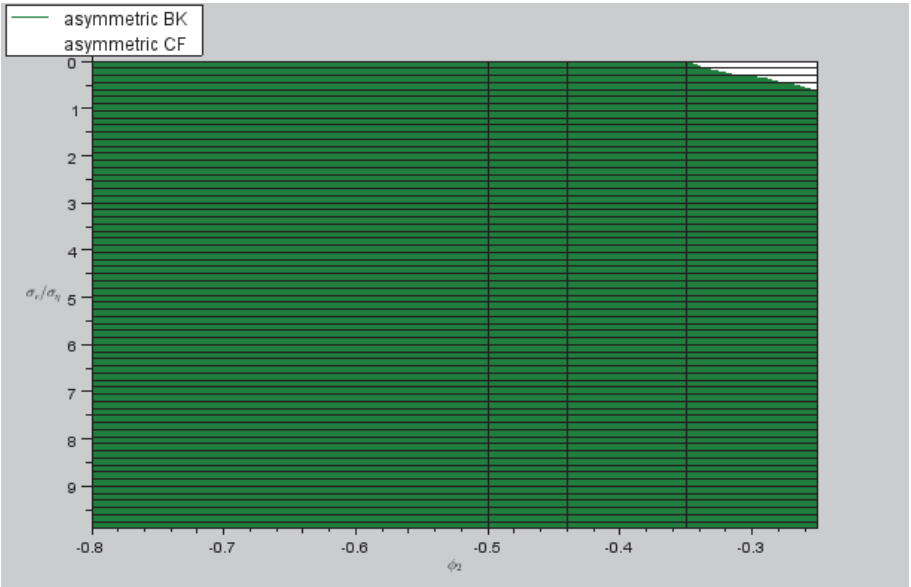


Figure 2: Relative superiority of the asymmetric BK and CF filters in the sample, except for the first and the last K=12 observations. The horizontal axis represent cycle length, while the vertical axis represent the importance of permanent component in the series. The results suggest that, on average in the sample, the asymmetric BK filter is superior to the asymmetric CF filter.

The simulation results show (not supplied to to space limitation) that the distance between

the average correlation between the true and estimated cyclical components for given $[\phi_1, \phi_2]$ and $\sigma_\epsilon/\sigma_\eta$ values from asymmetric BK and CF filters is practically nil at all points in the sample, when the first and the last K=12 observations are dropped. Figure 2 shows the regions of (small) relative superiority of the asymmetric BK and CF filters. Figure 2 shows that the asymmetric BK filter is superior to asymmetric CF filter for all $\sigma_\epsilon/\sigma_\eta$ values and for any interesting length of the cycle. A slightly surprising finding from Figure 2 is the inability of asymmetric CF filter to perform better than the asymmetric BK filter in the region of high influence of the permanent component (corresponds to lower part of the graph).

Now, let us compare the performance of the asymmetric filters for the K=12 observations of the sample, where the fixed-length symmetric filters can not be applied. Figure **??** shows the estimated correlation of the true and estimated cycles at each of the first three observations of the sample, calculated across the 10000 replications, and averaged over both symmetric ends. The correlation graphs at other observations are skipped due to space limitation. The results show that the filters give close result at points closer to the center of the sample. As the estimation point approaches the end of the sample, filters become more asymmetric, and the difference in their performance becomes more obvious. Thus, the asymmetric filters perform roughly equally well at points that are at least about 3 years (for quarterly data) away from the end of the sample. Otherwise, the asymmetric BK filter becomes increasingly superior to the asymmetric CF filter for any cycle length and for any share of permanent component in the input signal considered in the simulation. Thus, based on the results illustrated in Figure **??**, it is recommended to use the asymmetric BK filter rather than the asymmetric CF filter for the business-cycle frequency extraction in real time, i.e, at the very end of the sample.

As for the comparison of the true AR(2) process of the cycle, and the estimated AR(2) regression coefficients, Table 2 shows that the length of the cycle extracted by the filters, when the influence of the permanent component in the input series is sufficiently high, i.e., about $\sigma_\epsilon/\sigma_\eta > 5$, is about constant, regardless of the true length of the cycle. This result shows the potential limitation of the considered filters.

	ϕ_1	ϕ_2	Fundamental period of the cycle (yrs)
true	1.2	-0.25	$\approx \infty$
	1.2	-0.35	$\gg 8$
	1.2	-0.44	8.2
	1.2	-0.5	3.5
	1.2	-0.8	1.9
symmetric BK	1.699	-0.886	3.56
asymmetric BK	1.689	-0.884	3.48
asymmetric CF	1.696	-0.879	3.60
symmetric CF	1.623	-0.848	3.23

Table 2: The true AR(2) parameters and cycle length, and the estimated AR(2) parameters and cycle length by the four filters, when the influence of the permanent component in the input series is sufficiently high, i.e., about $\sigma_\epsilon/\sigma_\eta > 5$. In such case, the estimated AR(2) parameters and the length of the extracted cycle are about constant, regardless of the true length of the cycle.

Acknowledgement

This work has been supported by the European Social Fund within the project "Support for the implementation of doctoral studies at Riga Technical University". The author is thankful to his supervisor Viktors Ajevskis for guidance. Remaining errors are the author's own. The opinions expressed in this paper are those of the author and do not necessarily reflect the views of the Central Statistical Bureau of Latvia.

References

[1] AHAMADA, I., JOLIVALDT, P.: *Classical vs wavelet-based filters Comparative study and application to business cycle*. Documents de travail du Centre d'Economie de la Sorbonne 10027, Universite Pantheon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne, 2010.

[2] BAXTER, M., KING, R. G.: *Measuring business cycles: approximate band-pass filters for economic time series*. In The Review of Economics and Statistics, MIT Press, Vol. 81(4), pp. 575-593, 1999.

[3] BOX, G., JENKINS, G. M., REINSEL, G.: *Time Series Analysis: Forecasting and Control*, 3rd Edition, Prentice Hall, 1994.

[4] CAMPBELL, J. Y., PERRON, P.: *Pitfalls and opportunities: what macroeconomists should know about unit roots*. NBER Technical Working Papers 0100, National Bureau of Economic Research, Inc., 1991.

[5] CHRISTIANO, L. J., FITZGERALD, T. J.: *The band pass filter*. In International Economic Review, Vol. 44(2), pp. 435-465, 2003.

[6] GUAY, A., ST-AMANT, P.: *Do the Hodrick-Prescott and Baxter-King filters provide a good approximation of business cycles?*. In Annales d'Economie et de Statistique, issue 77, 2005.

[7] PRIESTLEY, M. B.: *Spectral Analysis and Time Series*, Academic Press, Inc., 1981.

[8] WATSON, M. W.: *Univariate detrending methods with stochastic trends*. In Journal of Monetary Economy, Elsevier, Vol. 18, pp. 49-75, 1986.

Current address

Ginters Buss, 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

or

Mathematical Support Division, Central Statistical Bureau of Latvia,
Lacplesa iela 1, Riga, LV-1301,
e-mail: ginters.buss@csb.gov.lv

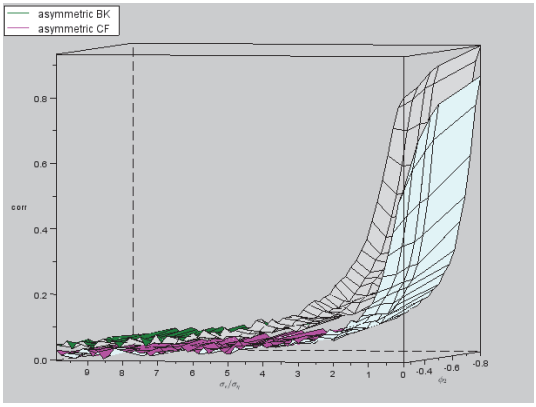


Figure 3: Correlation at obs. 3

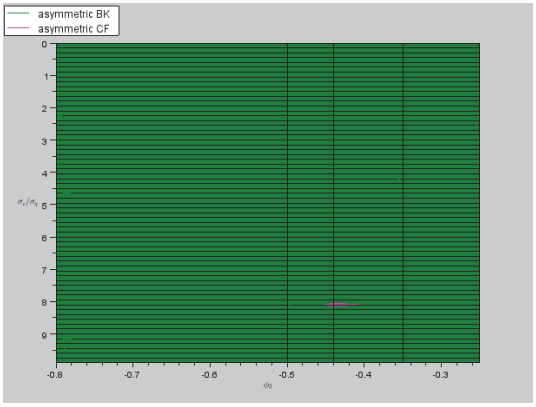


Figure 4: view at 3 from the top

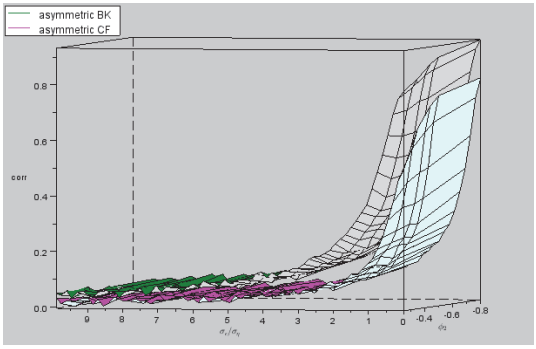


Figure 5: Correlation at obs. 2

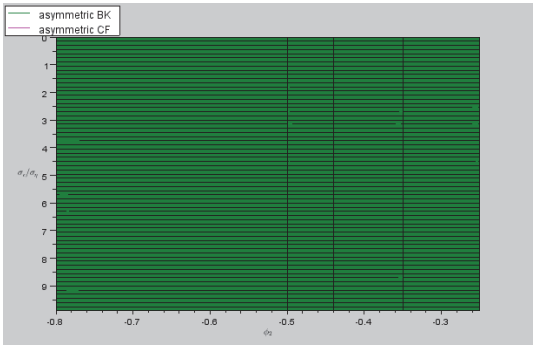


Figure 6: view at 5 from the top

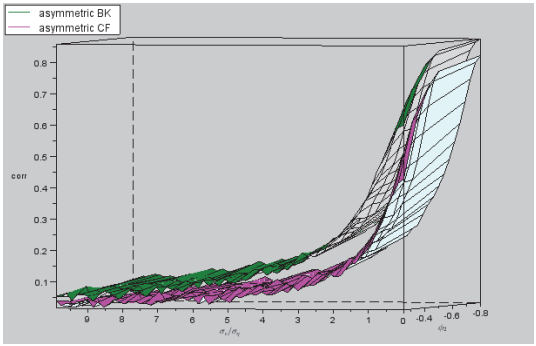


Figure 7: Correlation at obs. 1

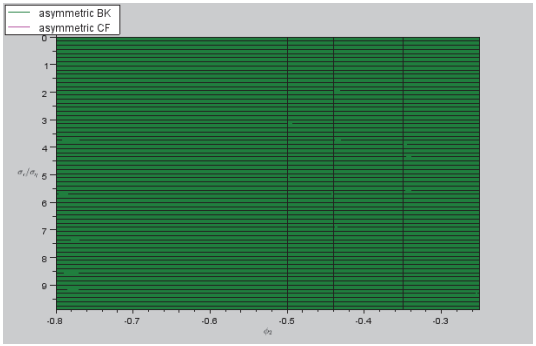


Figure 8: view at 7 from the top

Journal of Applied Mathematics

volume IV (2011), number III

THE M/G/∞ QUEUE BUSY PERIOD DISTRIBUTION EXPONENTIALITY

FERREIRA, Manuel Alberto M., (PT), ANDRADE, Marina, (PT)

Abstract. Infinite servers queuing systems are often used to solve a large number of practical problems, namely in Engineering, Biology, Management, Finance and Sociology. The study of the M/G/∞ systems busy period is very important. Yet it is a very difficult task. This paper presents a service time distributions collection for which the length of the busy period is almost exponential. It is also shown that in heavy-traffic conditions, for a certain class of service time distributions, the length of the busy period of the M/G/∞ systems is approximately exponentially distributed. Thus, in all these cases, the study of this busy period is much simpler. It is also exemplified, with the power function and the Pareto service distributions, that it is possible to have in some particular situations an approximately exponential behaviour for the M/G/∞ queue busy period length.

Key words. M / G / ∞ queue, busy period, exponential distribution, power function distribution, Pareto distribution.

Mathematics Subject Classification: 60K35.

1 Introduction

The M / G / ∞ queue is a system where customers arrive according to a Poisson process at rate λ. Each customer receives a service which length is a positive random variable with mean u and distribution function $G(.)$, and $\alpha = \int_0^\infty [1 - G(t)]dt$. When a customer arrives its service starts at once (i.e., there are infinite servers) and it is independent both from the services of the other customers and from the arrival process. The quantity $\rho = \lambda u$ is the traffic intensity.

Infinite servers queuing systems are often used to solve a large number of practical problems, namely in Engineering, Biology, Management, Finance and Sociology. See, for instance, (Syski, 1960, 1986), (Kleinrock, 1975), (Kelly, 1979), (Hershey, Weiss and Morris, 1981), (Carrillo, 1981), (Figueira and Ferreira, 1999) and (Ferreira, Andrade and Filipe, 2009).

In this paper some more results will be given about the busy period of the M / G / ∞ queue.

For any queuing system, a busy period begins when a customer arrives at the system, finding it empty, ends, when a customer leaves the system, letting it empty, and in it there is always at least one customer present. So, in a queuing system, there is a sequence of idle and busy periods.

The length of an idle period of a M / G / ∞ queue is a random variable, called I, which distribution is exponential with parameter λ, as happens in any queue where the arrival process is Poisson.

For the random variable B, length of the busy period of the M / G / ∞ queue, the situation is not so simple.
Tackács (1962) studied the busy cycle, a busy period followed by an idle period, for the M / G / ∞ queue. Its length is a random variable Z and, evidently, $Z = B + I$. That author showed that B and I are independent and deduced the expression for the Laplace-Stieltjes transform of Z. Using this expression and the fact that B and I are independent, (Stadje, 1985) deduced the expression

$$\overline{B}(s) = 1 + \lambda^{-1}\left(s - \frac{1}{\int_0^{\infty} e^{-st-\lambda\int_0^t [1-G(v)]dv}\, dt} \right) \tag{1}$$

for the Laplace-Stieltjes transform of B and it is possible to show that

$$E[B] = \frac{e^{\rho} - 1}{\lambda} \tag{2}$$

which is independent of the form of $G(.)$. This is not, however, valid for higher order moments.

The inversion of (1) to obtain $b(t)$, the probability density function of B, is very hard to carry out.

For a particular service distributions collections, see (Ferreira, 1998), it will be shown that it is easy to invert (1). And simple expressions for $B(t)$ related to the exponential distribution will be obtained. Then, it is deduced that for a large class of service distributions, since ρ and α are great enough, $b(t)$ is approximately exponential.

Finally, service distributions cases for which does not occur necessarily $\lim_{\alpha\to\infty} G(t) = 0$, as it happens with power function and Pareto ones, are studied in order to identify situations for which B has a similar behaviour.

2 The M/G/ ∞ Systems with Almost Exponential Busy Periods

Lemma 1

If

$$G(t)=1-\frac{\left(1-e^{-\rho}\right)(\lambda+\beta)}{\lambda e^{-\rho}\left(e^{(\lambda+\beta)t}-1\right)+\lambda},\qquad t\geq 0,\ -\lambda\leq\beta\leq\frac{\lambda}{e^{\rho}-1} \tag{3}$$

B is exponentially distributed at rate $e^{-\rho}(\lambda+\beta)$, with an atom at the origin which value is $\frac{e^{-\rho}(\lambda+\beta)-\beta}{\lambda}$.

Demonstration

Substituting (3) in (1) and inverting the consequent $\overline{B}(s)$, it is obtained the following probability density function:

$$b(t)=\frac{e^{-\rho}(\lambda+\beta)-\beta}{\lambda}\delta(t)+\left(1-\frac{e^{-\rho}(\lambda+\beta)-\beta}{\lambda}\right)e^{-\rho}(\lambda+\beta)e^{-e^{-\rho}(\lambda+\beta)t},\quad t\geq 0\ -\lambda\leq\beta\leq\frac{\lambda}{e^{\rho}-1}$$

($\delta(.)$ is the Dirac function).

Notes

- For $\beta=0$, $\overline{B}(s)=e^{-\rho}+\left(1-e^{-\rho}\right)\frac{\lambda e^{-\rho}}{\lambda e^{-\rho}+s}$ and then

$$b(t)=e^{-\rho}\delta(t)+\left(1-e^{-\rho}\right)\lambda e^{-\rho}e^{-\lambda e^{-\rho}t}.$$

- Note that for this distribution the Laplace-Stieltjes transform of Z is

$$\overline{Z}(s)=\frac{\lambda}{\lambda+s}\left(e^{-\rho}+\left(1-e^{-\rho}\right)\frac{\lambda e^{-\rho}}{\lambda e^{-\rho}+s}\right)=\frac{\lambda e^{-\rho}}{\lambda e^{-\rho}+s}.$$

That is: Z is exponentially distributed at rate $\lambda e^{-\rho}$ and so the points in time at which begin busy periods occur according to a Poisson process at rate $\lambda e^{-\rho}$.

- For $\beta=\frac{\lambda}{e^{\rho}-1}$, $\overline{B}(s)=\frac{\lambda\left(e^{-\rho}-1\right)^{-1}}{s+\lambda\left(e^{-\rho}-1\right)^{-1}}$ and B is exponential with mean $\frac{e^{\rho}-1}{\lambda}$. Now the situation for Z is not as simple as for $\beta=0$. It is easy to show, following the same procedures as in the precedent case that the probability density function of Z is

$$z(t)=\frac{\lambda}{e^{\rho}-2}\left(e^{-\frac{\lambda}{e^{\rho}-1}t}-e^{-\lambda t}\right),\quad t\geq 0\,.$$

3 Distributions of *B* when the Service Distribution is such that $\lim_{\alpha\to\infty} G(t)=0$

An example of a distribution for which $\lim_{\alpha\to\infty} G(t)=0$ is the exponential one.

From (1), see for instance (Stadje, 1985) for the distribution function of B the following expression is obtained

$$B(t)=1-\lambda^{-1}\sum_{n=1}^{\infty}\left[\frac{e^{-\lambda\int_0^t[1-G(v)]dv}\lambda(1-G(t)dt)}{1-e^{-\rho}}\right]^{*n}.(1-e^{-\rho})^n, t\geq 0 \tag{4}$$

where* is the operator convolution.

Fixing λ, if $\lim_{\alpha\to\infty} G(t)=0, 1-G(t)\cong 1$ for α great enough. And, if ρ is great enough, then $1-e^{-\rho}\cong 1$. Thus

$$\frac{e^{-\lambda\int_0^t[1-G(v)]dv}\lambda(1-G(t))dt}{1-e^{-\rho}}\cong\lambda e^{-\lambda t}$$

and

$$B(t)\cong 1-\lambda^{-1}\sum_{n=1}^{\infty}\left(\lambda e^{-\lambda t}\right)^{*n}\left(1-e^{-\rho}\right)^n.$$

The Laplace-Stieltjes transform of the second member is

$$\frac{1}{s}\frac{\lambda e^{-\rho}}{s+\lambda e^{-\rho}}+\frac{e^{-\rho}}{s+\lambda e^{-\rho}}$$

and, after its inversion, it is obtained

$$1-\left(1-e^{-\rho}\right)e^{-\lambda e^{-\rho}t}\,.$$

So it results that in such conditions $B(t)=1-e^{-\lambda e^{-\rho}t}$. Then

Lemma 2

For service distributions verifying $\lim_{\alpha\to\infty} G(t)=0$, fixing λ, for α and ρ great enough, B is approximately exponentially distributed.

In order to feel the meaning of α and ρ great enough, the coefficient of variation of B, $CV[B]$ the coefficient of symmetry of B, $\gamma_1[B]$ and the kurtosis of B, $\gamma_2[B]$, were computed for the

$M/G_1/\infty$ (service distribution given by (3) for $\beta = 0$), M/D/∞ (service distribution constant) and M/M/∞ (service distribution exponential) systems for $\rho = .5, 1, 10, 20, 50, 100$ with $\lambda = 1$ (therefore $\rho = \alpha$).

According to (Kendall and Stuart, 1979)

$$CV[B] = \frac{\sqrt{E[B^2] - E^2[B]}}{E[B]} \tag{5},$$

$$\lambda_1[B] = \frac{\left(E[B^3] - 3E[B]E[B^2] + 2E^3[B]\right)^2}{\left(E[B^2] - E^2[B]\right)^3} \tag{6}$$

and

$$\lambda_2[B] = \frac{\left(E[B^4] - 4E[B]E[B^3] + 6E^2[B]E[B^2] - 3E^4[B]\right)^2}{\left(E[B^2] - E^2[B]\right)^2} \tag{7}.$$

For an exponential distribution, $CV = 1$, $\gamma_2 = 4$ and $\gamma_2 = 9$.

From **Lemma 1** with $\beta = 0$, to the $M/G_1/\infty$ queue

$$E[B^n] = \left(1 - e^{-\rho}\right)\frac{n!}{\left(\lambda e^{-\rho}\right)^n}, n = 1,2,... \tag{8}.$$

With $n = 1$ (2) results from (8).

For the remaining systems, note that (1) is equivalent to $\left(\overline{B}(s) - 1\right)\left(C(s) - 1\right) = \lambda^{-1} s C(s)$ being $C(s) = \int_0^\infty e^{-st - \lambda\int_0^t [1 - G(v)]dv} \lambda\left(1 - G(t)\right)dt$. Differentiating n times, using Leibnitz's formula and making $s = 0$, results

$$E[B^n] = (1)^{n+1}\left\{\frac{e^\rho}{\lambda} n C^{(n-1)}(0) - e^\rho \sum_{\rho=1}^{n-1} (-1)^{n-p} \binom{n}{p} . E[B^{n-p}] C^{(p)}(0)\right\}, n = 1,2,... \tag{9}$$

being

$$C^{(n)}(0) = \int_0^\infty (-t)^n e^{-\lambda\int_0^t [1 - G(v)]dv} \lambda\left(1 - G(t)\right)dt, n = 0,1,2,... \tag{10}.$$

The expression (9) gives a recursive method to compute $E[B^n], n = 1,2,...$ as a function of $C^{(n)}(0), n = 0,1,2,...$

Making $n = 1$ (2) results from (9)

- For the M/D/∞ system

$$C^{(0)}(0) = 1 - e^{-\rho}$$

$$C^{(n)}(0) = -e^{-\rho}(-\alpha)^n - \frac{n}{\lambda} C^{n-1}(0), n = 1,2,... \tag{11}$$

and it is possible to compute any $E\left[B^n\right]$ exactly.

- For the M/M/∞ system it is not possible to obtain expressions as simple as (11) to the $C^{(n)}(0)$. It is mandatory to compute numerically integrals with infinite limits and so approximations must be done.

The results are:

System / **ρ**	$M\ /\ G_1\ /\ \infty$	$M\ /\ D\ /\ \infty$	$M\ /\ M\ /\ \infty$
.5	2.0206405	.40655883	1.1109224
1	1.4710382	.56798436	1.1944614
10	1.0000454	.99959129	1.1227334
20	1.0000000	.99999999	1.0544722
50	1.0000000	.99999999	1.0206393
100	1.0000000	.99999999	1.0101547

Table 1: ***CV*[*B*]**

System / **ρ**	$M\ /\ G_1\ /\ \infty$	$M\ /\ D\ /\ \infty$	$M\ /\ M\ /\ \infty$
.5	9.5577742	6.0360869	5.0972761
1	5.5867425	4.5899937	5.4821324
10	4.0000000	4.0000000	4.1511831
20	4.0000000	4.0000000	4.0326858
50	4.0000000	4.0000000	4.0049427
100	4.0000000	4.0000000	4.0012250

Table 2: γ_1**[*B*]**

System / **ρ**	$M\ /\ G_1\ /\ \infty$	$M\ /\ D\ /\ \infty$	$M\ /\ M\ /\ \infty$
.5	15.983720	11.142336	10.454678
1	10.878212	9.6137084	10.923071
10	9.0000000	9.0000000	9.1617573
20	9.0000000	9.0000000	9.0337903
50	9.0000000	9.0000000	9.0550089
100	9.0000000	9.0000000	9.0012250

Table 3: γ_2**[*B*]**

The parameters studied assume values that are typical of an exponential distribution, after $\rho = 10$, for the $M/G_1/\infty$ and $M/D/\infty$ queues. For the $M/M/\infty$ system, only after $\rho = 20$ it can be said that those values are the ones of an exponential distribution.

Finally note that the convolution of the exponentials with parameters λ and $\lambda e^{-\rho}$ gives an approximate expression for $Z(t)$, distribution function of Z, in the same condition of Lemma 2:

$$z(t) \cong \frac{1-e^{\rho}-e^{-\lambda t}+e^{\rho-\lambda e^{-\rho}t}}{1-e^{\rho}}, \qquad t \geq 0 \tag{12}.$$

4 Power Function Service Distribution

If the service distribution is a power function with parameter $C, C > 0$

$$G(t) = \begin{cases} t^c & 0 \leq t < 1 \\ 1, & t \geq 1 \end{cases},$$

$$\alpha = \frac{C}{C+1} \text{ and } 0 < \alpha < 1.$$

So $\lim_{C\to\infty} G(t) = \begin{cases} 0, & 0 \leq t < 1 \\ 1, & t \geq 1 \end{cases}$ and $\lim_{C\to\infty} \alpha = 1$

For this service time distribution, since ρ and C are great enough,

$$\frac{e^{-\lambda\int_0^t [1-G(v)]dv}\,\lambda(1-G(t))}{1-e^{-\rho}} \cong \lambda e^{-\lambda t}, 0 \leq t \leq 1.$$

Computing the Laplace-Stieltjes transform of (4) with this approximation it is obtained first

$$\int_0^1 e^{-st}\lambda e^{-\lambda t}dt = \lambda\int_0^1 e^{-(s+\lambda)t}dt = \lambda\left[-\frac{e^{-(s+\lambda)t}}{s+\lambda}\right]_0^1 = \lambda\frac{1-e^{-(s+\lambda)}}{s+\lambda}.$$

But, for λ great enough, this is approximately $\frac{\lambda}{\lambda+s}$. So, from (4), assuming the Laplace-Stieltjes transform of

$$\frac{e^{-\lambda\int_0^t [1-G(v)]dv}\,\lambda(1-G(t))}{1-e^{-\rho}}, 0 \leq t \leq 1 \text{ as being } \frac{\lambda}{\lambda+s}$$

the conclusion is that $B(t) \cong 1-e^{-\lambda e^{-\rho}t}$. That is:

Lemma 3

In an M/G/∞ queue where the service distribution is a power function, for α near 1, since α and ρ are great enough B is approximately exponential with mean $\frac{e^{\rho}}{\lambda}$.

For this system the values of $\gamma_1(B)$ and $\gamma_2(B)$ were computed for $\alpha = .25$, .5 and .8 making, in each case, ρ take values from .5 till 100. The results are:

ρ	$\alpha = .25$		$\alpha = .5$		$\alpha = .8$	
	$\gamma_{1(B)}$	$\gamma_{2(B)}$	$\gamma_{1(B)}$	$\gamma_{2(B)}$	$\gamma_{1(B)}$	$\gamma_{2(B)}$
.5	3.0181197	9.5577742	1.5035507	5.9040102	3.8933428	9.3287992
1	4.4211164	9.1402097	2.7111584	7.4994861	3.9854257	9.0702715
1.5	5.3090021	10.433228	3.3711526	8.2784408	3.9749455	8.9969919
2	5.8206150	11.140255	3.7332541	8.6924656	3.9751952	8.9815770
2.5	6.0803833	11.489308	3.9322871	8.9173048	3.9809445	8.9828631
3	6.1786958	11.619970	4.0388433	9.0369125	3.9871351	8.9877124
6	5.7006232	11.020248	4.0969263	9.1024430	3.9996462	3.9996459
7	5.5034253	10.774653	4.0765395	9.0804332	3.9999342	8.9999341
8	5.3382992	10.570298	4.0596336	9.0623268	3.9999992	8.9999992
9	5.2037070	10.404722	4.0467687	9.0486468	4.0000086	9.0000086
10	5.0944599	10.271061	4.0372385	9.0385796	4.0000068	9.0000068
15	4.7702550	9.8790537	4.0152698	9.0156261	4.0000005	9.0000005
20	4.6102777	9.6888601	4.0082556	9.0083980	4.0000000	9.0000000
50	4.3045903	9.3338081	4.0012425	9.0012513	4.0000000	9.0000000
100	4.1715617	9.1842790	4.0003047	9.0003057	4.0000000	9.0000000

Table 4: **$\gamma_1(B)$ and $\gamma_2(B)$ considering power function service distribution**

The analysis of the results shows a strong trend of $\gamma_1(B)$ and $\gamma_2(B)$, to 4 and 9, respectively, after $\rho = 10$. This trend is faster the greatest is the value of α .

5 Pareto Service Distribution

Through this section only some examples will be presented. So consider a Pareto distributions such that

$$1 - G(t) = \begin{cases} 1, & t < k \\ \left(\frac{k}{t}\right)^3, & t \geq k \end{cases}, k > 0 \tag{13}$$

Then, $\alpha = \frac{3}{2}k$ (see, for instance, (Murteira, 1979)).

The values calculated for $\gamma_1(B)$ and $\gamma_2(B)$ with $\lambda = 1$ and, so, $\rho = \alpha$ were

$\alpha = \rho$	$\gamma_1(B)$	$\gamma_2(B)$
.5	1028.5443	1373.4466
1	1474.7159	1969.0197
10	38.879220	54.896896
20	4.0048588	9.0049233
50	4.0000000	9.0000000
100	4.0000000	9.0000000

Table 5: **$\gamma_1(B)$ and $\gamma_2(B)$ considering the Pareto service distribution given by (13)**

and show a strong trend from $\gamma_1(B)$ and $\gamma_2(B)$ to 4 and 9, respectively, after $\rho = 20$. This is natural because, in this case, the convergence of α to infinite imposes the same behaviour to k. And so, after (13), it results $\lim_{\alpha \to \infty} G(t) = 0$.

But, considering now a Pareto distribution, such that

$$1 - G(t) = \begin{cases} 1, & t < .4 \\ \left(\frac{.4}{t}\right)^{\theta} & t \geq .4 \end{cases}, \qquad \theta < 1 \qquad (14),$$

$\alpha = \frac{.4\theta}{\theta - 1}$ (See, again, Murteira, 1979) and the values got for $\gamma_1(B)$ and $\gamma_2(B)$ in the same conditions of the previous case are

$\alpha = \rho$	$\gamma_1(B)$	$\gamma_2(B)$
.5	10.993704	16.675733
1	6.8553306	12.010791
10	4.5112470	9.5724605
20	4.4832270	9.5397410
50	4.4669879	9.5208253
100	4.4616718	9.5146406

Table 6: **$\gamma_1(B)$ and $\gamma_2(B)$ considering the Pareto service distribution given by (14)**

and do not go against the hypothesis of the existence of a trend from $\gamma_1(B)$ and $\gamma_2(B)$ to 4 and 9, respectively, although much slower than in the previous case. But, now, the convergence of α to infinite implies the convergence of θ to 1. So

$$\lim_{\alpha \to \infty} G(t) = \begin{cases} 0, & t < .4 \\ 1 - \frac{.4}{t} & t \geq .4 \end{cases}$$

and it is not possible to guarantee at all that for α great enough $1 - G(t) \cong 1$.

6 Conclusions

The exponential distribution is very simple and quite useful from a practical point of view. It has been frequently used to study queuing systems. Among its various properties, it is remarkable its lack of memory, i.e.: $P[T > t + y|T > y] = P[T > t]$, where T is an exponential random variable.

The determination of $b(t)$ is very fastidious for any kind of queuing system and not only for the M/G/ ∞ queue.

Conditions under which B is exponential or approximately exponential for the M/G/ ∞ queue were derived.

Many quantities of interest in queues are insensible. This means that they depend on the service distribution only by its mean. Thus it is indifferent which service distribution is being considered. But using those given by (3), result quasi-exponential or exponential busy periods. And, for these service distributions, all distributions related to the busy cycle have simple forms and are related to the exponential distribution.

In section 3, for a large class of distributions under conditions of heavy- traffic, it was proved that B is approximately exponential irrespectively of the service distribution.

But, for instance, if the service distribution is a power function, as it was seen, such conditions can not hold, at least in the same way. However, for α near 1 and λ and ρ great enough, it is possible to guarantee that B is approximately exponentially distributed.

And in the case of the Pareto distribution, where $G(t) \cong 0$ for α great enough does not necessary hold. Although it is not possible to give identical guarantees to those of the power function service, the results got for $\gamma_1(B)$ and $\gamma_2(B)$ are not against that, for ρ great enough, B is approximately exponentially distributed.

References

[1] CARRILLO, M. J.: *Extensions of Palm's Theorem: A Review*. In Management Science, Vol. 37, No 6, pp. 739-744, 1991.

[2] FERREIRA, M. A. M.: *Application of Ricatti Equation to the Busy Period Study of the $M/G/\infty$ System*. In Statistical Review, Vol I. INE, pp. 23-28, 1998.

[3] FERREIRA, M. A. M.: *$M/G/\infty$ Queue Heavy-Traffic Situation Busy Period Length Distribution (Power and Pareto Service Distributions)*. In Statistical Review, Vol. 1. INE, pp. 27-36, 2001.

[4] FERREIRA, M. A. M.: *The Exponentiality of the $M/G/\infty$ Queue Busy Period*. In Actas das XII Jornadas Luso-Espanholas de Gestão Científica, Volume VIII- Economia da Empresa e Matemática Aplicada. UBI, Covilhã, Portugal. pp. 267-272, 2002.

[5] FERREIRA, M. A. M.: *Statistical Queueing Theory*. Entry in Lovric, M. (ed.), International Encyclopaedia of Statistical Science. Springer. Forthcoming.

[6] FERREIRA, M. A. M. and Andrade, M.: *The Ties Between the M/G/∞ Queue System Transient Behaviour and the Busy Period*. In International Journal of Academic Research 1 (1), pp. 84-92, 2009.

[7] FERREIRA, M. A. M. and ANDRADE, M.: Looking to a *M/G/∞* system occupation through a Ricatti equation. In Journal of Mathematics and Technology, 1(2), pp. 58-62, 2010.

[8] FERREIRA, M. A. M. and ANDRADE, M.: *M/G/∞ System Transient Behavior with Time Origin at the Beginning of a Busy Period Mean and Variance*. In Aplimat- Journal of Applied Mathematics, 3(3), pp. 213-221, 2010.

[9] FERREIRA, M. A. M. and ANDRADE, M.: M/G/∞ Queue Busy Period Tail. In Journal of Mathematics and Technology, 1(3), 11-16, 2010.

[10] FERREIRA, M. A. M. e RAMALHOTO M. F.: *Estudo dos parâmetros básicos do Período de Ocupação da Fila de Espera $M/G/\infty$*. In A Estatística e o Futuro da Estatística. Actas do I Congresso Anual da S.P.E.. Edições Salamandra, Lisboa, 1994.

[11] FERREIRA, M. A. M., ANDRADE, M. and FILIPE, J. A.: The Ricatti Equation in the $M/G/\infty$ Busy Cycle Study. In Journal of Mathematics, Statistics and Allied Fields 2(1), 2008.

[12] FERREIRA, M. A. M., ANDRADE, M. and FILIPE, J. A.: *Networks of Queues with Infinite Servers in Each Node Applied to the Management of a Two Echelons Repair System*. In China-USA Business Review 8(8), pp. 39-45 and 62, 2009.

[13] FIGUEIRA, J. and FERREIRA, M. A. M.: *Representation of a Pensions Fund by a Stochastic Network with Two Nodes: An Exercise*. In Portuguese Revue of Financial Markets, Vol. I, No 3, 1999.

[14] HERSHEY, J. C., WEISS, E. N. and MORRIS, A. C.: *A Stochastic Service Network Model with Application to Hospital Facilities*. In Operations Research, Vol. 29, No 1, pp. 1-22, 1981.

[15] KELLY, F. P.: *Reversibility and Stochastic Networks*. New York: John Wiley and Sons, 1979.

[16] KENDALL and STUART: *The Advanced Theory of Statistics. Distributions Theory*. London. Charles Griffin and Co., Ltd. 4th Edition, 1979.

[17] KLEINROCK, L.: *Queueing Systems*. Vol. I and Vol. II. Wiley- New York, 1985.

[18] MURTEIRA, B.: *Probabilidades e Estatística*,. Vol. I. Editora McGraw-Hill de Portugal, Lda. Lisboa, 1979.

[19] STADJE, W.: *The Busy Period of the Queueing System $M/G/\infty$*. In Journal of Applied Probability, 22, pp. 697-704, 1985.

[20] SYSKI, R.: *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd- London, 1960.

[21] SYSKI, R.: *Introduction to Congestion Theory in Telephone Systems*. North Holland. Amsterdam, 1986.

[22] TACKÁCS, L.: *An Introduction to Queueing Theory*. Oxford University Press. New York, 1962.

Current address

Manuel Alberto M. Ferreira, Professor Catedrático
ISCTE – Lisbon University Institute
UNIDE - IUL
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

ISCTE – Lisbon University Institute
UNIDE - IUL
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

Journal of Applied Mathematics

volume IV (2011), number III

EXPLORING THE REGIONAL CZECH HOUSEHOLD INCOME DYNAMICS VIA REGRESSION MIXTURES

FORBELSKÁ Marie, (CZ)

Abstract. Model–based cluster analysis is popular methods for grouping observations into unobserved segments. The article deals with cluster analysis of household income dynamics based on the results of statistical survey EU SILC 2005–2008. In this paper we address the problem of clustering sets of trajectory data generated by households in different regions. The focus is to model curve data directly using a set of model–based curve clustering algorithms referred to as mixtures of regressions or regression mixtures. The R environment (R Development Core Team, 2010) is used for mixture model analysis.

Keywords. clustering, mixtures of regressions, generalised linear model, linear mixed models, household income, EU SILC.

Mathematics Subject Classification. Primary 62H30; Secondary 30C40.

1 Introduction

The EU-SILC (European Union Statistics on Income and Living Conditions) is an instrument aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. This instrument is anchored in the European Statistical System (ESS). For the first time this investigation was carried out by the Czech Statistical Office in 2005 under the name Living Conditions 2005.

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.

The analysis has been carried out on household income, adjusted for different household types using an equivalence scale. Personal equivalised income is obtained by dividing the total household disposable income by the equivalised size of the household, using modified OECD scale: 1 for the first person aged 14 or more; 0.5 for any subsequent person aged 14 or more; and 0.3 for persons aged less then 14.

2 Methods

Modelling clustered and longitudinal data with and without nested factors has gained importance in recent years. Early expositions are the books by Searle, Casella, and McCulloch (1992), Verbeke and Molenberghs (2001) and McCulloch and Searle (2001), which deal primarily with linear mixed models (*LMM*s). Hierarchical linear model (*HLM*) or multi-level formulations are discussed in Raudenbush and Bryk (2001), which can be rewritten as *LMM*s. Extensions to generalized *LMM* (*GLMM*) are considered in Molenberghs and Verbeke (2005), Fitzmaurice, Laird and Ware (2004) and an up-to-date mathematical treatment is given by Jiang (2007). A Bayesian perspective of *HLM*s is taken in Gelman and Hill (2006).

2.1 Linear Mixed Models (*LMM*s)

Linear mixed models extend classical linear models by incorporating random effects in the structure. Assume that the data set at hand consists of N subjects (here households). Let n_i denote the number of observations for the ith subject. $\mathbf{Y}_i$ is the $n_i \times 1$ vector of observations for the ith household ($1 \leq i \leq N$). The general linear mixed model is specified as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \qquad i = 1, \ldots, N. \tag{1}$$

$\boldsymbol{\beta}$ $(p \times 1)$ gives the p fixed-effects parameters. These are fixed, but unknown, regression parameters, common to all subjects. $\mathbf{b}_i$ $(q \times 1)$ is the vector with the random effects for the ith subject in the data set. The use of random effects reflects the belief that there is heterogeneity among subjects for a subset of the regression coefficients in $\boldsymbol{\beta}$. $\mathbf{X}_i$ $(n_i \times p)$ and $\mathbf{Z}_i$ $(n_i \times q)$ are the design matrices for the p fixed and q random effects, and $\boldsymbol{\varepsilon}_i$ $(n_i \times 1)$ contains the residual components for subject i. Independence between subjects is assumed. Here $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ also are assumed to be independent, and we follow the traditional assumption that they are normally distributed with mean vector $\mathbf{0}$ and covariance matrices, say, $\mathbf{D}$ $(q \times q)$ and $\boldsymbol{\Sigma}_i$ $(n_i \times n_i)$, respectively. Different structures for these covariance matrices are possible; an overview of some frequently used ones can be found in Verbeke and Molenberghs (2001). It is easy to see that $\mathbf{Y}_i$ then has a marginal normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}_i = Var(\mathbf{Y}_i)$, given by

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i. \tag{2}$$

In this interpretation it becomes clear that the fixed effects enter only the mean $E\mathbf{Y}_i$, whereas the inclusion of subject–specific effects specifies the structure of the covariance between obser-

vations on the same unit. Under the traditional normality assumptions,

$$\begin{aligned} \mathbf{Y}_i|\mathbf{b}_i &\sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_i), \\ \mathbf{b}_i &\sim N(\mathbf{O}, \mathbf{D}), \end{aligned} \tag{3}$$

it becomes clear that the residual terms model variability within a subject.

Denote the unknown parameters in the covariance matrix $\mathbf{V}_i$ with $\boldsymbol{\psi}$. Conditional on $\boldsymbol{\psi}$, a closed–form expression for the maximum likelihood estimator of $\boldsymbol{\beta}$ exists, namely,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i. \tag{4}$$

Conditional on $\boldsymbol{\psi}$, this is the *Best Linear Unbiased Estimator* (*BLUE*) for $\boldsymbol{\beta}$, where *best* is in the sense of minimum mean squared error. To predict the random effects, the mean of the posterior distribution of the random effects given the data, $\mathbf{b}_i|\mathbf{Y}_i$, is used. Conditional on $\boldsymbol{\psi}$, we have

$$\widehat{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{5}$$

which can be proven to be the *Best Linear Unbiased Predictor* (*BLUP*) of $\mathbf{b}_i$ (where *best* is again in the sense of minimum mean squared error). Estimation of $\boldsymbol{\psi}$ is mostly performed by maximum likelihood (*ML*) or restricted maximum likelihood (*REML*). The expression maximized by the *ML* (l_1), respectively *REML* (l_2), estimates is given by

$$l_1(\boldsymbol{\psi}; \mathbf{y}_1, \ldots, \mathbf{y}_N) = c_1 - \tfrac{1}{2}\sum_{i=1}^{N} \log(|\mathbf{V}_i|) - \tfrac{1}{2}\sum_{i=1}^{N} \mathbf{r}_i'\mathbf{V}_i^{-1}\mathbf{r}_i, \tag{6}$$

$$l_2(\boldsymbol{\psi}; \mathbf{y}_1, \ldots, \mathbf{y}_N) = c_2 - \tfrac{1}{2}\sum_{i=1}^{N} \log(|\mathbf{V}_i|) - \tfrac{1}{2}\sum_{i=1}^{N} \mathbf{r}_i'\mathbf{V}_i^{-1}\mathbf{r}_i - \tfrac{1}{2}\sum_{i=1}^{N} \log(|\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i|), \tag{7}$$

where $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i'(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i)^{-1}$ and c_1, c_2 are appropriate constants. Equations (6) and (7) are maximized using iterative numerical techniques such as Fisher scoring or Newton–Raphson (for full details, see Demidenko, 2004). In equations (4) and (5) the unknown $\boldsymbol{\psi}$ is then replaced with $\widehat{\boldsymbol{\psi}}_{ML}$ or $\widehat{\boldsymbol{\psi}}_{REML}$, leading to the *empirical BLUE* for $\boldsymbol{\beta}$ and the *empirical BLUP* for $\mathbf{b}_i$. For inference regarding the fixed and random effects and the variance components, appropriate likelihood ratio and Wald tests are explained in Verbeke and Molenberghs (2000).

The predictor for the conditional expectation $E(\mathbf{Y}_i|\mathbf{b}_i) = \boldsymbol{\mu}_{i|\mathbf{b}_i} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ is obtained from equation (4) and (5), namely,

$$\begin{aligned} \widehat{\boldsymbol{\mu}_{i|\mathbf{b}_i}} &= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) \\ &= (\mathbf{I}_{n_i} - \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}) + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i = \boldsymbol{\Sigma}_i\mathbf{V}_i^{-1}\mathbf{X}_i\widehat{\boldsymbol{\beta}} + (\mathbf{I}_{n_i} - \boldsymbol{\Sigma}_i\mathbf{V}_i^{-1}\mathbf{X}_i)\mathbf{Y}_i. \end{aligned}$$

Note that this expression is a weighted average of $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$ (related to the whole population) and $\mathbf{Y}_i$ (related to subject i).

2.2 Generalized Linear Mixed Models (*GLMM*s)

Generalized linear models (*GLM*s) are, as the name suggests, a generalization or extension of normal linear model. *GLM*s incorporate normal linear models as a special case, but also cater for other error distributions (binomial, Poisson, negative binomial, or gamma distribution). Nelder and Wedderburn (1972) were the first to propose the generalized linear model to encompass these different models under one unified mathematical framework.

The generalized linear mixed model (*GLMM*) (see McCullagh and Nelder, 1989) consists of three parts: a link function, a linear predictor, and a distributional model.

Given $\mathbf{b}_i$, the variables $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ are mutually independent with a density function (from the exponential family of distribution) given by

$$f(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}) = \exp\left\{\frac{y_{ij}\theta_{ij} - a(\theta_{ij})}{d_{ij}(\phi)} + c(y_{ij}, \phi)\right\} \tag{8}$$

where θ_{ij} is the canonical parameter and ϕ is the scale parameter. The functions d_{ij} and c are specific to each distribution.

The conditional mean and the conditional variance of Y_{ij} are given by

$$E(Y_{ij}|\mathbf{b}_i) = \mu_{ij|\mathbf{b}_i} = g^{-1}(\eta_{ij}) = g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i) \tag{9}$$

$$Var(Y_{ij}|\mathbf{b}_i) = v(\mu_{ij|\mathbf{b}_i})d_{ij}(\phi) \tag{10}$$

where g and v are, respectively, the link and the variance function, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are the j-th row of the matrix $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively.

The random effects $\mathbf{b}_1, \ldots, \mathbf{b}_N$, are mutually independent with a common underlying distribution G which depends on the unknown parameter $\boldsymbol{\psi}$. Next the vector of random effects $\mathbf{b}_i$ is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{D}$.

2.3 Finite Mixtures of Regression Models

For a mixture of K component distributions of *GLMM*s in proportions $\pi_1, \ldots, \pi_K$ ($\sum_{k=1}^{K} \pi_k = 1$), we have that the conditional density of the response variable $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ ($i = 1, \ldots, N$) given fixed and random covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$ is modeled by

$$f(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\Psi}_i) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\eta}_{i,k}), \tag{11}$$

with the vector $\boldsymbol{\eta}_{i,k} = (\eta_{i1,k}, \ldots, \eta_{in_i,k})'$ of linear predictors $\eta_{ij,k} = g(\mu_{ij,k|\mathbf{b}_{i,k}}) = \mathbf{x}'_i\boldsymbol{\beta}_k + \mathbf{z}'_i\mathbf{b}_{i,k}$ and vectors of unknown parameters $\boldsymbol{\Psi}_i = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\eta}'_{i,1}, \ldots, \boldsymbol{\eta}'_{i,K})'$ ($i = 1, \ldots, N$).

The log likelihood is given by $\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{N} \log\left\{\sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\eta}_{i,k})\right\}$ and the parameter vector $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\eta}'_{1,1}, \ldots, \boldsymbol{\eta}'_{N,K})'$ can be estimated by maximum likelihood

(*ML*) and can be obtained via the *Expectation–Maximization* (*EM*) algorithm of Dempster et al. (1977).

Using an estimate of the vector of all unknown parameters $\mathbf{\Psi}$, this approach gives a probabilistic clustering of $(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i)$ into K clusters in terms of estimates of the posterior probabilities of component membership

$$\omega_k(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i) = \frac{p_k f_k(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\eta}_{i,k})}{f(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{\Psi}_i)}, \tag{12}$$

where $\omega_k(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i)$ is the posterior probability that $(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i)$ belongs to the kth component of the mixture $(i = 1, \ldots, N,\ k = 1, \ldots, K)$. In the Bayesian framework, we use the rule which assigns observation $(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i)$ to the class for which it has the highest posterior probability.

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 also McLachlan and Peel (2000).

3 Model Based Clustering of Household Incomes Over Years 2005–2008

Clustering is typically used as a tool for understanding and exploring large data sets. Finite mixture models are commonly used as a basis for cluster analysis.

EU-SILC will gather comparative statistics on income distribution and social exclusion from the 25 EU members states, Norway and Iceland. EU-SILC data are highly complex and contain detailed information on the income of the sampled individuals and households. More information on EU-SILC can be found in Eurostat (2004).

EU-SILC will provide two types of annual data:

- Cross-sectional data pertaining to a given time or a certain time period with variables on income, poverty, social exclusion and other living conditions, and
- Longitudinal data pertaining to individual-level changes over time, observed periodically over a four years period.

3.1 Poverty Across Czech NUTS3 Regions

The standard poverty rate used in this section (60% of the national median equivalised disposable income) is a relative definition as it depends on the average income of the country. The national average at-risk-of-poverty rate, however, masks important differences within a country, including regional differences (see Table 1 and Figure 1).

The division of NUTS3 regions into sub–groups based on percentage of households below the poverty line, we used the following mixture of binomial logit regressions

$$f(y_{i,year}|n_{i,year}, year, \mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k Bi(y_{i,year}|n_{i,year}, year, \eta_{i,year,k})$$

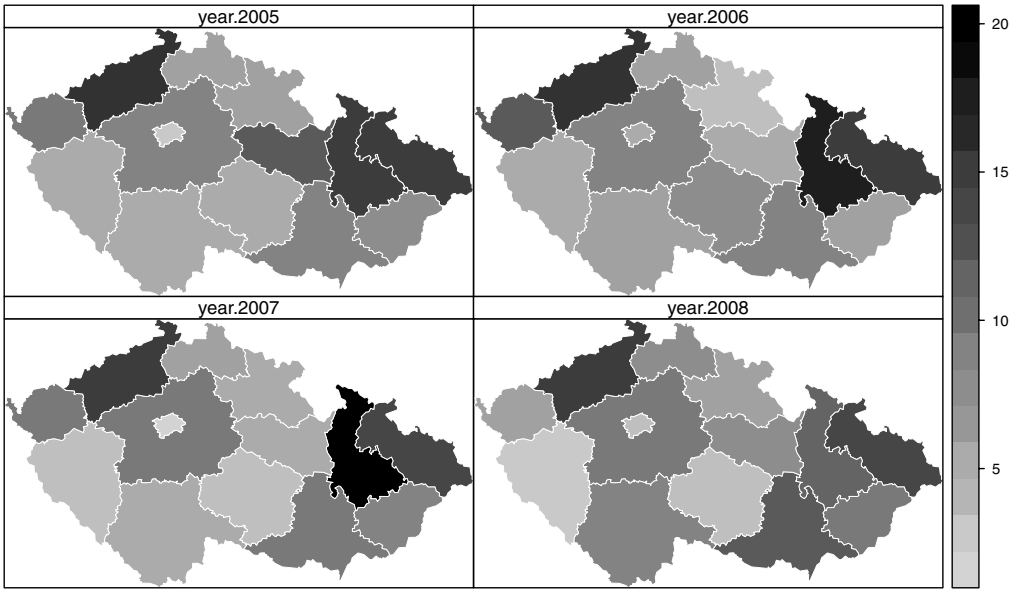


Figure 1: At-risk-of-poverty rates across Czech NUTS3 regions (in %)

Table 1: Regional poverty rates (in %) - Source: own calculations based on EU-SILC 2005–2008

Region	11	21	31	32	41	42	51	52	53	61	62	71	72	81	CZ
2005	3.4	9.2	5.1	5.8	10.1	16	7.1	6.4	12.1	5.2	9	15	7.4	15	9.4
2006	4.9	9.3	6.5	5.1	13.2	16	6.6	3.7	4.9	7.2	9	17	6.4	15	9.3
2007	2.2	9.7	5.8	4.3	9.8	15	6.9	4.8	5.0	3.6	10	19	8.5	14	8.9
2008	4.0	10.6	8.4	2.6	7.1	15	8.2	6.7	8.1	4.3	12	12	10.4	14	9.3

where $y_{i,year}$ is the at-risk-of-poverty rate of i-th region ($i = 1, \ldots, 14$), $n_{i,year}$ is the number of households of i-th region and $year = 2005, \ldots, 2008$ (categorial variable). In order to determine the suitable number of components K, the mixture is fitted with different numbers of components and the *BIC* information criterion is used to select an appropriate model. In this case a model with three components is selected. The fitted values for the model with three components are given in Figure 2, Figure 3 and Table 3. Calculations were performed using the package *flexmix* (Gruen and Leisch, 2007 a 2008).

3.2 Clustering of Regional Household Income Dynamics

Finally, we applied finite mixture model of linear mixed models on each NUTS3 region separately. The data is illustrated in the Figures 4, 5, 6 and the model is given by

$$f(\mathbf{y}_{i,region}|year, region, \mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k N(\mathbf{y}_{i,region}|\boldsymbol{\mu}_{i,region}, year, \boldsymbol{\beta}_{k,region}, \mathbf{b}_{i,region})$$

with

$$\boldsymbol{\mu}_{i,region} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_4 \end{pmatrix} \begin{pmatrix} \beta_{0,region} \\ \beta_{1,region} \end{pmatrix} + \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_4 \end{pmatrix} \begin{pmatrix} b_{0,region} \\ b_{1,region} \end{pmatrix},$$

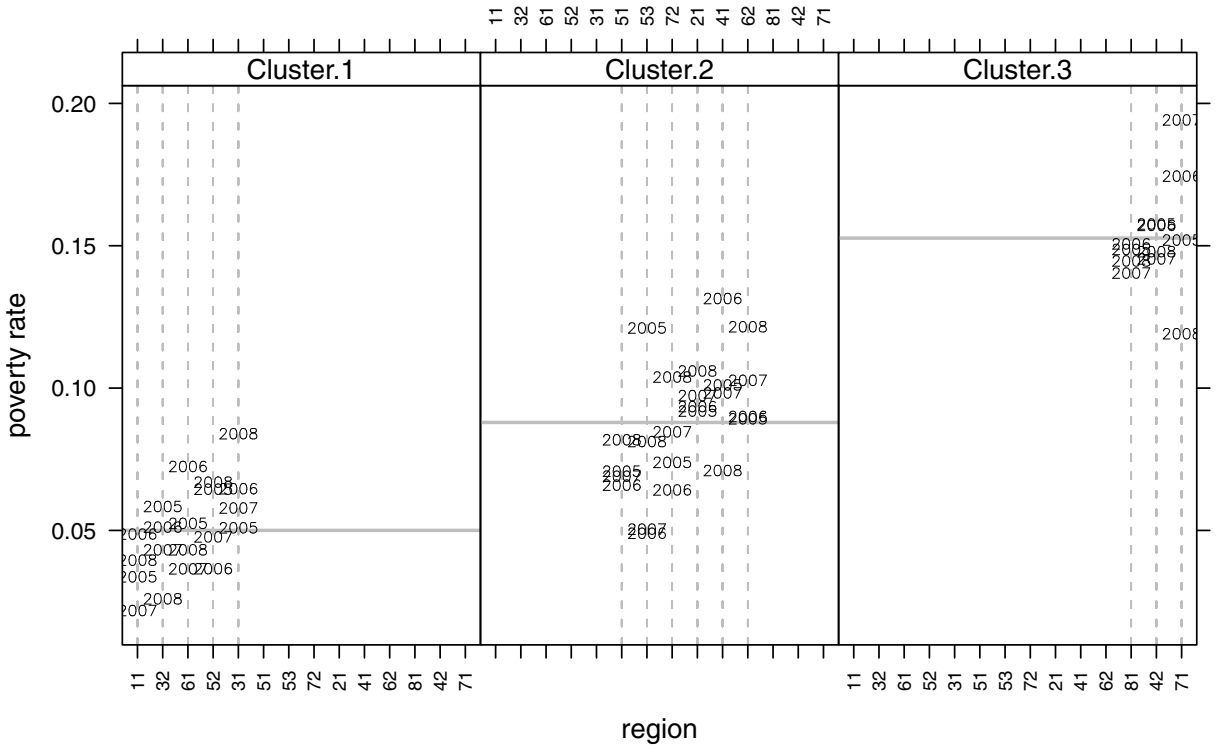


Figure 2: At-risk-of-poverty rate and classification of NUTS3 regions into three components. The regions are sorted by the mean poverty rates over years 2005–2008.

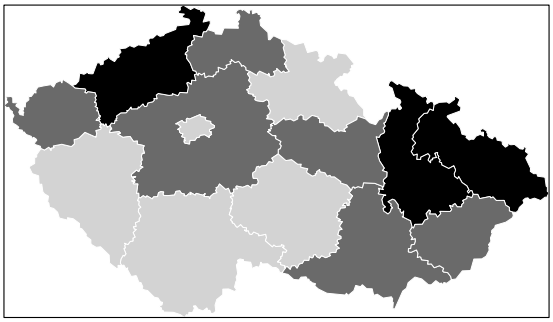


Figure 3: Classification of NUTS3 regions into three clusters: "*rich*" (light grey), "*middle*" (grey), "*poor*" (black)

Table 2: Estimated proportions

	Comp.1	Comp.2	Comp.3
π_k	0.357	0.429	0.214

Table 3: Czech NUTS3 regions

	Region		Region
11	Prague	52	Hradec Králové
21	Central Bohemia	53	Pardubice
31	South Bohemia	61	Vysočina
32	Plzeň	62	South Moravia
41	Karlovy Vary	71	Olomouc
42	Ustí nad Labem	72	Zlín
51	Liberec	81	Moravia-Silesia

where the vector $(t_1, \ldots, t_4)'$ contains the centered years and i denotes the i-th household within the region. We used the *mclust* package over all data for initial solution and *flexmix* package for each region. The number of components was again determined on the basis of *BIC* criteria. In this case a model with four component was selected. Results are plotted for different regions in the Figures 4, 5, 6.

As shown in Figures 4, 5, 6, the first three components contain almost all equalised income.

The fourth component only comprises between 0 and 8.2% of households. Using regression mixture, the *BIC* criterion divides households into four income categories: households with low income, average income, higher income and with extremely high income.

The relationship between household income categories and types of regions is described in Table 4. Type of region is mainly characterized by the percentage of households in the third income category.

Table 4: Percent range of households in each component for three types of regions

Regions	Comp.1	Comp.2	Comp.3	Comp.4
"*rich*"	14.2–33.3%	33–56.7%	25.4–42.3%	0–8.2%
"*middle*"	21–34.5%	49.1–58.9%	8.3–23.4%	2.6–6.6%
"*poor*"	23.9%–29.5%	55.1–63.9%	6.6–14.2%	1.1–1.2%

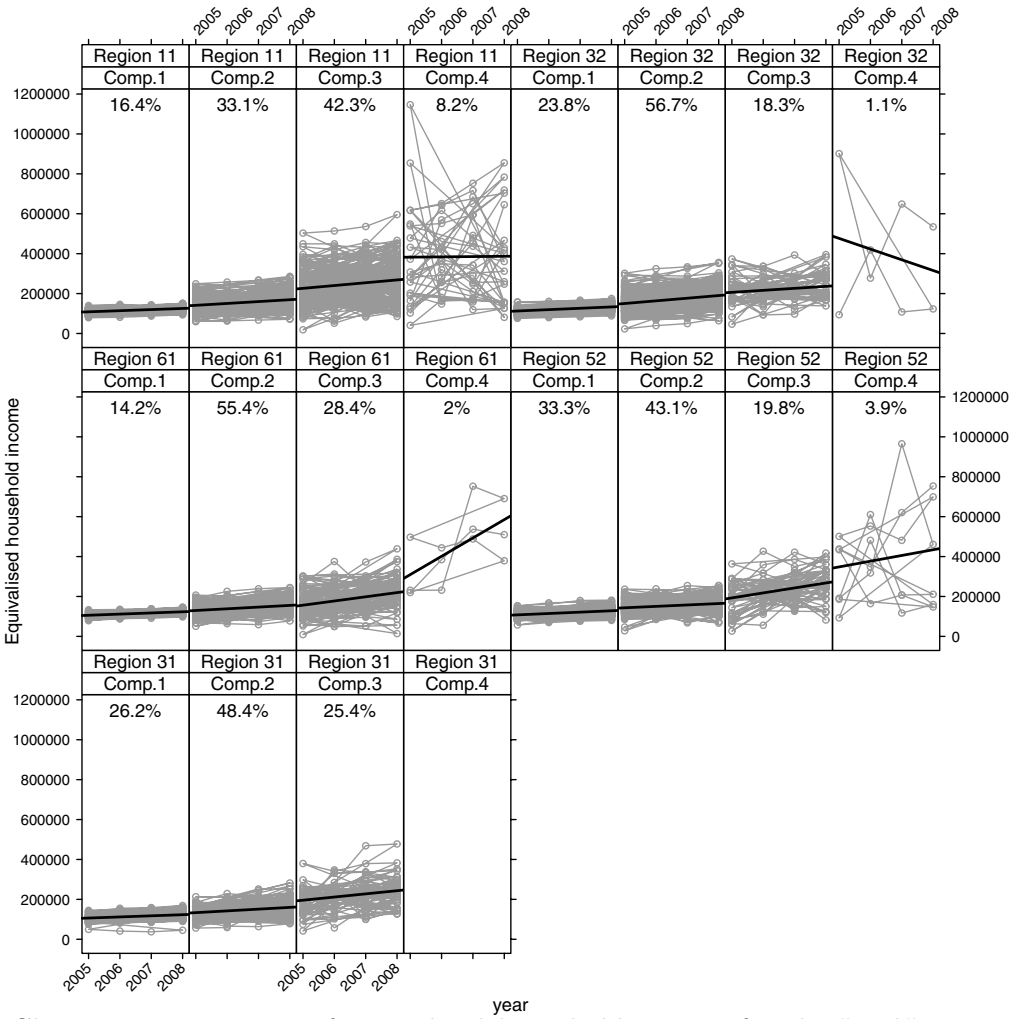


Figure 4: Clustering trajectories of equivalised household income for the "*rich*" regions into four components. Solid line is straight line with fixed parameters.

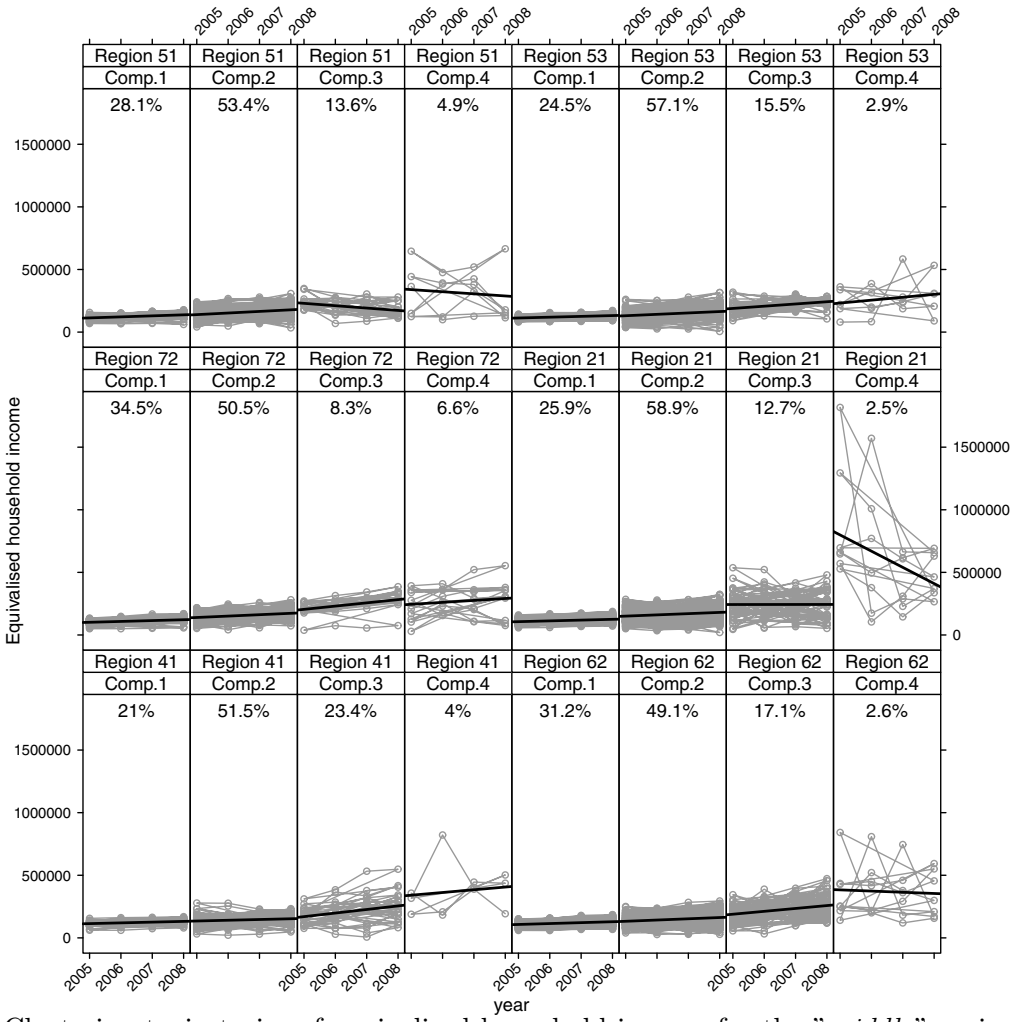


Figure 5: Clustering trajectories of equivalised household income for the "*middle*" regions into four components. Solid line is straight line with fixed parameters.

4 Conclusions

This paper introduces the use of regression mixture models in exploring household income dynamics over short time period. Within the regression mixture approach, classification of regions based on the risk of monetary poverty was also carried out.

Acknowledgement

The paper was supported by grant from Grant Agency of Czech Republic no. 420/09/05015 with title "Analysis and modelling of financial power of Czech and Slovak Households".

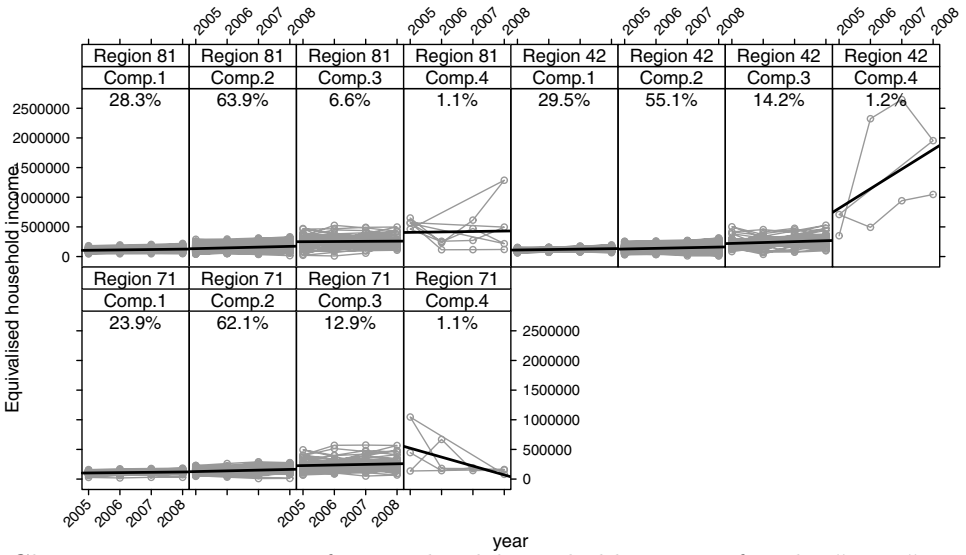


Figure 6: Clustering trajectories of equivalised household income for the "*poor*" regions into four components. Solid line is straight line with fixed parameters.

References

[1] BARTOŠOVÁ, J., FORBELSKÁ, M.: *Mixture Model Clustering for Household Incomes*. Aplimat - Journal of Applied Mathematics, Bratislava : Slovak University of Technology in Bratislava, 3, 163-172, 2010.

[2] CELEUX, G., LAVERGNE, C. , MARTIN, O.: *Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments*. Statistical Modelling 5, 243-267, 2005.

[3] DEMIDENKO, E.: *Mixed models: theory and applications*. NewYork¿ John Wiley & Sons. 2004.

[4] 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.

[5] Eurostat. *Description of Target Variables: Cross-sectional and Longitudinal*. EU-SILC 065/04, Eurostat, Luxembourg. 2004.

[6] FITZMAURICE, G. M., LAIRD, N. M. , WARE, J. H.: *Applied Longitudinal Analysis*. Wiley, New York. 2004.

[7] FORBELSKÁ, M.: *Modelování heterogenity ročních příjmů českých domácností. Analýza a modelování finančního potenciálu českch (slovenských) domácností*. Jindřichův Hradec : Vysoká škola ekonomická v Praze, Fakulta managementu Jindřichův Hradec, 14 s., 2009.

[8] FORBELSKÁ, M., BARTOŠOVÁ, J.: *Clustering of Czech Household Incomes Over Very Short Time Period*. Proceedings of COMPSTAT'2010. Book of abstract. 19th International Conference on Computational Statistics, 1023-1030, 2010.

[9] FRALEY, C., RAFTERY, A. E.: MCLUST: *Normal Mixture Modeling and Model-Based Clustering*. R package version 3.0-0; 2006.

[10] GELMAN, A. and HILL, J.: *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge. 2006.
[11] GRUEN, B., LEISCH, F.: *Fitting finite mixtures of generalized linear regressions in R*. Computational Statistics & Data Analysis, 51(11), 5247-5252, 2007.
[12] GRUEN, B., LEISCH, F.: *FLEXMIX: Version 2. Finite mixtures with concomitant variables and varying and constant parameters*. Journal of Statistical Software, 28(4), 1-35, 2008.
[13] JIANG, J.: *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, Berlin. 2007.
[14] LEISCH, F.: *FlexMix: A general framework for finite mixture models and latent class regression in R*. Journal of Statistical Software, 11(8), 1-18, 2004.
[15] McCULLAGH, P., NELDER, J. A.: *Generalized Linear Models. Chapman and Hall*, London 1994.
[16] McCULLOCH, C. E., SEALRE, S. R.: Generalized, Linear, and Mixed Models. Wiley, New York, 2001.
[17] McLACHLAN, G. J. , PEEL, D.: *Finite mixture models*. New York: Wiley & Sons, 2000.
[18] MOLENBERGHS, G., VERBEKE, G.: *Models for Discrete Longitudinal Data*. Springer, Berlin. 2005.
[19] NELDER, J. A., WEDDERBURN, R. W. M.: *Generalized linear models*. Journal of the Royal Statistical Society - Series A 135(3): 370384. 1972
[20] R Development Core Team: *R: A language and environment for statistical computing*. R. Foundation for Statistical Computing, Vienna, Austria. 2010. URL http://www.R-project.org
[21] RAUDENBUSH, S. W., BRYK, A. S.: *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA. 2001.
[22] SCHWARTZ, G.: *Estimating the Dimension of a Model*. In The Annals of Statistics, 6 (2), pp. 461-464, 1978.
[23] SEARLE, S., CASELLA, G., McCULLOCH, C.: *Variance components*. John Wiley and sons, Inc., New York, NY. 1992.
[24] VERBEKE, G., MOLENBERGHS , G.: *Linear Mixed Models for Longitudinal Data*. Springer, Berlin. 2001.
[25] WITKOVSKY, V.: *Matlab algorithm mixed.m for solving Henderson's mixed model equations*. Technical Report, Inst. of Measurement Science, Slovak Academy of Science, 2001.

Current address

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: forbel@math.muni.cz

Journal of Applied Mathematics

volume IV (2011), number III

LFLF FORECASTER AS NEW TOOL FOR TIME SERIES PREDICTION

HABIBALLA, Hashim, (CZ), PAVLISKA, Viktor, (CZ), DVOŘÁK, Antonín (CZ)

Abstract. The article describes software tool for prediction of time series based on various methods derived from F-transformation and linguistic rules. The tool has been created on Institute for Research and Applications of Fuzzy Modeling, University of Ostrava. It is developing into a really powerful application giving good results in some cases.

Key words. Fuzzy logic, time series prediction, F-transform, linguistic rules.

Mathematics Subject Classification: Primary 37M10; Secondary 94D05.

1. Introduction

Time series analysis and prediction is an important task that can be used in many areas of practice. The task of getting the best prediction to given series may bring interesting engineering applications in wide number of areas like economics, geography or industry. Solution to the problem of obtaining best results in prediction of time series can be based on well-known and simple methods like Winters or Linear method [2]. In this paper we present a tool based on two methods originally developed by members of Institute for Research and Applications of Fuzzy Modeling. The aim of the paper is not to present the details of the methods already published, but to present a tool implementing them. The first method is based on the notion of F-transform devised by the group of Prof. Perfilieva [3]. The second approach use the linguistic rules utilizing fuzzy logic and deduction that is a well-known formalism with very good results in variety of practical applications like industrial ones.

2. F-transform

The core idea of the F-transform technique is a fuzzy partition of the universe. It can be simply presented like set of intervals fulfilling some criteria. It is described in the following definition that is stated in [4].

Definition 1
Let $x_1 < \ldots < x_n$ be fixed nodes within $[a, b]$, such that $x_1 = a$, $x_n = b$ and $n \geq 2$. We say that fuzzy sets $A_1, \ldots, A_n$, identified with their membership functions $A_1(x), \ldots, A_n(x)$ defined on $[a, b]$, form a *fuzzy partition* of $[a, b]$ if they fulfil the following conditions for $k = 1, \ldots, n$:

(1) $A_k : [a, b] \longrightarrow [0, 1]$, $A_k(x_k) = 1$;
(2) $A_k(x) = 0$ if $x \notin (x_{k-1}, x_{k+1})$ where for the uniformity of denotation, we put $x_0 = a$ and $x_{n+1} = b$;
(3) $A_k(x)$ is continuous;
(4) $A_k(x)$, $k = 2, \ldots, n$, monotonically increases on $[x_{k-1}, x_k]$ and $A_k(x)$, $k = 1, \ldots, n-1$, monotonically decreases on $[x_k, x_{k+1}]$;
(5) for all $x \in [a, b]$

$$\sum_{k=1}^{n} A_k(x) = 1. \tag{1}$$

The membership functions $A_1(x), \ldots, A_n(x)$ are called *basic functions*.

These partitions form a base for F-transform which lead to the tuple of numbers representing original transformated function. The n-tuple can be obtained using the following notion.

Definition 2
Let $f \in V_l$ be given and $A_1, \ldots, A_n$, $n < l$, be basic functions which constitute a fuzzy partition of $[a, b]$. We say that the n-tuple of real numbers $[F_1, \ldots, F_n]$ is the F-transform of f with respect to $A_1, \ldots, A_n$ if

$$F_k = \frac{\sum_{j=1}^{l} f(p_j) A_k(p_j)}{\sum_{j=1}^{l} A_k(p_j)}. \tag{3}$$

To forecast a time series we will use its F-transform representation and separately forecast the next component Y_{n+1} of the F-transform(of y_t) and a respective residuum. We will consider three methods for the forecasting a component of the F-transform: the F-transform of the second order an extrapolation of the inverse fuzzy transform and a logical deduction [1].

3. Linguistic rules

The theory of linguistic term and variables is well-known approach in the fuzzy logic community. It enables to work with rules containing terms of natural language like small or big and modifiers like very, roughly etc. The rule interpretation is then done by logical deduction based on which is based on fuzzy set theory and fuzzy logic to enable to deduce conclusions on the basis of imprecise description of the given situation using the linguistically formulated fuzzy IF-THEN rules [1].

The usage of this theory within a frame of time series prediction lies in the learning of these rules from the serie and then application to future (predicted) members of the serie. These learning

algorithms are already prepared within the LFLC software [1], which is intended to perform logical deduction on linguistic rules. The core of the system serve also for the presented application.

Example

An example fuzzy IF-THEN rule is

IF *the obstacle is near* AND *the speed of the car is high*
THEN *the breaking force is very strong*.

The "obstacle", "speed" and "breaking force" are variables while "near", "high" and "very strong" are expressions characterizing vaguely the magnitude of the variable.

The part before THEN is called the *antecedent* and the part after it the *succedent*. The variables $X_1, \ldots, X_n$ are called *input*, or *independent* variables. The variable Y is called *output*, or *dependent* variable.

The fuzzy IF-THEN rules are usually put together to form the *linguistic description*

$$\begin{aligned} \mathcal{R}_1 &:= \text{IF } X_1 \text{ is } \mathcal{A}_{11} \text{ AND } \cdots \text{ AND } X_n \text{ is } \mathcal{A}_{1n} \text{ THEN } Y \text{ is } \mathcal{B}_1 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathcal{R}_m &:= \text{IF } X_1 \text{ is } \mathcal{A}_{m1} \text{ AND } \cdots \text{ AND } X_n \text{ is } \mathcal{A}_{mn} \text{ THEN } Y \text{ is } \mathcal{B}_m \end{aligned}$$

4. Time series tool

The software for prediction of time series based on the previously presented formalisms is currently in the development, but first alpha version is now complete. It was implemented on MS-Windows platform in C++ using free GUI library WxWidgets. The predecessor of the tool has been an console application without GUI, but shares the same core like the tool.

The main tasks of the tool are the following:

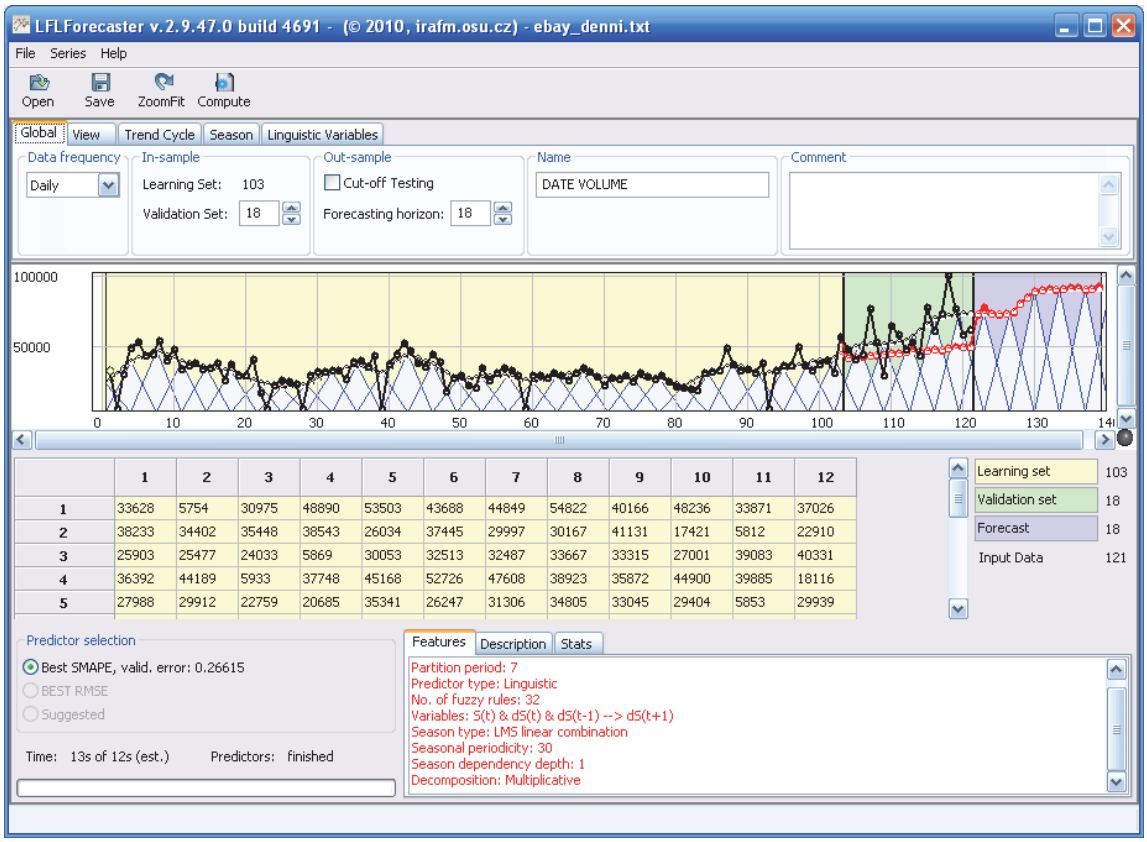
- Loading and presentation of a prepared file with a time serie in a graph.
- Setting up the methods and parameters for prediction.
- Computation of prediction according to the methods.
- Presentation and selection of predictions in a graph with reports generated during prediction (difference to the original serie etc.)
- Export of selected results.

The application is an SDI (single document interface) application divided into two main windows – Plot and Methods (Fig. 1). Plot window allows opening of user chosen serie through standard file open dialog (File menu). Then it performs desired predictions on loaded serie and presents it in a graph. The graph window contains two basic panels – graphical information panel and text information panel. Graphical information panel enables to present standard graph of the serie and predictions, it can be zoomed (important method is to fit the curve into graph which is connected to 's' key). Some basic graph operations can be obtained by Plot menu. Last menu item - Help - shows the application info and basics of application control (Fig.2).

Text information panel describes performed predictions projected in graphical information panel. It presents information concerning particular prediction in one line (color refers to the same color in graph). The information given could be presented on the following simple example:

No.4 Predictor, name: AvgTrend[k=5] {IntRule[v=9]}, error:0.0105458

It divides into predictor number, prediction description (methods used) – Trend (k = seasons number included for prediction), {method [specification]}, error of prediction. For our example it means:
4-th best prediction, trend is computed from average,[number of seasons included = 5]{logical deduction was used with [variable number = 9]}.



	1	2	3	4	5	6	7	8	9	10	11	12
1	33628	5754	30975	48890	53503	43688	44849	54822	40166	48236	33871	37026
2	38233	34402	35448	38543	26034	37445	29997	30167	41131	17421	5812	22910
3	25903	25477	24033	5869	30053	32513	32487	33667	33315	27001	39083	40331
4	36392	44189	5933	37748	45168	52726	47608	38923	35872	44900	39885	18116
5	27988	29912	22759	20685	35341	26247	31306	34805	33045	29404	5853	29939

Fig. 1:Time series tool layout

Methods window (Fig. 3) enables to set up the details for prediction process. It consists of four basic parts:

1. Trend computation selection.
2. Method computation selection.
3. Season part computation selection.
4. Operations with application – computation of prediction, export of the selected curve, export of all curves into file representation.

Trend could be selected either to be computed via standard average method or via inverse F-transform. There are basically four possibilities how to predict serie – standard Linear or standard

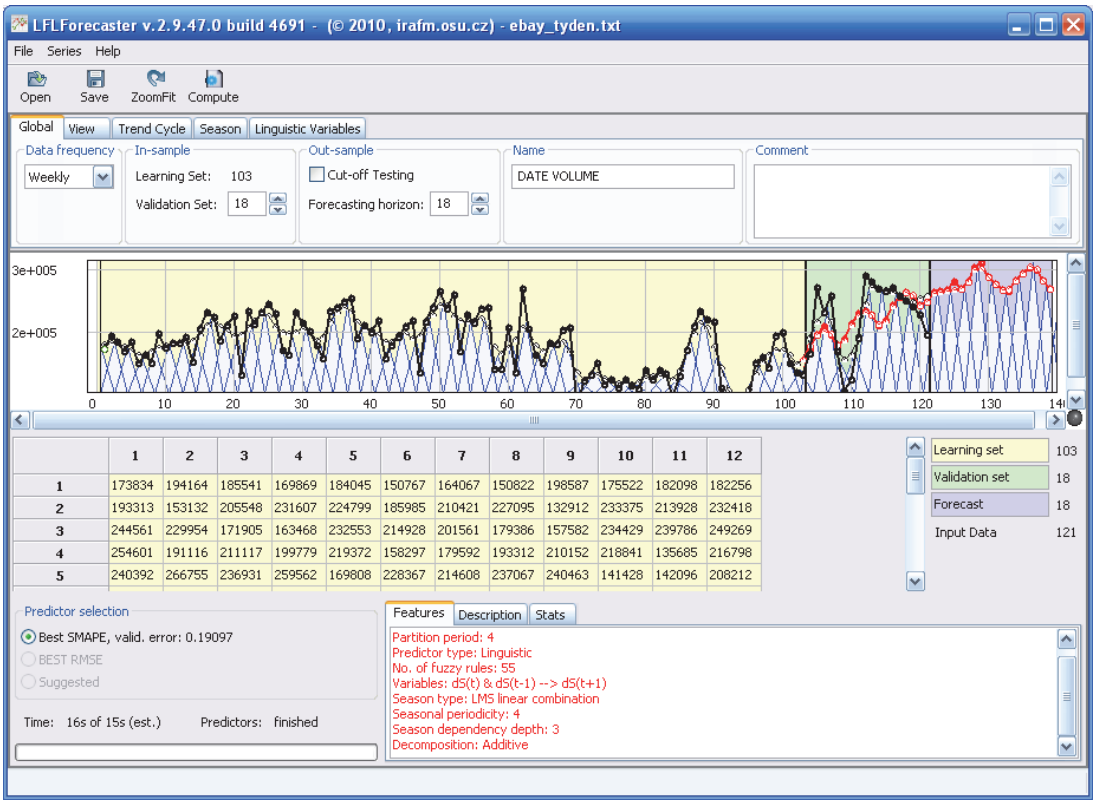


Fig. 2: Plot window

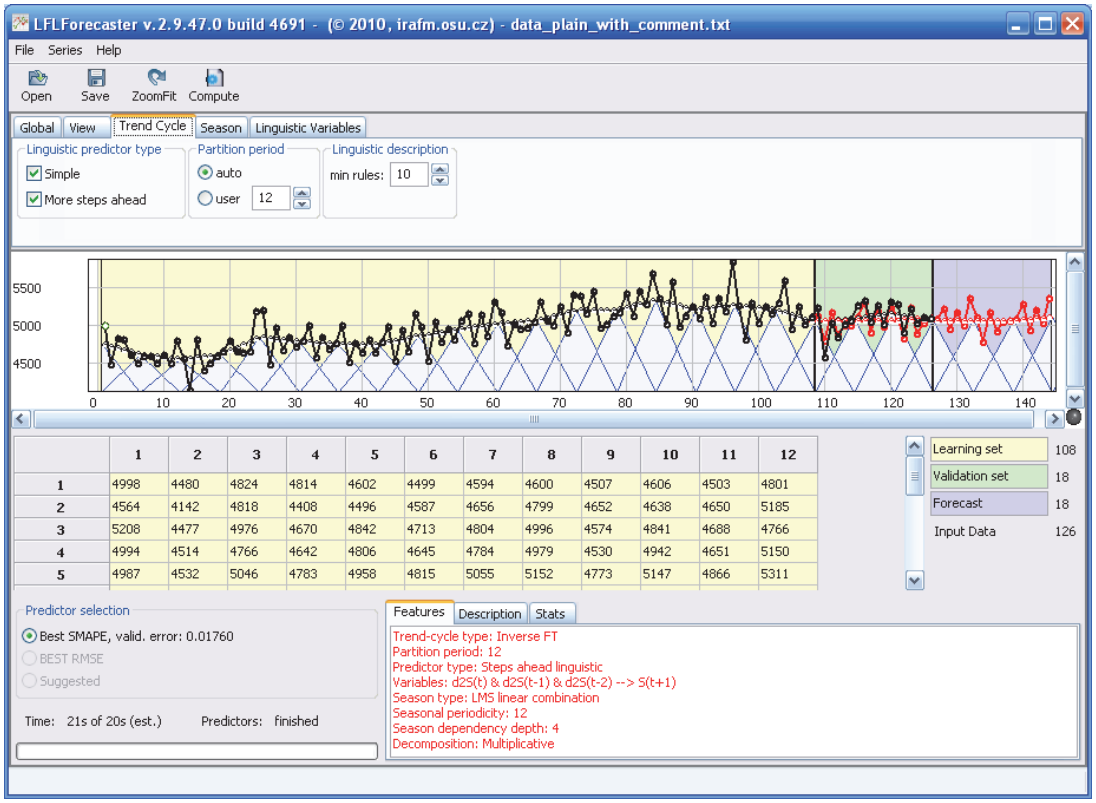


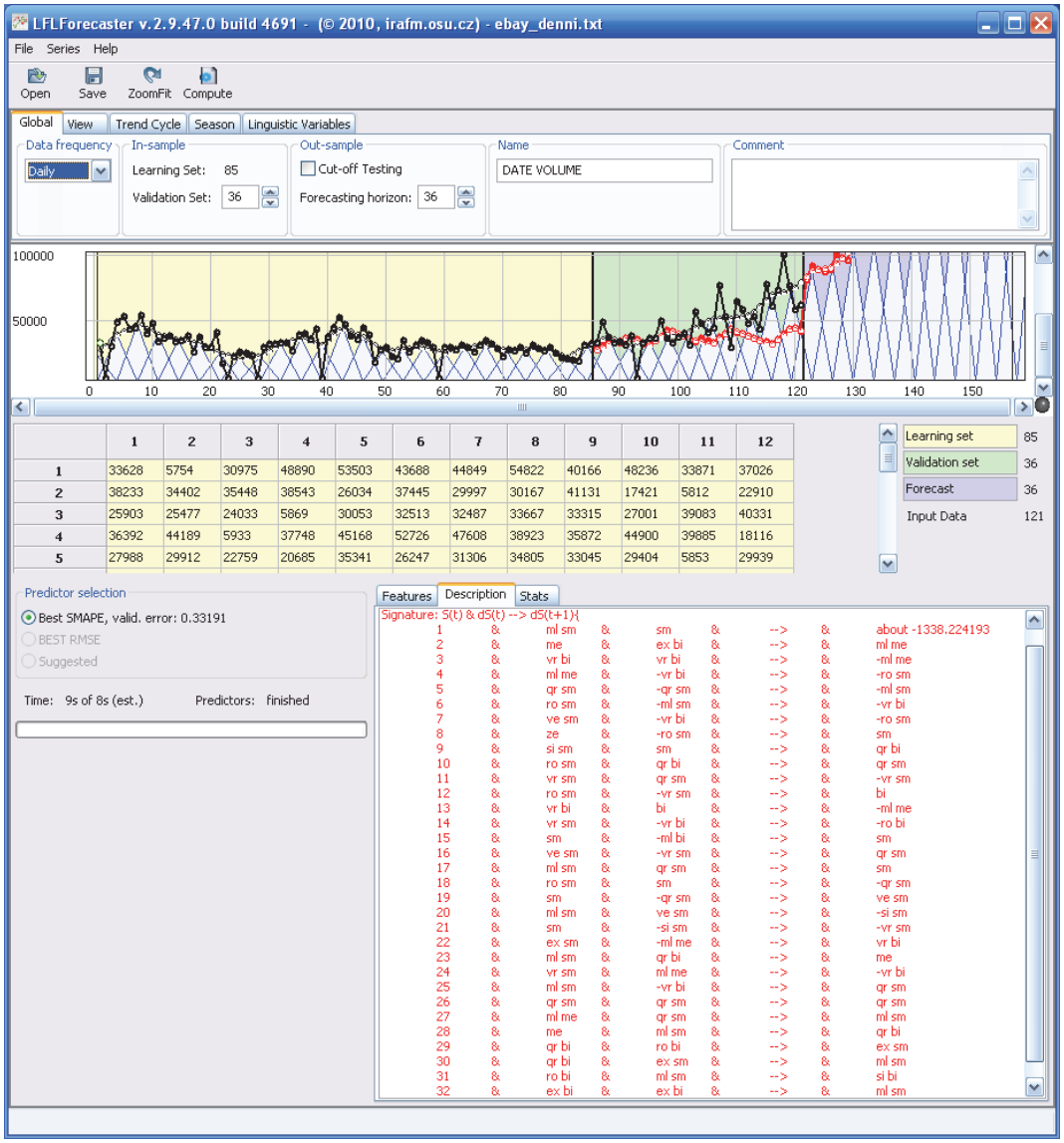
Fig. 3: Methods window

Winters method or new method of logical deduction or second order F-transform. Season computation can be additionally set up for season's dependency.

The input file with a serie should conform to a simple format, which could be observed from the following example:

```
NN3_101     4998  4480  4824  4814  4602  4499  4594  4600  4507  4606  4503
        4801  4564  4142  4818  4408  4496  4587  4656  4799  4652  4638  4650
```

First item should be name of the serie followed by unlimited number of real numbers delimited by a blank space (TAB, space, etc.).



5 Conclusion

The presented application is currently in initial stage of usage and experimentation. We believe it may bring interesting results in comparison with standard methods. Initial experiments show mainly promising values for logical deduction. Application will be soon deployed in its demo version and also there will be additions to the core of methods utilized. We present one of the applications of LFLForecaster in the following image.

Acknowledgement
The authors have been partially supported by the project MSM6198898701 and 1M0572 of the MSMT CR.

References

[1.] DVOŘÁK et. al.: The concept of LFLC 2000 - its specificity, realization and power of applications. Computers in Industry. 03/2003(51), Elsevier, Amsterdam, 2003, pp.269-280.
[2.] KŘIVÝ, I.: *Time-series*. University of Ostrava, Ostrava, 2005 (in czech).
[3.] PERFILIEVA, I., NOVÁK, V., DVOŘÁK, A. Fuzzy transform in the analysis of data. INT J APPROX REASON. 2008, sv. 48, s. 36-46. ISSN 0888-613X..
[4.] PERFILIEVA, I., NOVÁK, V., PAVLISKA, V., DVOŘÁK, A., ŠTĚPNIČKA, M. Analysis and Prediction of Time Series Using Fuzzy Transform. WCCI 2008 Proceedings. Hong Kong: IEEE Computational Intelligence Society, 2008. s. 3875-3879.

Current address

HABIBALLA Hashim, RNDr., PaedDr., Ph.D., PhD.
PAVLISKA Viktor, Mgr.
DVOŘÁK Antonín, Ing., Ph.D.
Institute for Research and Application of Fuzzy Modeling,
University of Ostrava,
30. dubna 22, 70103 Ostrava 1, CZ,
tel.: +420608886130,
email: hashim.habiballa@osu.cz

Journal of Applied Mathematics

volume IV (2011), number III

SARIMA MODELS FOR TEMPERATURE AND PRECIPITATION TIME SERIES IN THE CZECH REPUBLIC FOR THE PERIOD 1961–2008

HELMAN Karel, (CZ)

Abstract. The paper offers an analysis of monthly average temperature and precipitation sum time series recorded at 44 measurement stations in the Czech Republic over the period of 1961–2008. The two objectives to be achieved are the construction of SARIMA models based on Box-Jenkins methodology and a comparison of different models constructed according to the given factors of particular measurement stations' elevation, longitude and latitude.

Key words. temperatures, precipitation, Box-Jenkins, SARIMA models, time series, Czech Republic

Mathematics Subject Classification: Primary 62M10; Secondary 62P12

1 Introduction

The time series analysed in this paper are represented by monthly time series of average temperatures (in degrees Celsius) and precipitation amounts (in millimetres) taken from 44 measurement stations in the Czech Republic (i.e. 88 time series altogether) in the period between 1961 and 2008. The detailed description of measurement stations location can be found in [3]. The elevation of the selected measurement stations ranges from 158 metres above sea level (Doksany station) to 1322 metres above sea level (Lysá hora station). The southernmost station is that of Lednice (48°47'34"), the northernmost station is Bedřichov (50°48'54"). The Czech Republic "far west" region is represented by the station in the town of Aš (12°10'47") and the easternmost station is located at Lysá hora (18°26'52"). The input data were gained from the database of the Czech Hydrometeorological Institute CLIDATA. Other valuable information about this rich source of data as well as additional references to resources in English can be found e.g. in [4]. The location of the chosen measurement stations can be seen in Fig. 1 below.

The use of particular statistical tools derived from the so-called "Box-Jenkins methodology" in the climate and meteorological science is not common in the Czech Republic. *SARIMA* models were constructed, for instance, in [8], where the *SARIMA*(1,1,1)x(0,1,1)12 model was identified for

monthly average temperatures from Čáslav measurement station for the period 1876–1996. For time series of average year temperatures taken at Praha-Klementinum measurement station, model *AR*(9) in [5] was found.

When analysing the global warming phenomenon in [12], models with the term *AR*(9) were also constructed. The authors were working with global average yearly temperature time series and constructed *ARIMA*(9,1,0) and *ARIMA*(9,1,2) models.

A wider and more relevant use of *SARIMA* models can be traced in the field of hydrology and water management. (Perhaps the most important issue is the drought analysis. In work [1], where monthly streamflow data from an Ethiopian river were analysed, the *SARIMA*(0,1,1)x(0,1,1)12 model outperformed other models when capturing the severity of drought in the area. The same issue, having used data from Iran, was dealt with in work [9].)

Moving from climatology and meteorology to environmental science, a lot of works using models derived from Box-Jenkins methodology can be found. (From recent publications we can mention, for instance, [10] and [6], where the authors constructed *SARIMA* models for fractions of the atmospheric Particulate Matter (PM) concentrations in Brazil. As an example of inventive application of *SARIMA* models, see work [11], where quarterly data about container transshipment at German ports were examined.)

Figure 1. The location of the selected measurement stations in the Czech Republic territory

2 SARIMA models theory

In general, in a serial correlation theory, we will deal with specifications of the form

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta} + u_t \\ u_t &= \mathbf{z}_{t-1}'\boldsymbol{\gamma} + \varepsilon_t, \end{aligned} \tag{1}$$

where $\mathbf{x}_t$ is a vector of explanatory variables observed at time t, $\mathbf{z}_{t-1}$ is a vector of variables known in the previous period, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of parameters, u_t is a disturbance term and ε_t innovation in disturbance. Vector $\mathbf{z}_{t-1}$ may contain lagged values of u or those of ε or both. If there are no explanatory variables involved in the model, it is possible to replace u_t with y_t in the equations below.

The above mentioned Box-Jenkins methodology, or *ARIMA* models theory, is developed in work [2]. *ARIMA* (autoregressive integrated moving average) models are generalizations of a simple AR model that uses three tools for modelling serial correlation in disturbance. The first tool is an autoregressive, or *AR* term. Each *AR* term corresponds to the use of a lagged value of the residual in a forecasting equation for the unconditional residual. The autoregressive model of order p, *AR(p)* has the following form:

$$u_t = \varphi_1 u_{t-1} + \varphi_2 u_{t-2} + \ldots + \varphi_p u_{t-p} + \varepsilon_t \tag{2}$$

or, alternatively, with the use of a lag operator B:

$$(1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p) u_t = \varphi_p(B) u_t = \varepsilon_t , \tag{3}$$

where for B holds

$$B^j u_t = u_{t-j} . \tag{4}$$

The next tool is an integration order term. Each integration order corresponds to the differentiation of the series being forecast. The first-order integrated component means that the forecasting model is designed for the first difference of the original series. The second-order component corresponds to the second difference, etc. The third tool is *MA*, or a moving average term. The moving average forecasting model uses lagged values of a forecast error to improve the current forecast. The first-order moving average term uses the most recent forecast error, the second-order term uses the forecast error from two most recent periods, etc. *MA(q)* has the form:

$$u_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} . \tag{5}$$

Equivalently, it can be rewritten by the lag operator as follows:

$$u_t = (1 - \theta_1 B - \ldots - \theta_q B^q) \varepsilon_t = \theta_q(B) \varepsilon_t . \tag{6}$$

When modelling time series with systematic seasonal movements – which is the case of monthly average temperatures and monthly precipitation sums – Box and Jenkins in [2] recommended the use of seasonal autoregressive (*SAR*) and seasonal moving average (*SMA*) terms. The seasonal autoregressive model of order P can be written as

$$u_t = \Phi_1 u_{t-S} + \Phi_2 u_{t-2S} + \ldots + \Phi_P u_{t-SP} + \varepsilon_t \tag{7}$$

or

$$\Phi_P(B^S) u_t = \varepsilon_t . \tag{8}$$

The seasonal moving average model of order Q can be as written as

$$u_t = \varepsilon_t - \Theta_1 \varepsilon_{t-S} - \ldots - \Theta_Q \varepsilon_{t-SQ} \tag{9}$$

or, equivalently,

$$u_t = \Theta_Q(B^S)\varepsilon_t\,. \tag{10}$$

In all the four equations above, S denotes the length of seasonality, which is for the time series analysed in this paper equal to number 12.

Finally, we can write the most general *SARIMA(p,d,q)(P,D,Q)S* -with-constant model as

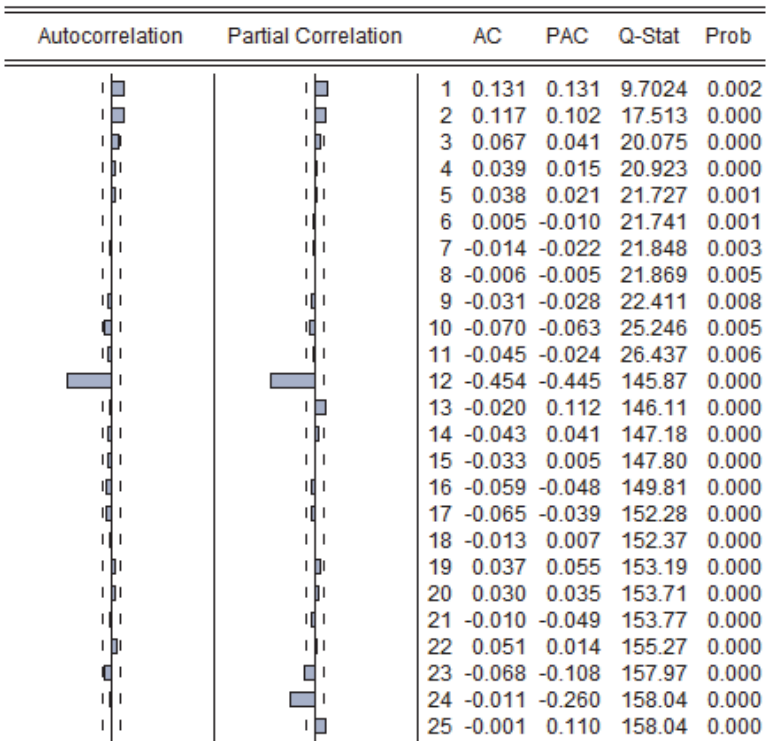
$$\varphi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D u_t = \varphi_0 + \theta_q(B)\Theta_Q(B^S)\varepsilon_t\,, \tag{11}$$

where the constant equals

$$\varphi_0 = \mu[(1-\varphi_1-\varphi_2-\ldots-\varphi_p)(1-\Phi_1-\Phi_2-\ldots-\Phi_p)]\,. \tag{12}$$

3 Model construction procedure, quick walkthrough

Monthly average temperature time series in the period 1961–2008 really show an increasing linear trend and seasonality with the length equal to 12, so we can start modelling by using seasonal differentiation where D = 1. Further, as the main tool, we use mainly the (residual) correlogram visual analysis method. In the correlogram, the values of the (residual) autocorrelation function (AC), partial autocorrelation function (PAC) and Ljung-Box Q-statistics values can be seen. An example of possible software output is presented in Fig. 2.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.131	0.131	9.7024	0.002
		2	0.117	0.102	17.513	0.000
		3	0.067	0.041	20.075	0.000
		4	0.039	0.015	20.923	0.000
		5	0.038	0.021	21.727	0.001
		6	0.005	-0.010	21.741	0.001
		7	-0.014	-0.022	21.848	0.003
		8	-0.006	-0.005	21.869	0.005
		9	-0.031	-0.028	22.411	0.008
		10	-0.070	-0.063	25.246	0.005
		11	-0.045	-0.024	26.437	0.006
		12	-0.454	-0.445	145.87	0.000
		13	-0.020	0.112	146.11	0.000
		14	-0.043	0.041	147.18	0.000
		15	-0.033	0.005	147.80	0.000
		16	-0.059	-0.048	149.81	0.000
		17	-0.065	-0.039	152.28	0.000
		18	-0.013	0.007	152.37	0.000
		19	0.037	0.055	153.19	0.000
		20	0.030	0.035	153.71	0.000
		21	-0.010	-0.049	153.77	0.000
		22	0.051	0.014	155.27	0.000
		23	-0.068	-0.108	157.97	0.000
		24	-0.011	-0.260	158.04	0.000
		25	-0.001	0.110	158.04	0.000

Figure 2. AC and PAC of seasonally differentiated monthly average time series, H2USTI01 measurement station

Ljung-Box test publicized in [7] uses a test statistic in the form

$$Q_{LB} = T(T+2)\sum_{j=1}^{k}\frac{\tau_j^2}{T-j}, \tag{13}$$

where T is the number of observations, τ_j is the sample autocorrelation in lag j and k is the number of lags to be tested. Q-statistic (13) in lag k is the test statistic for the null hypothesis about zero autocorrelation up to lag k. Q has the asymptotic chi-square distribution with degrees of freedom equal to k reduced by the number of *MA* and *AR* terms employed in the model.

As can be seen in Fig. 2, most statistically significant are the values of AC and PAC in lag 12 and the value of PAC in lag 24. So the next step is the employment of the *SMA*(12) term. Or alternatively, Q = 1 and S = 12 in the equation (10) term. Now we can speak about the *SARIMA*(0,0,0)x(0,1,1)12 model. In the next step – according to the residual correlogram – some term for the lag 1 should be added to the model. Two types of criteria were used:

1) primary: statistical significance of the estimated coefficient on the 5% level considered according to the *t*-test
2) secondary: maximization of the adjusted coefficient of determination and/or minimization of the Akaike information criterion.

Secondary criterions were used when there was no possibility to decide after a primary criterion had been used, or when it was necessary to decide which term was the best one to be added to the model in a certain phase.

4 Results and discussion

4.1 Monthly average temperatures

First, let us have a look at the results for *SARIMA* models of monthly average temperature time series. Table 1 contains the summary of estimated parameters for all 44 measurement stations sorted out from the lowest to highest elevation. In the first column there is an identification symbol for each measurement station. In the second column can be found a model that meets the requirement of non-significant Q-statistics for both all the lags investigated and coefficients estimated as significant on the 5% significance level in both cases. This model is then rewritten in the third column. In the last (fourth) column, there is a list of other terms that can be transposed to the model on the 5% significance level. If there were only one or two terms respectively, the values of estimated coefficients can be found in the table as well. In Fig. 3, the data presented in Table 1 can be shown in a different way: a percentage of *AR* and *MA* terms (except for lags 1 and 12) employed in *SARIMA* models of monthly average temperatures according to a particular lag is noticeable.

As can be seen from Table 1, the "base" of *SARIMA* models was the same for all the models constructed for monthly average temperature time series in the period 1961–2008; *SARIMA*(1,0,0)x(0,1,1)12-with-constant. When pursuing the goal of non-significant Q-statistics for all lags, in approximately one quarter of measurement stations the employment of terms with lag 2 or 3 was necessary.

A somewhat unexpected result is the possibility of the employment (at 5% significance level) of terms with lag 13 in the models for all the measurement stations except the Klatovy station. In most cases it is the *MA*(13) term. In the same way, for two-thirds of measurement stations the use of terms with lag 26 is possible.

Table 1. *SARIMA* models for monthly average temperature time series, measurement stations from the lowest to highest elevation

Station ID	Model, Q-statistics not significant, 5% significance level	Model value of y_t	Other possible terms, 5% significance level[1]
U1DOKS01[2]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.038 + 0.163y_{t-1} + y_{t-12} - 0.163y_{t-13} - 0.969\varepsilon_{t-12} + \varepsilon_t$	MA(13), MA(16), AR(26)
B2LEDN01[3]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.032 + 0.183y_{t-1} + y_{t-12} - 0.183y_{t-13} - 0.965\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.099, MA(23) = -0.109
P2BRAN01[3]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.035 + 0.160y_{t-1} + y_{t-12} - 0.160y_{t-13} - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.101, AR(26) = -0.085
O3PRER01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.026 + 0.214y_{t-1} + y_{t-12} - 0.214y_{t-13} - 0.965\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.105
O2OLOM01	SARIMA(1,0,0)x(0,1,1)12 with const.., MA(2)	$\ldots + 0.213y_{t-1} \ldots\ldots - 0.963\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.113
U1ZAT001[4]	SARIMA(1,0,0)x(0,1,1)12 with const.., AR(3)	$\ldots + 0.177y_{t-1} \ldots\ldots - 0.961\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.101, AR(26) = -0.083
B1HOLE01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.029 + 0.202y_{t-1} + y_{t-12} - 0.202y_{t-13} - 0.964\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.11
P1PKAR01[3]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.035 + 0.143y_{t-1} + y_{t-12} - 0.143y_{t-13} - 0.968\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.095, AR(26) = -0.098
P2SEMC01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.027 + 0.165y_{t-1} + y_{t-12} - 0.165y_{t-13} - 0.968\varepsilon_{t-12} + \varepsilon_t$	MA(13), MA(16), MA(26), MA(36)
B2BTUR01[5]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.031 + 0.199y_{t-1} + y_{t-12} - 0.199y_{t-13} - 0.965\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.123
O1MOSN01[6]	SARIMA(1,0,0)x(0,1,1)12 with const.., MA(2)	$\ldots + 0.186y_{t-1} \ldots\ldots - 0.966\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.101, MA(36) = -0.111
O1OPAV01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.024 + 0.213y_{t-1} + y_{t-12} - 0.213y_{t-13} - 0.963\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.113, AR(36) = -0.109
H3HRAD01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.031 + 0.163y_{t-1} + y_{t-12} - 0.163y_{t-13} - 0.966\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.111, AR(36) = -0.089
O1LUCI01[3]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.027 + 0.183y_{t-1} + y_{t-12} - 0.183y_{t-13} - 0.964\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.093, AR(36) = -0.092
U2JAPO01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.032 + 0.182y_{t-1} + y_{t-12} - 0.182y_{t-13} - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13), AR(16), AR(26), MA(36)
B2KUCH01[7]	SARIMA(2,0,0)x(0,1,1)12 with const..	$\ldots + 0.167y_{t-1} \ldots\ldots - 0.966\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.119, MA(23) = -0.088
O3VALM01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.023 + 0.186y_{t-1} + y_{t-12} - 0.186y_{t-13} - 0.964\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.104, MA(36) = -0.088
P1PRUZ01[8]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.028 + 0.171y_{t-1} + y_{t-12} - 0.171y_{t-13} - 0.965\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.115, AR(26) = -0.084
C2CBUD01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.034 + 0.150y_{t-1} + y_{t-12} - 0.150y_{t-13} - 0.968\varepsilon_{t-12} + \varepsilon_t$	AR(13), MA(23), MA(26), MA(36)
U2LIBC01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.03 + 0.178y_{t-1} + y_{t-12} - 0.178y_{t-13} - 0.968\varepsilon_{t-12} + \varepsilon_t$	AR(13) = 0.107, AR(26) = -0.104
H2USTI01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.041 + 0.167y_{t-1} + y_{t-12} - 0.167y_{t-13} - 0.966\varepsilon_{t-12} + \varepsilon_t$	AR(13) = 0.091, AR(26) = -0.085
L1KLAT01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.029 + 0.138y_{t-1} + y_{t-12} - 0.138y_{t-13} - 0.966\varepsilon_{t-12} + \varepsilon_t$	AR(23) = -0.112, AR(26) = -0.097
B2VMEZ01[2]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.029 + 0.157y_{t-1} + y_{t-12} - 0.157y_{t-13} - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13), MA(16), MA(26)
P3HAVL01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.034 + 0.155y_{t-1} + y_{t-12} - 0.155y_{t-13} - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13), AR(16), MA(26), MA(36)
C2TABO01[3]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.02 + 0.161y_{t-1} + y_{t-12} - 0.161y_{t-13} - 0.968\varepsilon_{t-12} + \varepsilon_t$	AR(13) = 0.104, AR(26) = -0.086
L1DOMA01[9]	SARIMA(1,0,0)x(0,1,1)12 with const.., MA(3)	$\ldots + 0.128y_{t-1} \ldots\ldots - 0.968\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.126, AR(26) = -0.092
L2KRAL01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.037 + 0.139y_{t-1} + y_{t-12} - 0.139y_{t-13} - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13), AR(26), MA(36)
L1NEPO01[10]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.029 + 0.145y_{t-1} + y_{t-12} - 0.145y_{t-13} - 0.968\varepsilon_{t-12} + \varepsilon_t$	MA(13), AR(23), AR(26), MA(36)
L3CHEB01[10]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.032 + 0.151y_{t-1} + y_{t-12} - 0.151y_{t-13} - 0.966\varepsilon_{t-12} + \varepsilon_t$	MA(13), AR(23), AR(26), MA(36)
P3ONDR01[7]	SARIMA(2,0,0)x(0,1,1)12 with const..	$\ldots + 0.147y_{t-1} \ldots\ldots - 0.968\varepsilon_{t-12} + \varepsilon_t$	MA(13), AR(26), MA(36)
P3PRIB01[8]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.029 + 0.155y_{t-1} + y_{t-12} - 0.155y_{t-13} - 0.966\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.114, AR(26) = -0.083
B2KMYS01[11]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.032 + 0.131y_{t-1} + y_{t-12} - 0.131y_{t-13} - 0.968\varepsilon_{t-12} + \varepsilon_t$	MA(13), MA(23), MA(26), MA(32), MA(36)
B2BYSP01[12]	SARIMA(1,0,0)x(0,1,1)12 with const.., MA(2)	$\ldots + 0.150y_{t-1} \ldots\ldots - 0.966\varepsilon_{t-12} + \varepsilon_t$	MA(3), MA(13), MA(26)
C2NADV01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.026 + 0.119y_{t-1} + y_{t-12} - 0.119y_{t-13} - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13), AR(23), AR(26), MA(36)
O2STKU01	SARIMA(1,0,0)x(0,1,1)12 with const.., MA(2)	$\ldots + 0.151y_{t-1} \ldots\ldots - 0.965\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.135, AR(25) = 0.083
L3AS0001[6]	SARIMA(1,0,0)x(0,1,1)12 with const.., MA(2)	$\ldots + 0.127y_{t-1} \ldots\ldots - 0.965\varepsilon_{t-12} + \varepsilon_t$	MA(13), MA(23), MA(25), MA(26), MA(36)
B2NEDV01[3]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.03 + 0.142y_{t-1} + y_{t-12} - 0.142y_{t-13} - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.12, AR(26) = -0.083
H3SVRA01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.034 + 0.124y_{t-1} + y_{t-12} - 0.124y_{t-13} - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13), MA(25), AR(26)
L2PRIM01[10]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.034 + 0.122y_{t-1} + y_{t-12} - 0.122y_{t-13} - 0.966\varepsilon_{t-12} + \varepsilon_t$	MA(13), AR(23), MA(26), MA(36)
O1CERV01[3]	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.03 + 0.156y_{t-1} + y_{t-12} - 0.156y_{t-13} - 0.965\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.11, MA(25) = 0.097
U2BEDR01[6]	SARIMA(1,0,0)x(0,1,1)12 with const.., MA(2)	$\ldots + 0.155y_{t-1} \ldots\ldots - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.106, MA(25) = 0.111
U1MIL001	SARIMA(2,0,0)x(0,1,1)12 with const..	$\ldots + 0.143y_{t-1} \ldots\ldots - 0.967\varepsilon_{t-12} + \varepsilon_t$	MA(13), MA(25), MA(26), MA(36)
C1CHUR01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.033 + 0.100y_{t-1} + y_{t-12} - 0.100y_{t-13} - 0.969\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.112, MA(25) = 0.137
O1LYSA01	SARIMA(1,0,0)x(0,1,1)12 with const..	$0.029 + 0.113y_{t-1} + y_{t-12} - 0.113y_{t-13} - 0.970\varepsilon_{t-12} + \varepsilon_t$	MA(13) = 0.12, MA(25) = 0.121

[1]The employment of these terms will slightly affect the value of estimated coefficients in the previous column.
[2]The employment of more terms leads to significant Q-statistic for lag six.
[3]The employment of more terms leads to significant Q-statistic for lag five.
[4]The coefficient for AR(3) term is significant only on 10% significance level.
[5]The employment of MA(13) term leads to significant Q-statistic for lag four.
[6]The coefficient for MA(2) term is significant only on 10% significance level.
[7]The coefficient for AR(2) term is significant only on 10% significance level.
[8]The employment of more terms leads to significant Q-statistic for lags five and six.
[9]The coefficient for MA(3) term is significant only on 10% significance level.
[10]The employment of more terms leads to significant Q-statistic for lag seven.
[11]The employment of more terms leads to significant Q-statistic for lag eight.
[12]The employment of more terms leads to coefficient for MA(2) term significant only on 10% significance level.

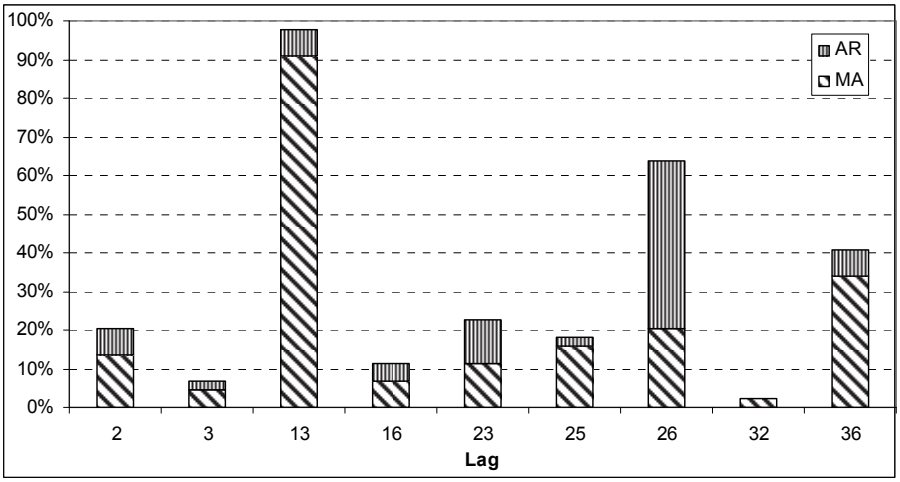


Figure 3. Percentage of *AR* and *MA* terms (except for lags 1 and 12) employment in SARIMA models of monthly average temperatures recorded at 44 measurement stations, according to the lag,

In all 44 models, however, *SMA*(12) and *AR*(1) terms are employed, so we can use "their" coefficients Θ_1 and ϕ_1 (to be found in the second and third column of Table 1) to examine whether the relation between the estimated model and elevation or geographical location really exists. The results of this part of the analysis can be seen in Fig. 4 and 5. In the case of longitude it is obvious according to Fig. 4 that the estimated values of the Θ_1 coefficient (blue colour) are mostly independent of the measurement stations´ longitude; only between 12° and 17° there is a certain parabolic dependence. With the exception of the easternmost measurement station Lysá hora (the highest located station with its elevation of 1322 meters above sea level) we can state that the values of the estimated coefficient ϕ_1 (red colour) tend to be higher for higher longitudes than for lower latitudes on average. In other words, the dependence of monthly average temperatures on the previous month value is on average higher at the stations located in the east of the Czech Republic than at those in the west.

But on the other hand – based on the measurement stations´ latitude – the following conclusions can be drawn. The estimated coefficients' values oscillate around a constant and the analysis shows the mutual independence of their values and the measurement stations´ latitude.

Fig. 5 shows the relation between the estimated *SARIMA* models' coefficients (as can be seen in the second and third column in Table 1) and the elevation of particular measurement stations. As can be seen in Fig. 5, while the estimated values of the Θ_1 coefficient (triangles) are independent of the measurement station elevation, the values of the coefficient ϕ_1 (points) tend to decrease with increasing measurement station elevation. This means that monthly average temperatures at the measurement stations with lower elevation are more dependent on the previous months´ average temperatures than those recorded at the measurement stations with higher elevations.

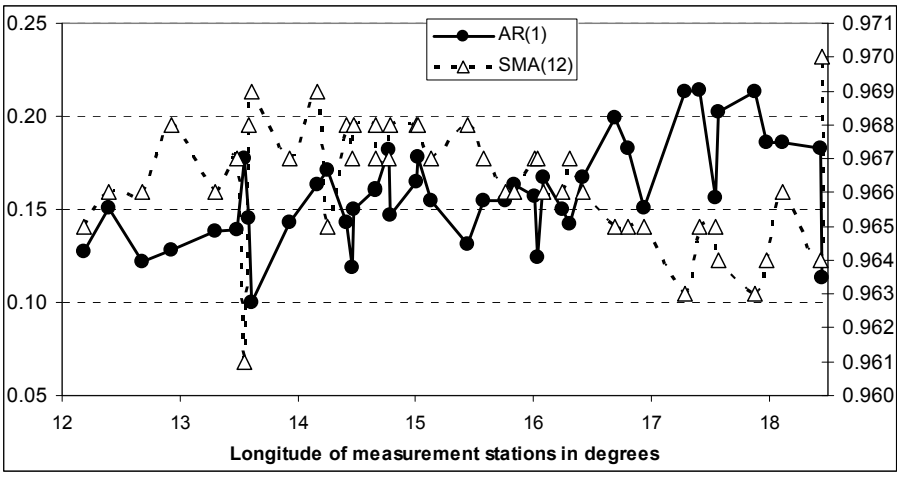


Figure 4. The values of estimated *AR*(1) and *SMA*(12) coefficients in *SARIMA* models of monthly average temperature time series according to the longitude of measurement stations

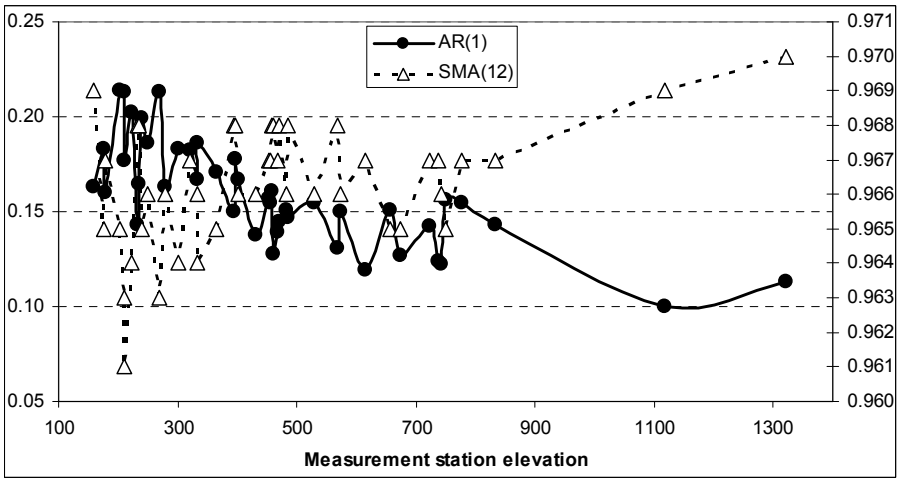


Figure 5. The values of estimated *AR*(1) and *SMA*(12) coefficients in *SARIMA* models of monthly average temperature time series according to measurement station elevation

4.2 Monthly precipitation sums

There are two reasons why modelling of monthly precipitation sum time series started by seasonal differentiation. Firstly, other procedures, e.g. the *SARIMA*(0,0,0)x(1,0,1)12 model, have led to a distinctively lower adjusted coefficient of determination (higher Akaike's information criterion) and secondly, seasonal differentiation allows a better comparison between the models for temperature and precipitation time series.

The same analysis as in the case of the temperature time series was also performed for the precipitation time series. A common "base" of *SARIMA* models for all the measurement stations in the latter case was the *SARIMA*(0,0,0)x(0,1,1)12 model. Further – in contrast with the models for the temperature time series – there are considerable differences between *SARIMA* models for the precipitation time series. This can be seen in Fig. 6.

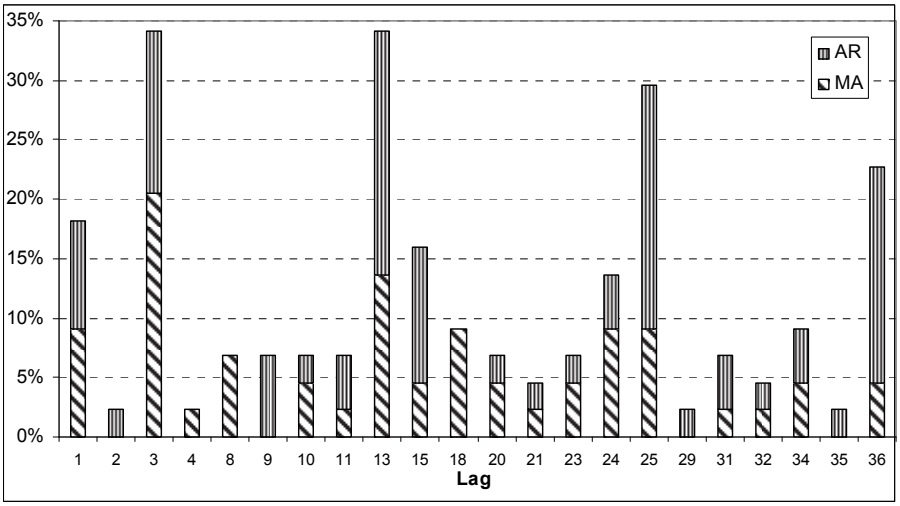


Figure 6. Percentage of AR and MA terms employment (except for lag 12) in SARIMA models of monthly precipitation sums recorded at 44 measurement stations, according to the lag,

When discussing the results, we can state that no dependence between the elevation of measurement stations and the appearance of constructed *SARIMA* models was found. However,, we got some unforeseen results when examining the dependence on the latitude and longitude of the measurement stations.

As for the latitude, we can see that eleven out of fifteen stations for which the lag 3 term was the part of the *SARIMA* model, belong to the fifteen southernmost stations. All three stations for which the term *MA*(8) was employed in the *SARIMA* model are situated in maximum proximity to one another – between 49°34'33" and 49°36'42". Six from other seven measurement stations for which the lag 15 term was the part of the model also lie in close proximity (between 49°34'58" and 49°58'49"). The three norhernmost stations are the only ones for which the lag 31 term became the part of the model, etc. Thus, the conclusion can be made that there is a considerable connection between the latitude of measurement stations and *SARIMA* models.

When investigating dependence on the longitude, we obtained similar results. Let us point out some of them. All fifteen measurement stations for which the lag 3 term was involved in the *SARIMA* model are among the westernmost half of the measurement stations. In the quaternion of neighbouring measurement stations (between 17°52'34" and 18°26'33") are all three stations with the lag 20 term in the *SARIMA* model, etc. Thus, as for the longitude, a similar conclusion can be drawn – a link between the longitude of measurement stations and *SARIMA* models exists.

Acknowledgement
Financial backing from the research grant supporting FIS VŠE long-term scientific and research development is gratefully acknowledged.

References

[1.] ABEBE, A., FOERCH, G.: Stochastic simulation of the severity of hydrological drought. *Water and Environment Journal*, Vol. 22, No. 1, pp. 2-10(9), March 2008.

[2.] BOX, G. E. P., JENKINS, G. M.: *Time Series Analysis: Forecasting and Control*. Revised Edition, Oakland, CA: Holden-Day, 1976.

[3.] HELMAN, K., Analýza vývoje sezónnosti vybraných srážkových časových řad v České republice pro období 1961 - 2008. Praha 17.09.2009 - 18.09.2009. In: *Mezinárodní statisticko-ekonomické dny na VŠE* [CD-ROM], Praha, 2009.

[4.] HELMAN, K., Analýza meteorologických časových řad s využitím databáze ČHMÚ CLIDATA. Praha 23.02.2006. In: *Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze*. Praha : Oeconomica, 158–166 (2006).

[5.] HLÁVKA, Z.: *Analýza hydrometeorologických časových řad* [Diplomová práce]. Univerzita Karlova, Fakulta matematicko-fyzikální, Praha, 1995.

[6.] LIMA, E. A. P., GUIMARAES, E. C., POZZA, S. A., BARROZO, M. A. S., COURY, J. R.: A study of atmospheric particulate matter in a city of the central region of Brazil using time-series analysis. *International Journal of Environmental Engineering*, Vol. 1, Nr. 1, June 2009.

[7.] LJUNG, G. M., BOX, G. E. P.: On a Measure of a Lack of Fit in Time Series Models. *Biometrika*, Vol. 65, pp. 297–303, 1978.

[8.] METELKA, J.: Náhodný trend v klimatologických řadách. *Meteorologické zprávy*, 1999, roč. 52, s. 41-49.

[9.] MODARRES, R.: Streamflow drought time series forecasting. *Stochastic Environmental Research and Risk Assessment*, Vol. 21, Nr. 3, February 2007 , pp. 223-233(11).

[10.] POZZA, S. A., LIMA, E. P., COMIN, T. T., GIMENES, M. L., COURY, J. R.: Time series analysis of PM 2.5 and PM 10−2.5 mass concentration in the city of Sao Carlos, Brazil. *International Journal of Environment and Pollution*, Vol. 41, Nrs. 1-2, March 2010 , pp. 90-108(19).

[11.] SCHULZE, P., PRINZ, A.: Forecasting container transshipment in Germany. *Applied Economics*, Vol. 41, Nr. 22, October 2009 , pp. 2809-2815(7).

[12.] WOODWARD, W. A., GRAY, H. L.: Global Warming and the Problem of Testing for Trend in Time Series Data. *Journal of Climate*, 1993, Vol. 6, s. 953-962.

Current address

Ing. Karel Helman
University of Economics Prague,
W.Churchill Sq.4, 13067 Prague, CZ
e-mail: helmank@vse.cz

Journal of Applied Mathematics

volume IV (2011), number III

COMPOSITE INDICATORS AND WEIGHTING SCHEME: THE CASE OF EUROPE 2020 INDICATORS

HUDRLIKOVA Lenka, FISCHER Jakub, CZ

Abstract. The paper is focused on issue of aggregating Europe 2020 indicators to the composite indicator. Generally speaking, we can compare countries by multi-dimensional methods and/or by a composite indicator (the second alternative has been chosen for the paper). When constructing the composite indicator, we need to select indicators, select a method of aggregation and assign weights for partial indicators. In this paper some methods for aggregation and weighting and the different composite indicators are compared.

Key words. Composite indicator, Europe 2020, weighting scheme, international statistical comparison

Mathematics Subject Classification: Primary 62P20; Secondary 91B82.

1 Introduction

For the international statistical comparison, we can generally use two types of methods. Firstly, the wide spectrum of multidimensional methods is at a disposal. As another approach we can construct the composite indicator. Constructing of composite indicator is a very difficult process with several steps (data selection, imputation of missing data, multivariate analysis, normalisation, weighting and aggregation, uncertainty and sensitivity analysis, visualisation)[1]. The aim of this paper is to compare some weighting and aggregation methods at the constructing of composite indicator based on Europe 2020 indicators[2] and to compare the countries using different indicators.
We compare countries using Equal Weighing Method, Principal Component Analysis and Factor Analysis, and Benefit of Doubt Approach. Methods which need subjective input are not included to our analysis.

[1] See also OECD (2008), pp. 20 – 21.
[2] See Eurostat (2010), http://epp.eurostat.ec.europa.eu/portal/page/portal/europe_2020_indicators/headline_indicators

2 Data and Methodology

We use 8 main indicators from *Europe 2020: A strategy for smart, sustainable and inclusive growth*.[3] These indicators contain:

Employment rate by gender, age group 20-64	(EMP)
Gross domestic expenditure on R&D	(GERD)
Greenhouse gas emissions, base year 1990	(GH)
Share of renewables in gross final energy consumption	(RE)
Energy intensity of the economy	(EN)
Early leavers from education and training by gender	(EL)
Tertiary educational attainment by gender, age group 30-34	(TE)
Population at risk of poverty or exclusion	(POV)

Weighing scheme could have a crucial influence for the composite indicator. We can divide methods of aggregation and weighing to the two main types:
(i) methods based on statistical methods and
(ii) methods based on opinions of researchers.

The second type of methods we do not consider at this paper, only the methods based on statistical methods are compared.

We compare these three methods:

2.1 Equal weighting (EW)

Using this method, the equal weight is assigned for each indicator:

$$w_q = \frac{1}{Q} \tag{1}$$

where w_q is weight for q^{th} sub-indicator ($q = 1,...,Q$) and for country c ($c = 1,...,M$).

It means all sub-indicators are given the same weight for all countries. Linear aggregation is used in the summation to composite indicator:

$$CI_c = \sum_{q=1}^{Q} I_{qc} w_q$$

2.2 Principal Component Analysis (PCA) a Factor Analysis (FA)

Principal components analysis, and more specifically factor analysis, groups together individual indicators which are collinear to form a composite indicator that captures as much as possible of the information common to individual indicators. The individual indicators must have the same unit

[3] See also http://ec.europa.eu/europe2020/index_en.htm.

of measurement. Each factor (usually estimated using principal components analysis) reveals the set of indicators with which it has the strongest association. The idea under PCA/FA is to account for the highest possible variation in the indicator set using the smallest possible number of factors. Therefore, the composite no longer depends upon the dimensionality of the data set but rather is based on the "statistical" dimensions of the data.[4]

Principal component analysis and factor analysis need a significant correlation between partial indicators, because the weights are set in accordance to correlation between indicators. We can see a correlation matrix (see table 1)[5].

Table 1. Correlation Matrix of Europe 2020 Indicators

	EMP	GERD	GH	RE	EN	EL	TE	POV
EMP	1.000	-0.073	-0.095	-0.457	0.241	-0.037	-0.142	-0.110
GERD	-0.073	1.000	0.018	0.334	-0.115	0.207	-0.058	0.350
GH	-0.095	0.018	1.000	-0.071	0.215	0.240	-0.115	-0.089
RE	-0.457	0.334	-0.071	1.000	-0.373	-0.012	0.184	0.221
EN	0.241	-0.115	0.215	-0.373	1.000	0.190	-0.495	-0.252
EL	-0.037	0.207	0.240	-0.012	0.190	1.000	-0.273	-0.044
TE	-0.142	-0.058	-0.115	0.184	-0.495	-0.273	1.000	0.196
POV	-0.110	0.350	-0.089	0.221	-0.252	-0.044	0.196	1.000

Source: Computation of authors

The correlations between indicators are relatively small. Despite this fact, weights for 4 factors are experimentally computed in the next part of the paper.

2.3 Benefit of Doubt approach (BOD)

BOD is a method a based on adjusted Data envelope analysis (DEA) which is used mainly at production issues. Using BOD, the composite indicator is defined as the ratio of a country's actual performance to its benchmark performance.

Data should be standardised using „min-max" method. Each value x_{qn}^t of indicator q for country c and time t (in this paper for year 2008) is transformed using the formula

$$I_{qc}^t = \frac{x_{qc}^t - \min_c\left(x_q^t\right)}{\max_c\left(x_q^t\right) - \min_c\left(x_q^t\right)} \tag{2}$$

After standardisation, all the values lies between 0 (laggard, $x_{qc}^t = \min_c\left(x_q^t\right)$) and 1 (leader, $x_{qc}^t = \max_c\left(x_q^t\right)$).

[4] Quoted directly from OECD (2008), p. 89.
[5] See also Hudrlikova (2010).

There is not one weighing scheme for all countries. For each country, they are used weights which are optimal for this country. It guarantees the best position for the associated country vis-à-vis all other countries in the sample. With any other weighting profile, the relative position of that country would be worse. Optimal weights are obtained by solving the following constrained optimisation[6]:

$$CI_c^* = \max \sum_{q=1}^{Q} I_{qc} w_{qc} \tag{3}$$

s. t.

$$\sum_{q=1}^{Q} I_{qk} w_{qk} \leq 1, \qquad w_{qk} \geq 0, \ \forall k = 1...M, \ \forall q = 1...Q$$

where k means countries and q means sub-indicators.

3 Results

3.1 Principal component analysis and its comparison with equal weights

Setup of weights using PCA is based on eigenvalues and then from the optimal numbers of components[7]. In tables 2 and 3 one can see factor loadings for 4 principal components and squared factor loadings.

Table 2 Factor loadings based on principal components

	Factor 1	Factor 2	Factor 3	Factor 4
EMP	-0.510	-0.306	-0.607	-0.209
GERD	0.363	0.624	-0.48	0.006
GH	-0.266	0.463	0.464	-0.623
RE	0.705	0.332	0.184	0.331
EN	-0.765	0.221	-0.053	0.083
EL	-0.247	0.698	0.02	-0.034
TE	0.604	-0.431	0.153	-0.432
POV	0.542	0.199	-0.495	-0.354
Eigenvalues	2.265	1.57	1.12	0.861

Source: Computation of authors

Table 3 Squared factor loading (scaled to unity sum); re-scaled weight

	Factor 1	Factor 2	Factor 3	Factor 4	WEIGHT
EMP	0.115	0.060	0.329	0.051	13%
GERD	0.058	0.248	0.206	0.000	13%
GH	0.031	0.137	0.192	0.451	15%
RE	0.219	0.070	0.030	0.127	13%
EN	0.258	0.031	0.003	0.008	11%
EL	0.027	0.310	0.000	0.001	9%
TE	0.161	0.118	0.021	0.217	13%
POV	0.130	0.025	0.219	0.146	12%
Expl.var	2.265	1.57	1.12	0.861	

[6] See also OECD (2008), p. 83.
[7] See also Hudrlikova (2010).

Note: Weights are normalized by squared factor loading, which is the portion of the variance of the factor explained by the variable. Source: Computation of authors

This is a case of setup of weights using principle component analysis. Using maximum likelihood method the weights could be different. We also can use factor analysis: the weights are influenced both by the extraction method and the rotation method as well.

As we said, neither PCA nor FA is feasible for weighing at a case of Europe 2020 indicators due to small correlations between sub-indicators. However, weights by Equal Weight Method and PCA are compared in table 4.

Table 4 Weights for the Europe 2020 indicators based on different methods

geo	EMP	GERD	GH	RE	EN	EL	TE	POV
EW	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
PCA	0.132	0.129	0.153	0.129	0.111	0.095	0.131	0.121

Source: Computation of authors

3.1 Weights by Benefit of Doubt Approach

Table 5 Weights by BOD approach applied to Europe 2020 indicators, Composite Indicators

geo	EMP	GERD	GH	RE	EN	EL	TE	POV	**CI**
Denmark	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	**1.00**
Cyprus	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	**1.00**
Latvia	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	**1.00**
Netherlands	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	**1.00**
Poland	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	**1.00**
Sweden	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**1.00**
Slovenia	0.000	0.000	0.000	0.000	0.000	0.998	0.000	0.006	**0.99**
Italy	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	**0.95**
Lithuania	0.000	0.000	0.974	0.000	0.000	0.000	0.072	0.000	**0.92**
Greece	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	**0.92**
Finland	0.000	0.958	0.000	0.000	0.000	0.000	0.050	0.000	**0.91**
Estonia	0.000	0.000	0.971	0.000	0.000	0.000	0.000	0.084	**0.91**
Spain	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	**0.91**
Czech Republic	0.000	0.000	0.000	0.000	0.000	0.057	0.000	0.953	**0.90**
Portugal	0.000	0.000	0.000	0.096	0.959	0.000	0.000	0.000	**0.84**
United Kingdom	0.000	0.000	0.150	0.000	0.898	0.000	0.000	0.000	**0.81**
Romania	0.000	0.000	0.920	0.000	0.000	0.116	0.000	0.000	**0.77**
Bulgaria	0.000	0.000	0.920	0.000	0.000	0.116	0.000	0.000	**0.75**
Slovakia	0.000	0.000	0.228	0.000	0.000	0.830	0.000	0.000	**0.71**
Austria	0.000	0.000	0.000	0.084	0.730	0.290	0.000	0.000	**0.59**
Ireland	0.000	0.000	0.000	0.000	0.643	0.441	0.000	0.000	**0.57**
Germany	0.000	0.045	0.584	0.000	0.548	0.000	0.000	0.000	**0.55**
Luxembourg	0.000	0.000	0.030	0.000	0.384	0.000	0.000	0.628	**0.53**
EU - 27	0.000	0.000	0.456	0.042	0.620	0.066	0.000	0.000	**0.51**
Malta	0.000	0.000	0.000	0.000	0.400	0.000	0.000	0.633	**0.48**
France	0.000	0.000	0.247	0.000	0.376	0.565	0.000	0.000	**0.43**
Belgium	0.000	0.000	0.247	0.000	0.376	0.565	0.000	0.000	**0.42**
Hungary	0.000	0.000	0.247	0.000	0.376	0.565	0.000	0.000	**0.40**

Source: Computation of authors

Weights derived from equal weights method and PCA on one hand are not comparable with weights derived from benefit of doubt approach on the other hand. At BOD, weights are assigned individually for each country. In table 5, there are weights for each country and in the last column there is a component index value for each country. The value of CI lies between 0 and 1, when the value 1 is considered as a benchmark.
The main disadvantage of this method is that without setup of borders (min-max values) the weight is given by an indicator in which the country is the best. For some countries the value of composite indicator is equal to 1. The results are influenced by the fact that countries with only one best value are considered as successful.
On the other hand, setup of borders needs to take subjective opinion into account. For example, it is possible to state that the weight of each indicator has to be between 5 % and 30 %. This approach will lead to significant change of weights.
The results obtained by BOD approach show in which indicator is the individual country „strong". E. g., the Czech Republic has a good value at indicator POV (poverty and social exclusion).

3.3 Comparison of countries

In table 6 (see below), one can see a comparison of individual countries depending of weighing scheme. While the ranking using EW and PCA is similar (see also table 4), the ranking based on BOD approach is quite different. Note that Sweden ranks first with all used methods. Slovakia has quite similar ranking when EW (20th), PCA (21st) and BOD (19th) are used. One the other hand, Finland ranks second with EW and also with PCA but only 11th according to BOD. Clearly with BOD method there is a problem with countries that has one leading sub-indicator. This is an example of Cyprus, Latvia and Poland. So these countries have CI equals to 1 and took the first place according to BOD.

4 Discussion

Advantages and disadvantages of the methods have been indicated. As it can be seen from table 6, the ranking is strongly influenced by the methods used for weighting. We considered just the methods which do not need subjective opinion. From the second type of methods, we can use for example budget allocation process, public opinion approach, analysis hierarchy process or conjoint analysis.

5 Conclusion

Methods for setup of weights for construction of composite indicator have been compared in empirical case of Europe 2020 Indicators (which are used for monitoring of Europe 2020 Strategy). Different methods can be used and on this empirical case we can see that a comparison of countries using a composite indicator strongly depends on setup of weights. There is no general consensus which weighting scheme is the best. In the case of indicators Europe 2020, equal weighting can not face the problem that there are eight indicators but they represent only 5 targets of EU policy. PCA seems as inadequate because of the assumption of correlation between sub-indicators. BOD has two cons. One of them was mentioned above. It is value 1 for CI of countries with leading sub-indicator.

It implies a problem of compensation of individual sub-indicators. The solution could be in setting boundaries - minimal and maximal weight for one sub-indicator in CI. However, setting boundaries change the character of the method from an objective to a subjective one.
Subjective methods for comparison of countries could be also taken into account, their using in thi case will be a subject of further research. To choose the appropriate weighting and aggregation method there is needed the next step in constructing CI – uncertainty and sensitivity analysis.

Table 6 EU country rankings based on different weighting methods

	EW	PCA	BOD
Sweden	1	1	1
Finland	2	2	11
Denmark	3	3	1
Austria	4	5	20
Netherlands	5	4	1
France	6	7	26
Germany	7	6	22
Slovenia	8	10	7
Estonia	9	8	12
Luxembourg	10	12	23
Belgium	11	11	27
Ireland	12	14	21
United Kingdom	13	9	16
Lithuania	14	13	9
EU (27 countries)	15	16	24
Latvia	16	15	1
Czech Republic	17	17	14
Cyprus	18	18	1
Spain	19	19	13
Slovakia	20	21	19
Portugal	21	20	15
Poland	22	22	1
Greece	23	23	10
Italy	24	24	8
Hungary	25	25	28
Bulgaria	26	26	18
Romania	27	27	17
Malta	28	28	15

Source: Computation of authors

Acknowledgement

The paper was supported by grant from Internal Grant Agency of University of Economics, Prague, project No. IGS F4/19/2011 “One-factor and multi-factor productivity analysis in context of input-output analysis and international comparison”.

References

[1] CHERCHYE L., MOESEN W., ROGGE N., Van PUYENBROECK T.: Constructing a Knowledge Economy Composite Indicator with Imprecise Data. 2009. [online]

[2] Eurostat (2010): Europe 2020 Indicators: http://epp.eurostat.ec.europa.eu/portal/page/portal/europe_2020_indicators/headline_indicators

[3] FISCHER Jan, FISCHER Jakub: EU – ostrý start statistiky nové ekonomie. Statistika, 2002, roč. 39, č. 11–12, s. 435–445. ISSN 0322-788X

[4] HEBÁK P. a kol.: Vícerozměrné statistické metody (3). Praha: Informatorium, 2007. 271 s. ISBN 80-7333-039-3.

[5] http://www.econ.kuleuven.be/eng/ew/discussionpapers/Dps09/Dps0915.pdf

[6] HUDRLÍKOVÁ L.: Indikátory Evropa 2020 a možnosti redukce proměnných. Forum Statisticum Slovacum, 2010, roč. VI, č. 5, s. 69–73. ISSN 1336-7420.

[7] NARDO M., SAISANA M., SALTELLI A., TARANTOLA S.: Tools for Composite Indicators Building. 2005. [online]

[8] http://composite-indicators.jrc.ec.europa.eu/Document/EUR%2021682%20EN_Tools_for_Composite_Indicator_Building.pdf

[9] OECD: Handbook on Constructing Composite Indicators. Metodology and User Guide. Paris: Organisation for Economic Co-operation and Development, 2008. 158 s. ISBN 978-92-64-04345-9.

[10] SAISANA M., TARANTOLA S.: State-of-the-Art report on Current Methodologies and Practices for composite Indicator Development. 2002. [online] http://composite-indicators.jrc.ec.europa.eu/Document/state-of-the-art_EUR20408.pdf.

Current address

Lenka Hudrlikova, Ing.
University of Economics, Prague,
nam. W. Churchilla 4, 130 67 Praha 3, Czech Rep.,
e-mail: xhudl05@vse.cz

Jakub Fischer, doc. Ing. Ph.D.
University of Economics, Prague,
nam. W. Churchilla 4, 130 67 Praha 3, Czech Rep.,
e-mail:fischerj@vse.cz

Journal of Applied Mathematics

volume IV (2011), number III

DETECTION OF CHANGE POINT IN STATISTICAL PROCESS CONTROL

JAROŠOVÁ Eva, (CZ)

Abstract. The paper deals with the Bayesian approach to the statistical control. The shift of the process mean is detected via a high value of the posterior probability. The average run length and the risk of false alarm are computed numerically by simulation for various levels of the shift and for different sample sizes. Both known and unknown process standard deviation are considered. The results of simulation show that the method performs better than the Shewhart control chart and confirm its usability in short run processes with the exception of individual values from the process with an unknown standard deviation.

Key words. Bayesian approach, average run length, risk of false alarm, Shiryayev-Roberts statistic.

Mathematics Subject Classification: 62P10, 62-07

1 Introduction

Statistical control of industrial processes is one of the most frequently used tools in quality control. A process is monitored through samples of relatively small size drawn at regular intervals. Sample characteristics are plotted against their order and compared to limits in the control chart. When a point falls beyond the control limits a signal is given that parameters of the process may have changed and so the process is out of control. Shewhart control charts for averages are very often applied. The control limits are positioned at $\pm 3\sigma / \sqrt{n}$ away from the central line which corresponds e.g. to the target process mean, standard deviation σ represents variation of the process that is under control and n denotes the size of subgroups.

When standard deviation σ is not known, 20 or 25 subgroups should be taken before the control limits are constructed. It cannot be accomplished in short run processes that are typical for modern business strategies. Various approaches were suggested to solve this problem. They include self-starting CUSUM chart [3], Q-charts [4] and others. Some methods are based on the Bayesian approach [2] and two of them were examined in [1]. To assess performance of different methods some characteristics are evaluated. The average run length (ARL) is the average number of subgroups taken until a point indicates an out-of-control condition. ARL is determined for several levels of shift δ in the process mean including $\delta = 0$. It is obvious that for $\delta = 0$ a fairly long ARL

is desirable while for a non-zero shift ARL should be as short as possible. The risk of false alarm (RFA) is the probability that a signal occurs without the process mean being shifted. RFA must be fairly small to avoid overcontrol. These characteristics must usually be determined numerically by simulation.
The aim of this paper is to examine one of the Bayesian methods in more detail. Some findings from [1] are used and other simulations are performed to explore the effect of the sample size and of the unknown process σ.

2 Detection of the change point

Suppose that a process is monitored at regular intervals and that means are determined in samples of size *n*. The sample means are assumed to have a normal distribution with mean μ_0 and variance σ^2 / n when the process is under control. Suppose the process mean changed from μ_0 to μ_1 at some time t_0 and remains at this level since then. Time t_0 of the change in the process mean is called the change point. Kenett and Zacks [2] present the following approach.
The probability that the change point occurred before or at sampling time *t* is determined repeatedly and its large value indicates the existence of a change point. A random discrete parameter τ is defined, where

$\tau = 0$ when the change point occurred before the first sampling time,

$\tau = i$ $(1 \le i < t)$ when the change point occurred between the i-th and (i+1)st sampling time,

$\tau = t$ when the change point occurred after time *t*.

The modified geometric prior distribution of this parameter at time *t* is used, defined by

$$\pi_t(\tau) = \begin{cases} \pi & \tau = 0, \\ (1-\pi)p(1-p)^{i-1} & \tau = i,\ 1 \le i < t, \\ (1-\pi)(1-p)^{t-1} & \tau = t. \end{cases} \tag{1}$$

Here π denotes the probability that the shift in the process occurred before the first sampling time and p is the probability of success on each trial (i.e. the probability that the shift occurs within the time interval between two successive samplings). Contrary to the ordinary geometric distribution, the set of values of τ is finite. We will assume that no shift occurred before the process started to be monitored. Then $\pi = 0$ and formulas (1) become simpler

$$\pi_t(\tau) = \begin{cases} 0 & \tau = 0, \\ p(1-p)^{i-1} & \tau = i,\ 1 \le i < t, \\ (1-p)^{t-1} & \tau = t. \end{cases} \tag{2}$$

The posterior probability function of τ at sampling time *t* given sample means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_t$ is

$$\pi_t(\tau \mid \bar{X}_1, \bar{X}_2, \ \dots\ , \bar{X}_t) = \frac{\pi_t(\tau) L_t(\tau; \bar{X}_1, \bar{X}_2, \ \dots\ , \bar{X}_t)}{\sum_{\tau} \pi_t(\tau) L_t(\tau; \bar{X}_1, \bar{X}_2, \ \dots\ , \bar{X}_t)}\ , \tag{3}$$

where the likelihood $L_t(\tau; \bar{X}_1, ..., \bar{X}_t)$ is given by

$$L_t(\tau; \bar{X}_1, ..., \bar{X}_t) = \begin{cases} \prod_{j=1}^{t} f(\bar{X}_j; \mu_1) & \tau = 0 \\ \prod_{j=1}^{\tau} f(\bar{X}_j; \mu_0) \ \prod_{j=\tau+1}^{t} f(\bar{X}_j; \mu_1) & 1 \le \tau < t \\ \prod_{j=1}^{t} f(\bar{X}_j; \mu_0) & \tau = t \end{cases} \tag{4}$$

Functions $f(\bar{X}_j; \mu_0)$ and $f(\bar{X}_j; \mu_1)$ are densities of normal distributions $N(\mu_0, \sigma^2/n)$ and $N(\mu_1, \sigma^2/n)$, respectively. At the sampling time t we are interested in the posterior probability $P(\tau < t \mid \bar{X}_1, ..., \bar{X}_t)$ that the change point has occurred. Using equations (2), (3) and (4), we have

$$P(\tau < t \mid \bar{X}_1, ..., \bar{X}_t) = \frac{\sum_{i=1}^{t-1} p(1-p)^{i-1} \prod_{j=1}^{i} f(\bar{X}_j; \mu_0) \ \prod_{j=i+1}^{t} f(\bar{X}_j; \mu_1)}{\sum_{i=1}^{t-1} p(1-p)^{i-1} \prod_{j=1}^{i} f(\bar{X}_j; \mu_0) \ \prod_{j=i+1}^{t} f(\bar{X}_j; \mu_1) \ + (1-p)^{t-1} \prod_{j=1}^{t} f(\bar{X}_j; \mu_0)}. \tag{5}$$

It can be rewritten as

$$P(\tau < t \mid \bar{X}_1, ..., \bar{X}_t) = \frac{\frac{p}{(1-p)^{t-1}} \sum_{i=1}^{t-1} (1-p)^{i-1} \prod_{j=i+1}^{t} R_j}{\frac{p}{(1-p)^{t-1}} \sum_{i=1}^{t-1} (1-p)^{i-1} \prod_{j=i+1}^{t} R_j + 1}, \tag{6}$$

where

$$R_j = \frac{f(\bar{X}_j; \mu_1)}{f(\bar{X}_j; \mu_0)} = \exp\left(-\frac{n\delta^2}{2\sigma^2} + \frac{n\delta}{\sigma^2} (\bar{X}_j - \mu_0) \right). \tag{7}$$

Kenett and Zacks [ ] use an approximate expression

$$P(\tau < t \mid \bar{X}_1, ..., \bar{X}_t) \cong \frac{\sum_{i=1}^{t-1} \prod_{j=i+1}^{t} R_j}{\sum_{i=1}^{t-1} \prod_{j=i+1}^{t} R_j + 1}, \tag{8}$$

where $\sum_{i=1}^{t-1} \prod_{j=i+1}^{t} R_j = W_t$ is Shiryayev-Roberts statistic.

In paper [1] the original expression (6) was considered. Putting

$$\frac{p}{(1-p)^{t-1}}\sum_{i=1}^{t-1}(1-p)^{i-1}\prod_{j=i+1}^{t}R_j = pZ_t, \tag{9}$$

Z_t can be determined recursively

$$Z_t = \frac{R_t}{1-p}(Z_{t-1}+1) \tag{10}$$

and the probability $P(\tau < t \mid \bar{X}_1,...,\bar{X}_t)$ is given by

$$P(\tau < t \mid \bar{X}_1,...,\bar{X}_t) = \frac{pZ_t}{pZ_t+1} \tag{11}$$

If $P(\tau < t \mid \bar{X}_1,...,\bar{X}_t)$ is larger than some stopping threshold π^* a signal is given that a change point has occurred that is that the process mean has shifted.
When σ of the process must be estimated, a recursive formula for sample size $n \geq 2$

$$w_t^2 = \frac{1}{t}\sum_{l=1}^{t}s_l^2 = \frac{(t-1)w_{t-1}^2 + s_t^2}{t} \tag{12}$$

can be used, where s_l^2 is the sample variance at the l[th] sampling time.

3 Simulation study

The prior distribution (2) with p = 0.05 was used based on the simulation study in [1]. To compute R_j according to (7), the deviation from μ_0 which is to be identified, i.e. $\delta = \mu_1 - \mu_0$ has to be set. The size of shift δ corresponded gradually to σ, 1.5σ, and 2σ, where σ is the standard deviation of the pocess. Based on [1], the stopping threshold equal to 0.99865 was chosen. This value imitating the one-sided risk of false alarm in the Shewhart control chart seemed to guarantee a sufficiently low risk of false signal.
The aim of the simulation study was to evaluate ARL for different sample sizes n and for both known and unknown σ. Monte Carlo method was used to simulate drawing subgroups from a process within SPC. Three situations were considered:

a) a process under control with the mean equal to the target value; in this case 1000 samples from N(10, 9) were generated in one cycle,
b) process with a shift of the mean equal to δ that occurred between sampling times $t = 5$ and $t = 6$; first 5 samples came from N(10, 9), the remaining 95 samples from N$(10+\delta, 9)$,
c) process with a shift of the mean equal to δ that occurred between sampling times $t = 10$ and $t = 11$; first 10 samples came from N(10, 9), the remaining 90 samples from N$(10+\delta, 9)$.

The sample size changed from 2 to 5 and in case of known σ also individual values were considered. For all conditions 1 000 cycles were performed every time and the number of samples until $P(\tau < t \mid \bar{X}_1,...,\bar{X}_t) > \pi^*$ were recorded. Results are given in tables 1 to 4.

Table 1. Empirical ARL0 based on 1000 subgroups.

n	known σ			unknown σ		
	3	4.5	6	3	4.5	6
1	956	976	982	-	-	-
2	982	988	991	923	932	952
3	988	991	996	963	977	984
4	987	992	994	979	986	989
5	988	994	997	978	990	996

Table 2. Empirical RFA based on 1000 subgroups.

n	known σ			unknown σ		
	3	4.5	6	3	4.5	6
1	83	42	33	-	-	-
2	40	21	14	100	79	54
3	26	16	6	49	34	20
4	27	14	9	36	19	14
5	22	9	6	31	13	8

Table 3. Empirical ARL , change point between 5^{th} and 6^{th} sampling time

n	known σ			unknown σ		
	3	4.5	6	3	4.5	6
1	14.052	6.964	4.067	-	-	-
2	7.661	3.601	1.936	7.305	3.430	1.926
3	5.275	2.416	1.219	5.163	2.287	1.202
4	3.953	1.736	0.792	4.002	1.686	0.795
5	3.258	1.260	0.492	3.146	1.285	0.522

Table 4. Empirical ARL , change point between 10^{th} and 11^{th} sampling time

n	known σ			unknown σ		
	3	4.5	6	3	4.5	6
1	13.687	6.741	3.974	-	-	-
2	7.527	3.578	2.000	7.439	3.522	2.007
3	5.295	2.351	1.193	5.258	2.395	1.213
4	4.052	1.740	0.787	4.075	1.701	0.823
5	3.200	1.305	0.478	3.202	1.308	0.521

Table 5. ARL in Shewhart control chart, standards given

n	σ	1.5σ	2σ
1	44	15	6
2	18	5	2
3	10	3	1
4	6	2	1
5	4	2	1

4 Conclusion

The simulation study confirmed good properties of the method. All values of ARL are smaller than those of the classical Shewhart control chart (Table 5). The fact that estimating σ practically does not affect ARL is important. Based on two simulated alternatives with the change point located between the 5^{th} and the 6^{th} sampling times or between the 10^{th} and the 11^{th} sampling times, it seems that the Bayesian method performs well even for quite short sequences of samples.
As for individual observations, ARL of the Bayesian method is much better than ARL of the Shewhart control chart when σ of the process is known. The problem arises, though, when σ is to be estimated. The estimation based on moving ranges used in the control charts for individuals is not applicable in the recurrent formula because the change of the process mean at the change point is expected to induce a large value of the corresponding moving range and thus to bias the estimate of σ. A possible excluding this "unsuitable" moving range seems to be quite intricate.

References

[1.] JAROŠOVÁ, E.: Bayesian approach to the short run process control. Demanovská Dolina, 25.-29.8.2010. AMSE 2010 Applications of Mathematics and Statistics in Economy, to be published

[2.] KENETT, R.S., ZACKS, S.: *Modern Industrial Statistics*. Brooks/Cole Publishing Company, Pacific Grove, 1998.

[3.] MONTGOMERY, D.C.: *Statistical Quality Control: A Modern Introduction*. John Wiley & Sons, Hoboken, 2009.

[4.] QUESENBERRY, Ch.: On Properties of Q Charts for Variables. Journal of Quality Technology 27, pp. 204-213, 1995.

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

Journal of Applied Mathematics

volume IV (2011), number III

DISTRIBUTION OF INCOMES PER CAPITA OF THE CZECH HOUSEHOLDS FROM 2005 TO 2008

MALÁ Ivana, (CZ)

Abstract. In the article two and three parameter lognormal and three parameter Dagum distributions are fitted into data of net annual income per head (in CZK) for the Czech households. Data from the survey Statistics of Income and Living Conditions (EU-SILC) organized by the Czech Statistical Office in the years 2005 - 2008 are used. The maximum likelihood estimates are evaluated and the fits are compared with the use of AIC information criterion. Furthermore estimated and sample characteristics are evaluated in order to compare results of different models. Estimated probability densities of income per capita for three parameter lognormal and Dagum distributions are presented in the figure, where the similar progress of distributions during treated four years is obvious. Dagum distribution shows better fit to data than lognormal distributions in all analysed years.

Key words. income distribution, lognormal distribution, Dagum distribution, maximum likelihood estimation

Mathematics Subject Classification: Primary 62-07; 62E17

1 Introduction

Knowledge of characteristics of income is very important not only for the government, experts and professionals in economy but also for the large public. The development of incomes and their characteristics in time is also important as it can well describe the progress in the economy and quantify level or variance of income or differences between subpopulations. More detailed information can be obtained from the modelling of the whole distribution of incomes. It provides not only characteristics but also probability density, distribution or quantile functions.

From the point of view of statistics, different probability distributions or its mixtures can be fit into data. Such distributions are sometimes called income distributions. For the successful model of wages or incomes a flexible, skewed distribution with high variability is necessary. The most frequently used distributions are lognormal with two, three and four parameters [1]-[4],[9], Dagum distribution, generalized lambda distribution [7] or mixtures of distributions mentioned above (for

the finite mixtures of lognormal distributions see [8]). For the high incomes extreme value distribution or Pareto distribution are usually used.

2. Methods

In the text two and three parameter lognormal distributions are used as a model for income distribution. If X is a random variable, then the two parametric lognormal distribution with the parameters μ and σ^2 (expected value and variance of the normal distribution of ln$X$) has the density given by a formula ($x > 0$)

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right). \tag{1}$$

If the third (shift) parameter θ is included into the model, the density is of the form

$$f(x;\mu,\sigma^2,\theta) = \frac{1}{\sqrt{2\pi}\sigma(x-\theta)}\exp\left(-\frac{(\ln(x-\theta)-\mu)^2}{2\sigma^2}\right),\ \ x>\theta. \tag{2}$$

There exist close formulas for maximum likelihood estimates of the unknown parameters in the case of the two parametric lognormal distribution

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln x_i\ ,\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\ln x_i-\hat{\mu}\right)^2. \tag{3}$$

In the case of the three parametric lognormal distribution a procedure from Cohen [5] is used in order to find maximum likelihood estimates of the unknown parameters.
The expected value $E(X)$ and percentiles x_P are computed with the use of formulas

$$E(X) = \theta+\exp(\mu+\sigma^2/2),\ \ x_P = \ \theta+\exp\left(\mu+\sigma u_P\right)\quad 0<P<1. \tag{4}$$

where u_P is a 100P % quantile of a standard normal distribution. The level of the random variable can be expressed (except for the expectation and the median) also by a mode. In the lognormal distribution the value x_{mode}, where the maximum of the density occurs, is given by the formula

$$x_{\text{mode}} = \theta+e^{\mu-\sigma^2}. \tag{5}$$

For the two parameter distribution $\theta = 0$ is used in (4) and (5).
The third distribution that is used in this text is three parameter Dagum distribution, called also Inverse Burr´s. This distribution is frequently used in actuarial applications, but it can be used also for the modelling of the distribution of wages and incomes with very good results. The density of this distribution is given by the formula

$$f(x;\alpha,\beta,p) = \frac{\alpha p y^{\alpha p-1}}{\left(\beta^{\alpha p}(1+(x/\beta)^{\alpha}\right)^{p+1}}, x>0, \tag{6}$$

where α,β and p are positive parameters. The distribution function of this distribution can be written in the form

$$F(x) = \left(1 + (x/\beta)^{-\alpha}\right)^{-p}. \tag{7}$$

Expected value (at least in the set of values of parameters we want to use in the modelling) is evaluated from the formula

$$E(X) = \left(b\,\Gamma\left(p\ +\ 1/\alpha\right)\,\Gamma\left(1\ -\ 1/\alpha\right)\right)\ /\ \Gamma(p). \tag{8}$$

After the straightforward calculation from (7) quantiles x_P are given by

$$x_P = F^{-1}(P) = \sqrt[\alpha]{\frac{\beta^{\alpha}}{\sqrt[p]{P^{-1}} - 1}}. \tag{9}$$

The maximum of density function occurs in

$$x_{\text{mode}} = b\left(\frac{\alpha p - 1}{\alpha + 1}\right)^{1/\alpha}. \tag{10}$$

Three unknown parameters α, b and p are estimated with the use of maximum likelihood method. The logarithmic likelihood function is maximized in the program R in order to obtain maximum likelihood estimates $\hat{\alpha}, \hat{\beta}$ and $\hat{p}$. It is known [6], that maximum likelihood estimates are sensitive to isolated observations (and there are isolated large incomes in our data) but the sample sizes are large enough to obtain reasonably good fits. In this text all estimates are made with the use of maximal likelihood and the models are compared by values of logarithmic likelihood in the final solution and Akaike criterion, that reflects number of parameters in the model distribution (two for three parameters in the lognormal distribution and three for Dagum distribution).

3 Data Analysis and Results

In the paper individual data of Czech households from the survey Results of the Living Conditions Survey (a national module of the European Union Statistics on Income and Living Conditions (EU-SILC)) for years 2005, 2006, 2007 and 2008 are analysed. These surveys have been organized regularly by the Czech Statistical Office since 2005. Data for various types of income of the Czech households are collected together with descriptive characteristics of a household as status of head of household, number of persons in job and number of dependent children, household type, age and education of head of household, size of municipality or region of living. For this text a net anual income per capita was evaluated as a ratio of a net total income of a household (in CZK) and a number of persons in a household. This variable is a ratio of two random variables, its value was evaluated for each household and this dataset was treated data as a random sample (not stratified as it in fact is).

Table 1: Maximum likelihood estimates of unknown parameters

Distribution	Lognormal (2 parameters)		Lognormal (3 parameters)			Dagum (3 parameters)		
Parameter	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
2005	11.503	0.454	11.503	0.454	-2.050	4.091	96,366.6	1.060
2006	11.542	0.446	11.542	0.446	-8.805	4.150	99,319.8	1.084
2007	11.623	0.436	11.623	0.435	-42.288	4.243	107,219.1	1.098
2008	11.702	0.422	11.703	0.421	-171.167	4.330	113,878.9	1.159

There are 4,351 observations for the year 2005, 7,483 for 2006, 9,675 for the year 2007 and 11,294 form the year 2008 in the dataset. From these households 3,314 are included in all years, there are in total 13,504 households in all data sets.
Estimated parameters of all fits for analysed years are given in the Table 1 and estimated densities are shown in the Figure 2. Parameters of lognormal distributions are almost equal and the test of the hypothesis that parameter θ in three parametric distribution is equal to zero is nonsignificant on sigificance level $\alpha = 0.01$ (values not included in the text). In order to compare sample and theoretical characteristic based on fits and values of parameters from the Table 1, sample and theoretical values of characteristics of the level of income are given in the Table 2 (only for three parameter distributions, characteristics for the two parameter lognormal distribution are very similar to those of three parameter distribution).

Table 2: Sample and estimated characteristics of the net annual income per capita (CZK)

Distribution	Year	10% quantile	25% quantile	Median	75% quantile	90% quantile	mean	mode
Sample	**2005**	58 120	79 600	97 050	124 068	171 833	111 024	
	2006	60 832	82 998	100 640	128 000	174 904	114 945	
	2007	68 147	90 000	108 744	138 000	189 505	123 806	
	2008	76 571	97 160	117 497	148 937	202 327	132 877	
lognormal	**2005**	55 379	72 938	99 047	134 502	177 145	109 781	80 624
	2006	58 156	76 233	102 976	139 100	182 331	113 734	84 416
	2007	63 875	83 209	111 621	149 730	195 034	122 719	92 344
	2008	70 374	90 910	120 810	160 524	207 304	132 014	101 170
Dagum	**2005**	58 385	75 622	98 284	128 138	167 379	108 678	86 923
	2006	61 424	78 982	102 028	132 356	172145	112 507	90 514
	2007	67 484	86 139	110 517	142 467	184 187	121 365	98 522
	2008	74 496	93 908	119 272	152 602	196 067	130 563	106 708

Table 3 contains values of logarithmic likelihood functions (referred as loglikelihood in the table) for all years and all distributions together with the value of AIC statistics defined by

$$AIC = 2 * \text{number of parameters} - 2 * \text{loglikelihood}. \tag{11}$$

Table 3: Comparison of fits

Distribution	$LN(\mu;\sigma^2)$		$LN(\mu;\sigma^2;\theta)$		Dagum	
Year	loglikelihood	*AIC*	loglikelihood	*AIC*	loglikelihood	*AIC*
2005	-52 785	105 573	-52 784	105 574	-52 608	105 221
2006	-90 942	181 887	-90 940	181 886	-90 612	181 230
2007	-118 135	236 273	-118 129	236 265	-117 689	235 383
2008	-138 429	276 862	-138 408	276 822	-137 849	275 703

All values of *AIC* are similar but the value is lesser for Dagum distribution in all analysed years.
In the Figure 1 the development of estimated percentiles 10%, 25%, 50%, 75% and 90 % obtained from fitted distributions are compared with sample quantiles The estimates of the characteristics of

the level of income are very close for both fits and well coincide with the sample values. We can see that lognormal distribution underestimates values of quantiles 10 and 25 percent and overestimate quantiles above median. The correspondence for Dagum distribution is similar, estimated and sample values are again closer than for lognormal distribution.

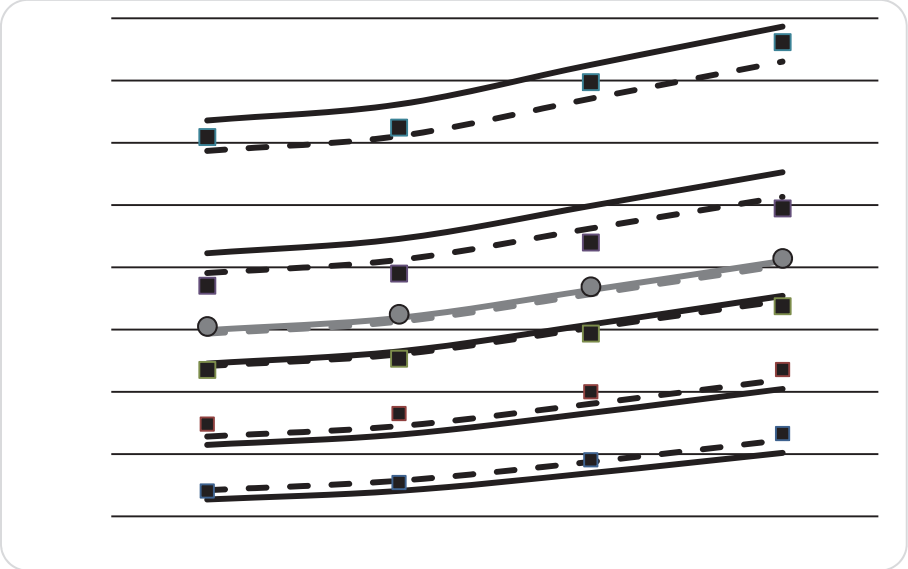


Figure 1: Sample quantiles of the net year income per capita in CZK (squares), estimated quantiles (Dagum dotted line, lognormal solid line). From the bottom 10%, 25%, 50%, 75% and 90%. Mean value (circle) and estimated expectation (lines) in grey colour.

Estimated densities from the three parametric lognormal and Dagum distribution are shown in the Figure 2. Solid lines are lognormal densities and dotted lines mean Dagum densities.

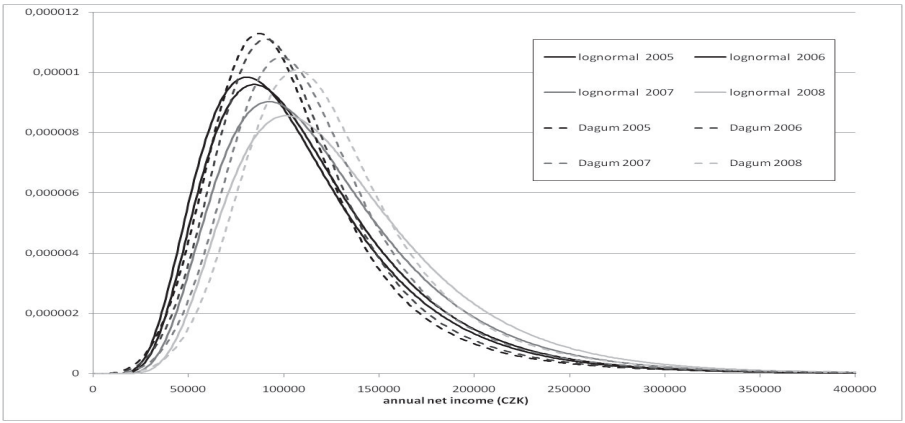


Figure 2: Estimated densities **(**Dagum distribution (dotted line) and three parameter lognormal distribution (solid line)). From the left to the right 2005 to 2008.

The development during years is obvious, curves moves from the left (year 2005) to the right (2008). Furthermore the densities are lower and lower in time that reflects increasing variability in incomes in years.

4 Conclusions

The Dagum distribution is suitable for the modelling of the net income per capita in the Czech Republic. According to the Akaike´s criterion the fit of Dagum distribution is superior to the lognormal distribution. The three parameter lognormal distribution provides comparable results to the distribution with two parameters. Estimates of shift parameter theta are negative for all years in the analysis, but large standard errors of estimates (if compared to the standard errors of estimates of other parameters) lead to the nonsignificant test of the parameter.

Acknowledgement

The paper was supported by the means of institutional support of a long-term conceptual advancement of science and research at the Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic.

References

[1] BARTOŠOVÁ, J. Logarithmic-Normal Model of Income Distribution in the Czech Republic. *Austrian Journal of Statistics*, Vol. 35, Iss. 23, s. 215 − 222. ISSN 1026-597x. 2006.

[2] BARTOŠOVÁ, J., BÍNA, V.: Modelling of Income Distribution of Czech Households in Years 1996 – 2005. *Acta Oeconomica Pragensia*, Vol. 17, Iss. 4, s. 3 − 18. ISSN 0572-3043. 2009

[3] BÍLKOVÁ, D. Application of Lognormal Curves in Modeling of Wage Distributions. *Journal of Applied Mathematics*, Vol. 1, Iss. 2, pp. 341 − 352. ISSN 1337-6365. 2008.

[4] BÍLKOVÁ, D. Modelování mzdových rozdělení v České republice v leeetech 2004 a 2005 s využitím logaritmicko-normálních křivek a křivek Pearsonova a Johnsonova systému. *Statistika*, roč. 88, č. 2, s. 149 − 166. ISSN 0322-788x. 2008.

[5] COHEN, A.C., JONES WHITTEN, B., Estimation in the three-parameter lognormal distribution, *Journal of American Statistical Association*, 399-404, Vol.75. 1980.

[6] KLEIBER, C., Kotz, S.: *Statistical Size Distributions in Economics and Actuarial Sciences,* Wiley-Interscience, New York. ISBN 0-471-15069-9. , 2003.

[7] PACÁKOVÁ, V., SIPKOVÁ, L. Generalized Lambda Distributions of Household\s Incomes. *E + M Ekonomie a Management*, roč. X, č. 1, s. 98 – 107. ISSN 1212-3609. 2007.

[8] PAVELKA, R.: Application of density mixture in the probability model construction of wage distributions. Applications of Mathematics and Statistics in Economy: AMSE 2009. Uherské hradiště, pp.341-350. ISBN 9788024516004. 2009,

Current address

Ivana Malá
University of Economics, Prague
nám. W.Churchilla 4, Prague 3
Czech Republic
tel. 420224095486,p
e-mail: malai@vse.cz

Journal of Applied Mathematics

volume IV (2011), number III

COMPARISON OF PENALIZED SPLINE REGRESSION WITH NONLINEAR REGRESSION

MARCINKO Tomáš, (CZ)

Abstract. Basically, semiparametric regression methods, which try to combine the advantages of both parametric and nonparametric approach to regression analysis, are often concerned with a flexible incorporation of nonlinear functional relationships in regression analysis. In particular, penalized spline regression uses the idea of nonparametric spline smoothing and it is in fact just a generalization of smoothing splines that should allow more flexibility in a choice of the spline model, the basis functions, and the penalty. The purpose of this article is to compare a penalized spline regression fit with an estimation based on a nonlinear regression model, which is calculated by a method of nonlinear least squares. Furthermore, some advantages and possible extensions of the penalized spline regression method are discussed.

Key words. regression analysis, semiparametric approach, spline smoothing, penalized splines, nonlinear regression

Mathematics Subject Classification: Primary 62G08; Secondary 62J02.

1 Introduction

Arguably, one of the most important questions in many fields of science is modeling of a general relationship between a response variable and one or more explanatory variables. Basically, this estimation of a conditional expectation of a response variable can be done in two ways. The widely used parametric approach assumes that the conditional mean function is of some specific functional form, e.g. in case of linear regression a line with unknown slope and intercept. The advantages of this parametric approach are a thorough theoretical framework and ease of interpretability.

An alternative approach tries to estimate the conditional mean function nonparametrically, i.e. without any prior assumptions about its functional form. These nonparametric methods are mostly only data-driven, which has an obvious advantage – these methods do not rely on a specification of the model describing a relationship between a response variable and explanatory variables.
Some techniques try to combine the ideas of both parametric and nonparametric regression methods and the penalized spline regression is definitely one of them. The biggest advantage of the penalized

spline regression is that, on the one hand, there is no a priori need to know a functional form of the underlying relationship, and on other hand, the estimation based on the penalized spline method is similar to a parametric estimation.

Since semiparametric regression methods are often concerned with a flexible incorporation of nonlinear relationships into a regression model, the main aim of this article will be a comparison between a nonlinear regression model estimated by the method of nonlinear least squares and a penalized spline regression fit. This comparison will be demonstrated on the dataset taken from Ratkowsky (1983). This dataset contains observed data, where the response variable is onion bulb dry weight, and the explanatory variable is growing time.

2 Penalized Spline Regression

2.1 Spline Model

The spline regression is essentially based on spline functions, i.e. piecewise polynomial functions. For example, the simple p-degree spline model can be defined as

$$m(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk} (x - \kappa_k)_+^p , \tag{2.1}$$

where $(x - \kappa_k)_+$ is the positive part of the function $x - \kappa_k$, sometimes referred to as a truncated line, and the value of κ_k corresponding to this function is usually called a knot. Basically, the number and the location of the knots can be arbitrary, but an automatic knot selection is reasonable, especially when using more elaborate penalized spline models. Therefore, Ruppert et al. (2003) proposed a default number of knots in the form

$$K = \min\left(\frac{1}{4} \times \text{number of unique } x_i , 35\right), \tag{2.2}$$

with knots being located at equally spaced quantiles, but in practice a lower number of knots is often sufficient.

However, a spline function defined without any restrictions can be too rough. One way to obtain a smooth curve is to constrain the influence of the knots. This can be achieved by adding a constraint on the coefficients β_{pk} of the form

$$\sum_{k=1}^{K} \beta_{pk}^2 < C . \tag{2.3}$$

Using the notation

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \ldots & x_1^p & (x_1 - \kappa_1)_+^p & \ldots & (x_1 - \kappa_k)_+^p \\ 1 & x_2 & \ldots & x_2^p & (x_2 - \kappa_1)_+^p & \ldots & (x_2 - \kappa_k)_+^p \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & \ldots & x_n^p & (x_n - \kappa_1)_+^p & \ldots & (x_n - \kappa_k)_+^p \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}$$

and $\mathbf{D} = \mathrm{diag}(\mathbf{0}_{p+1}, \mathbf{1}_K)$ it can derived (e.g. using the method of Lagrange multipliers) that the least squares fit for the penalized spline regression model is given by

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + h^2\mathbf{D}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{S}_h\mathbf{y}, \tag{2.4}$$

where h is a smoothing parameter and $\mathbf{S}_h$ is a smoother matrix.

An interesting fact about the penalized spline estimators described here is that they can be defined as a best linear unbiased predictor of a mixed model. Using the notation

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^{\,2} & \ldots & x_1^{\,p} \\ 1 & x_2 & x_2^{\,2} & \ldots & x_2^{\,p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^{\,2} & \ldots & x_n^{\,p} \end{pmatrix}, \ \mathbf{Z} = \begin{pmatrix} (x_1-\kappa_1)_+^p & \ldots & (x_1-\kappa_K)_+^p \\ (x_2-\kappa_1)_+^p & \ldots & (x_2-\kappa_K)_+^p \\ \ldots & \ldots & \ldots \\ (x_n-\kappa_1)_+^p & \ldots & (x_n-\kappa_K)_+^p \end{pmatrix}$$

the model (2.1) can be also formulated as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{2.5}$$

where the coefficients $\mathbf{u}$ are treated as random with the covariance matrix

$$\mathrm{cov}(\mathbf{u}) = \sigma_u^2\mathbf{I}, \text{ where } \sigma_u^2 = \frac{\sigma_\varepsilon^2}{h^2}. \tag{2.6}$$

The consequence of this mixed model representation of penalized splines is that it can be fit using mixed model software with restricted maximum likelihood method used to select the appropriate amount of smoothing as

$$\hat{h}_{REML} = \left(\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_u^2}\right)^{1/2p}. \tag{2.7}$$

2.2 Choice of a smoothing parameter

The smoothing parameter selection based on eq. (2.7) depends on the mixed model representation of penalized splines. However, many other nonparametric smoothing techniques do not have such a representation and therefore several other procedures, which can help a researcher with selection of an appropriate amount of smoothing, have been proposed.

Probably the most common approach is based on a cross-validation function. This function is defined as

$$CV(h) = \sum_{i=1}^{n}\left[y_i - \hat{m}_{(-i)}(x_i; h)\right]^2, \tag{2.8}$$

where $m_{(-i)}$ denotes the leave-one-out estimator, i.e. the regression estimator applied to data with (x_i, y_i) omitted. This leave-one-out estimator can be calculated using a smoother matrix $\mathbf{S}_h$ as

$$\hat{m}_{(-i)}(x_i;h)=\frac{\sum_{j\neq i} S_{h,ij}\,y_j}{\sum_{j\neq i} S_{h,ij}}, \tag{2.9}$$

The CV choice of the smoothing parameter h is the one that minimizes $CV(h)$ over $h \geq 0$.

One of the other possible methods that can guide the choice of the smoothing parameter is the penalizing function approach. This approach is based on minimizing the "penalized" weighted residual sum of squares using some penalizing function. Details of this approach are described, for example, in Härdle et al. (2004).

2.3 Construction of Confidence and Prediction Intervals

Suppose a penalized spline regression model

$$y_i = m(x_i)+\varepsilon_i, \tag{2.10}$$

with independent and identically distributed homoscedastic normal errors

$$\varepsilon_i \sim \mathrm{N}(0,\sigma_\varepsilon^2). \tag{2.11}$$

Considering that eq. (2.4) is in fact a linear smoother, the estimation of any $m(x)$ can be written as

$$m(x)=\mathbf{s}_x^{\mathrm{T}}\mathbf{y}\,. \tag{2.12}$$

It can be shown that under these circumstances it is possible to use the following approximation

$$\frac{\hat{m}(x)-E[\hat{m}(x)]}{\hat{\sigma}_\varepsilon\|\mathbf{s}_x\|}\sim \mathrm{t}_{[df_{\mathrm{res}}]}, \tag{2.13}$$

where $[df_{\mathrm{res}}]$ is the closest integer to the value of residual degrees of freedom, which can be formally defined as

$$df_{\mathrm{res}} = n-2\mathrm{tr}(\mathbf{S}_h)+\mathrm{tr}(\mathbf{S}_h\mathbf{S}_h^{\mathrm{T}}). \tag{2.14}$$

Therefore, a confidence interval can be written as

$$\hat{m}(x)\pm \mathrm{t}_{1-\alpha/2}\hat{\sigma}_\varepsilon\|\mathbf{s}_x\|. \tag{2.15}$$

Analogously to a parametric regression, a prediction interval for any observation at x is

$$\hat{m}(x)\pm \mathrm{t}_{1-\alpha/2}\hat{\sigma}_\varepsilon\sqrt{1+\|\mathbf{s}_x\|^2}\,. \tag{2.16}$$

Moreover, it is convenient to keep in mind that for large samples the normal approximation works reasonably well even if the errors are not necessarily Gaussian. Also, in case of heteroscedastic errors, the variance function estimation may be considered.

2.4 Additive model

The theoretical framework described so far can be easily extended so that it will allow multiple explanatory variables. Since the only assumption made will be the assumption of additivity of explanatory variables are the following models referred to as additive models.

For example, an additive linear spline model with two explanatory variables can be defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sum_{k=1}^{K_1} \beta_{1k} \left(x_{i1} - \kappa_{k1}\right)_+ + \sum_{k=1}^{K_2} \beta_{2k} \left(x_{i2} - \kappa_{k2}\right)_+ + \varepsilon_i . \tag{2.17}$$

The vector of fitted values for this additive penalized spline model is given by

$$\hat{\mathbf{y}} = \mathbf{C}\left(\mathbf{C}^{\mathrm{T}}\mathbf{C} + \boldsymbol{\Delta}\right)^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{y} , \tag{2.18}$$

where $\mathbf{C}$ is a design matrix and

$$\boldsymbol{\Delta} = \mathrm{diag}\left(0, 0, 0, h_1^2 \mathbf{1}_{K_1}, h_2^2 \mathbf{1}_{K_2}\right), \tag{2.19}$$

Another possibility is to use an additive partial linear model, which can be in case of two explanatory variables formulated as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sum_{k=1}^{K_2} \beta_{2k} \left(x_{i2} - \kappa_{k2}\right)_+ + \varepsilon_i . \tag{2.20}$$

The estimation of such a model is straightforward.

To conclude, it is important to keep in mind that these additive models, as already mentioned, rely on an assumption of additivity. If this assumption is not justified, other models dealing with possible interactions may be necessary, e.g. interaction models proposed in Ruppert et al. (2003). A test for additivity can be conducted by comparing the additive model to the interaction model and checking whether the interaction model offers a significant improvement in fit.

For further details on spline smoothing and penalized spline regression see Eubank (1999) and Ruppert et al. (2003) respectively.

3 Comparison of the Spline Model with the Nonlinear Model

3.1 Nonlinear model

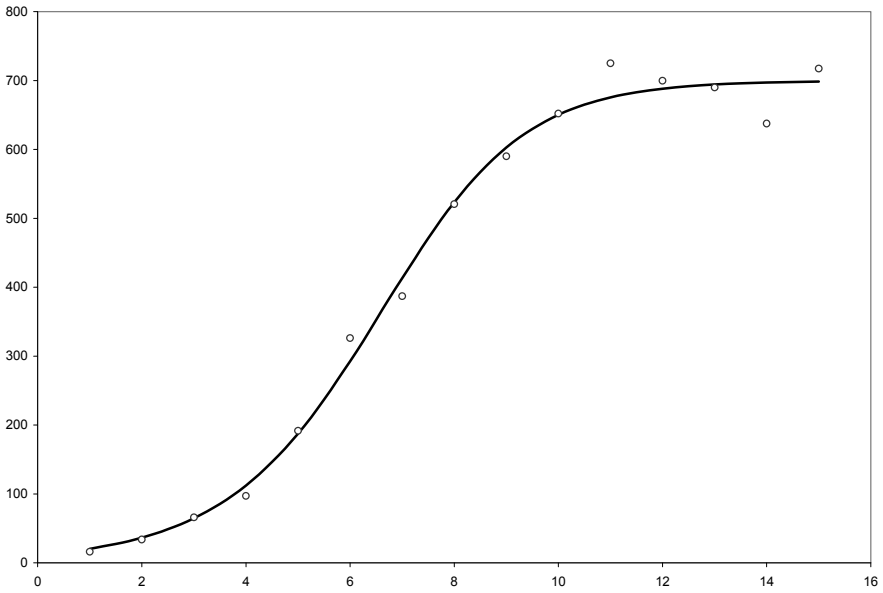
The dataset used for this comparison has 15 observations and the proper nonlinear regression model contains 4 parameters and is of the following functional form

$$y_i = \frac{\beta_1}{\left[1 + \exp\left(\beta_2 - \beta_3 x_i\right)\right]^{1/\beta_4}} + \varepsilon_i . \tag{3.1}$$

This model was estimated by the method of nonlinear least squares. The certified values of the parameters provided by Information Technology Laboratory of National Institute of Standards and Technology (NIST) are:

$$\hat{\beta}_1 = b_1 = 699{,}6415$$
$$\hat{\beta}_2 = b_2 = 5{,}2771$$
$$\hat{\beta}_3 = b_3 = 0{,}7596$$
$$\hat{\beta}_4 = b_4 = 1{,}2792.$$

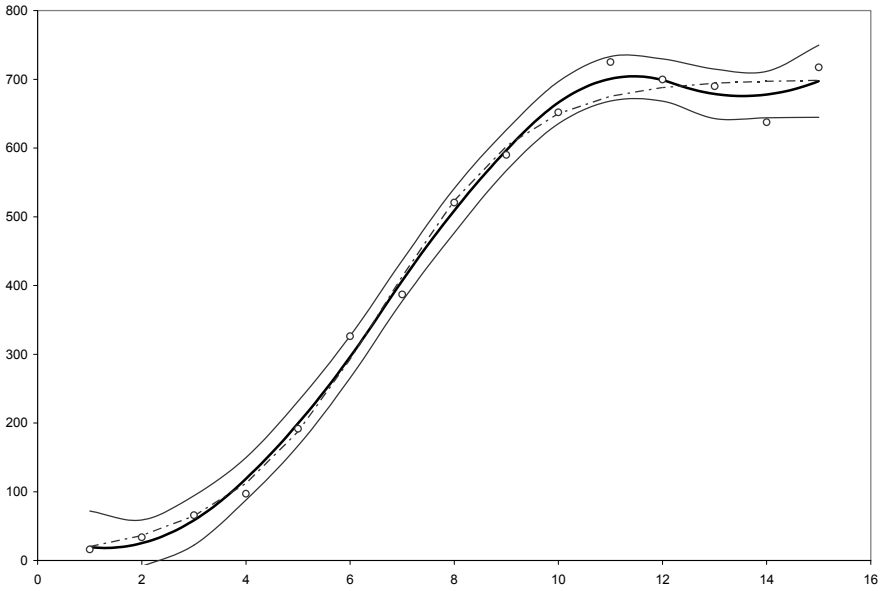
As far as it is not the main aim of this article, the well-known method of nonlinear least squares is not described in this paper. For any details of this method see, for example, Seber and Wild (1989). The estimated nonlinear relationship between onion bulb dry weight and growing time is depicted in the following figure.



3.2 Penalized spline model

The penalized spline regression model with a quadratic spline basis and four knots was used for the estimation of the relationship between onion bulb dry weight and growing time. The appropriate smoothing parameter was selected by minimizing the cross-validation function and the value of the chosen smoothing parameter was $h = 1{,}05$.

The corresponding penalized spline regression fit, estimated using eq. (2.4), is depicted with its confidence interval in the following graph.



It is obvious that there are some discrepancies between the smooth curve estimated by the penalized spline regression method and the estimation obtained by nonlinear least squares. These differences are the most noticeable for the values of growing time higher than 10.

It is possible to test for the statistical significance of this difference, e.g. by using the approximate F-test, which is based on the following statistic

$$F = \frac{R^2_{\text{larger}} - R^2_{\text{smaller}}}{\left(1 - R^2_{\text{larger}}\right)\left(df_{\text{res, smaller}} - df_{\text{res, larger}}\right) / df_{\text{res, larger}}}, \tag{3.2}$$

where R^2 is a square of the correlation coefficient between $\mathbf{y}$ and $\hat{\mathbf{y}}$. Under the null hypothesis (parametric model), this statistic will have an approximate F-distribution with $df_{\text{res, smaller}} - df_{\text{res, larger}}$ and $df_{\text{res, larger}}$ degrees of freedom.

For the given dataset the value of the corresponding F-statistic is 2,165 and the corresponding p-value is 0,177, which means that the difference between the penalized spline estimator and the nonlinear fit is neither statistically nor practically significant.

4 Conclusion

This paper describes the so-called penalized spline regression and tries to compare its performance with the estimation based on the parametric nonlinear model. Since the nonlinear function was carefully chosen, both of these methods gave satisfactory results modeling the nonlinear relationship between the dry weight of onion bulbs as the response variable and the growing time as the explanatory variable.

One of the biggest advantages of the penalized spline regression is that the relationship between a response variable and explanatory variables is modeled in a nonparametric manner, i.e. no prior

choice of the functional form of the model is needed. That is why it is very sensible at least to use this method to guide the choice of an appropriate parametric model or to test for the adequacy of any linear or nonlinear model. Moreover, in many cases the use of the penalized spline regression alone is often sufficient.

To conclude, there are several possible modifications of the simple spline model, if some of the assumptions are violated. Variance function estimation can be used in case of heteroscedasticity in the data or interaction models can be used when the assumption of additivity is not justified. Other modifications include generalized additive models, spatially adaptive models or Bayesian semiparametric regression. For further details see Ruppert et al. (2003).

Acknowledgement

This paper was supported by grant no. IG410040 from Grant Agency of University of Economics in Prague.

References

[1.] EUBANK, R. L.: *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York, 1999.
[2.] HÄRDLE, W., MÜLLER, M., SPERLICH, S., WERWATZ, A.: *Nonparametric and Semiparametric Models*. Springer-Verlag, Berlin Heidelberg, 2004.
[3.] RATKOWSKY, D. A.: *Nonlinear Regression Modeling: A Unified Practical Approach*. Marcel Dekker, New York, 1983.
[4.] RUPPERT, D., WAND, M. P., Carroll, R. J.: *Semiparametric Regression*. Cambridge University Press, New York, 2003.
[5.] SEBER, G. A. F., WILD, C. J.: *Nonlinear Regression*. John Wiley and Sons, New York, 1989.
[6.] National Institute of Standards and Technology, http://www.nist.gov/itl/

Current address

Tomáš Marcinko, Ing.
Vysoká škola ekonomická v Praze,
nám. W. Churchilla 4, 130 67 Praha 3,
e-mail: xmart14@vse.cz

Journal of Applied Mathematics

volume IV (2011), number III

LABOR MARKET AND SIMULTANEOUS EQUATIONS SOLVED BY TSLS

MISKOLCZI Martina, (CZ), LANGHAMROVÁ Jitka, (CZ)

Abstract. Labor market (employment and unemployment) is influenced by various economic indicators. The paper is concerned with trend on labor market measured by number of employed and unemployed individuals based on changes in selected economic variables.
First, multiple linear regression model is proposed and estimated by OLS method using data representing the Czech Republic 2003-2009. It confirms that inflation and wage are statistically significant for number of unemployed individuals.
Second, simultaneous equations are defined where number of employed and unemployed individuals depend on each other. Unknown coefficients are estimated by TSLS method. Proposed model consisting of three equations and including both exogenous and endogenous variables has to be simplified because of threat of multicolinearity. Relationship between basic macroeconomic indicators is complex and complicates adaptation of theoretical macroeconomic constructions to current real data and real situation. Results show that number of unemployed, export, wage and number of employed individuals in previous quarter are statistically significant variables for number of employed individuals, investments and GDP growth rate are statistically insignificant. In 2nd equation there are statistically significant number of employed individuals (both for previous and actual quarter), number of unemployed individuals (delayed) and GDP growth rate, inflation and wage are statistically insignificant.
Further, elasticity coefficient and prediction for 2010-2011 is introduced in the paper.

Key words. labor market, employment, unemployment, linear regression, OLS, simultaneous equations, TSLS

Mathematics Subject Classification: Primary 62P20, 93E24; Secondary 91B40, 91B64, 91B70, 91B99.

First Section

Supply side of labor market is formed by economic active (EA) people who could be either employed (E) or unemployed (U). Labor supply is composed of individuals offering their workforce; labor demand is created by employers searching best employees for offered job positions. Both employment and unemployment affect performance of each economy and are

affected by economic performance, respectively. Their balance is directly connected with social and consequently political stability.

1.1 Objectives Definition

Objective of this article is to propose model of simultaneous equations that would describe labor market in the Czech Republic and estimate its parameters. The econometric model introduces set of macroeconomic and other indicators that influence changes on labor market.
First, there will be proposed model of linear regression which will be estimated by ordinary least square method (OLS). Second, there will be proposed model of simultaneous equations which will be tested for multicolinearity and estimated by two-stage least square method (TSLS, 2SLS). All results will be tested at significance level $\alpha = 0.05$.
Data for estimate come from Labor Force Sample Survey (LFSS) conducted quarterly and from Czech Statistical Office (CZSO, www.czso.cz) for period 2003/Q1 – 2009/Q4. Calculated GDP growth rate is based on quarter-of-quarter (QoQ) increase.

Multiple Linear Regression

1.2 Economic Assumptions

For one-equation model the *number of unemployed* will be chosen as dependent variable.
Number of unemployed individuals depends on business cycle phase (growth/recession) which is measured by GDP trend and *GDP growth rate*. It also depends on *inflation* (Phillips curve) and level of *wages*. Other possible indicators were eliminated from model because of threat of multicolinearity which would disqualify any estimates.
Assumptions:

- as inflation grows number of unemployed decreases (inversely proportional). This assumption comes from Phillips curve economic theory,
- as wages grow number of unemployed decreases (inversely proportional). If economy is in growing phase it is connected with creation of new job positions, demand after workforce and growth of wages in order to attract more job applicants. Reverse interdependence between declining wages and increasing number of unemployed individuals is not obvious in modern economies because wages are not elastic in case of necessary reduction during recession and economic crisis; more likely their growth decelerates or stops. This fact could affect results.
- as GDP growth rate grows number of unemployed decreases because economy is in its growing phase (inversely proportional). Similarly to previous assumption there is need of new job applicants. Reversely, negative GDP growth rate (recession) induces increasing number of unemployed people and unemployment rate because it is connected with reduction of job positions.

1.3 Econometric model

One-equation linear model with constant and stochastic variable:

$$y_t = \beta_1 x_{1t} + \beta_{10} x_{10,t} + \beta_{11} x_{11,t} + \beta_{15} x_{15,t} + u_t , \quad (1)$$

where

y	number of unemployed individuals [000]	Source: LFSS
x_1	unit vector	
x_{10}	inflation [%]	Source. CZSO
x_{11}	average gross monthly wage - full time equivalent [000 CZK]	Source. CZSO
x_{15}	GDP growth rate (quarter-of-quarter) [%] (GDP: previous year average prices, seasonally adjusted), own calculation	Source: CZSO

(see Appendix A)

1.4 Estimate by OLS

Estimate was obtained using OLS method:

$$y_t = 634.196 - 18.857\, x_{10,t} - 11.923\, x_{11,t} + 1.483\, x_{15,t}\,, \tag{2}$$

1.5 Verification

Economic Verification

- If inflation grows by 1 percentage point number of unemployed decreases by 18,857 individuals. Assumption made for effect of inflation was confirmed.
- If average monthly wage increases by 1 thousand CZK number of unemployed decreases by 11,923 individuals. Assumption made for effect of wage was confirmed even though null hypothesis was rejected narrowly. This reflects low elasticity of wages in case of decline mentioned above.
- If GDP growth rate increases by 1 percentage point number of unemployed decreases by 1,483 individuals. Assumption made for GDP growth rate was not confirmed. Real data show that correlation coefficient between number of unemployed and QoQ GDP growth rate based on GDP in real prices and seasonally adjusted is +0.290 although correlation between number of unemployed and GDP itself is –0.788 as expected from theory. This could be explained by long period that is analyzed here (28 quarters) and fact that GDP was already modified – seasonally adjusted, in real prices whereas unemployment shows high level of seasonality.

Statistical Verification

Individual t-tests for all estimates show that constant, inflation and wage are statistically significant variables and should be included in a model whereas GDP growth rate is statistically insignificant.

variable	estimate	standard error	t-value	critical value	
x_1	634.196	81.415	7.790	2.064	**
x_{10}	-18.857	4.938	3.819		**
x_{11}	-11.923	4.041	2.950		*
x_{15}	1.483	6.336	0.234		

Coefficient of determination: $R^2 = 64.84\%$
It shows that 64.84% of variance of dependent variable is explained by estimated model.

Durbin-Watson test for autocorrelation of residuals: $DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = 0.4128$
Null hypothesis of no autocorrelation in residuals is rejected. There is autocorrelation of first order in residuals in model. This systematic part could be eliminated by including delayed dependent variable for example.

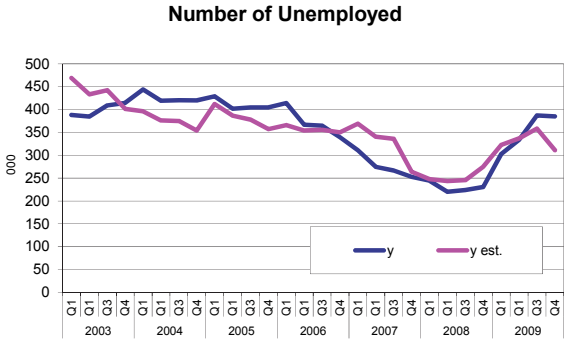


Figure 1: Number of unemployed – observations and predicted time series

Simultaneous Equations

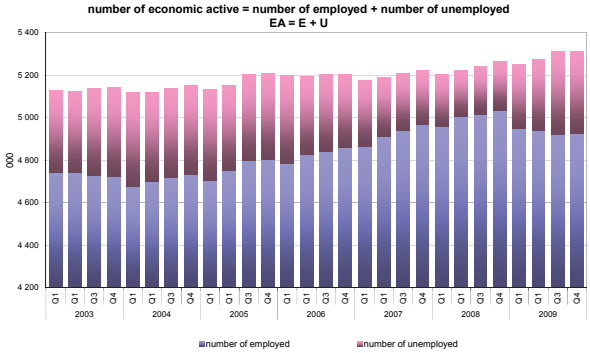


Figure 2: Number of employed and unemployed individuals (CR, 2003/Q1 – 2009/Q4)
Source: LFSS (CZSO)

1.6 Economic Assumptions

In simultaneous econometric model it is possible to propose more complicated construction where dependent variables – *number of employed* and *number of unemployed individuals* – depend on each other.
There is set of economic indicators that may affect number of employed and unemployed individuals (exogenous variables): rate of unemployment, trend in GDP, GDP delayed and its growth rate, final consumption of households (C) and government (G), investments (I), export (Ex), import (Im), net foreign trade, inflation and wages. Also delayed endogenous variables are incorporated in the analysis: number of employed and unemployed individuals in previous quarter in order to deal with possible autocorrelation.

1st and 2nd equation:
- number of employed and unemployed are reciprocal variables (inversely proportional),
- as unemployment rate grows number employed decreases,
- as GDP growth rate grows number of unemployed decreases because economy is in its growing phase (inversely proportional) and number of employed increases (from previous one-equation model it is obvious that relation between GDP and employed or unemployed individuals holds but indicator GDP growth rate does not have evidential influence for entire analyzed period),
- growth of consumption of households, investments, export or net export is connected with increase of employment and decrease of unemployment. These variables characterize phase of business cycle (growth/recession),
- as inflation grows number of employed increases (direct proportion) and number of unemployed decreases (inversely proportional, Phillips curve economic theory),
- as wages grow number of employed increases (direct proportion) and number of unemployed decreases (inversely proportional) with problematic trend during recession as wages are not elastic in case of decline,
- delayed endogenous variables have positive (directly proportional) effect on actual variable.

3rd equation: E + U = EA (identity, no unknown parameters are included in this equation)

1.7 Econometric Model

Simultaneous dynamic model is composed of three equations. First and second equations are stochastic, third is identity.

$$y_{1t} = \beta_{12}\, y_{2t} + \gamma_{11}\, x_{1t} + \gamma_{13}\, x_{3t} + \gamma_{14}\, x_{4t} + \gamma_{15}\, x_{5t} + \gamma_{16}\, x_{6t} + \gamma_{18}\, x_{8t} + \gamma_{1,11}\, x_{11t} + \gamma_{1,13}\, x_{13,t} + \gamma_{1,15}\, x_{15,t} + u_{1t}$$
$$y_{2t} = \beta_{21}\, y_{1t} + \gamma_{21}\, x_{1t} + \gamma_{22}\, x_{2t} + \gamma_{23}\, x_{3t} + \gamma_{2,10}\, x_{10,t} + \gamma_{2,11}\, x_{11,t} + \gamma_{2,12}\, x_{12,t} + \gamma_{2,13}\, x_{13,t} + \gamma_{2,14}\, x_{14,t} + \gamma_{2,15}\, x_{15,t} + u_{2t}$$
$$y_{3t} = y_{1t} + y_{2t}\,, \qquad (3)$$

where

y_1	number of employed (E) [000]	Source. CZSO
y_2	number of unemployed (U) [000]	Source. CZSO
y_3	number of economic active (EA) [000]	Source. CZSO
x_1	unit vector	
x_2	unemployment rate [%]	Source. CZSO
x_3	GDP, previous year average prices, seasonally adjusted [bn CZK]	Source. CZSO
x_4	final consumption (C) - households [bn CZK]	Source. CZSO
x_5	final consumption (G) - government [bn CZK]	Source. CZSO
x_6	investments [bn CZK]	Source. CZSO
x_7	net foreign trade [bn CZK]	Source. CZSO
x_8	export [bn CZK]	Source. CZSO
x_9	import [bn CZK]	Source. CZSO
x_{10}	inflation [%]	Source. CZSO
x_{11}	average gross monthly wage - full time equivalent [000 CZK]	Source. CZSO
x_{12}	GDP, previous year average prices, seasonally adjusted (t-1) [bn CZK]	Source. CZSO
x_{13}	number of employed (E) (t-1) [000]	Source. CZSO
x_{14}	number of unemployed (U) (t-1) [000]	Source. CZSO
x_{15}	GDP growth rate (quarter-of-quarter) [%] (GDP, previous year average prices, seasonally adjusted), own calculation	Source: CZSO

(see Appendix A)

1.8 Analysis of Correlation

Analysis of correlation shows that large number of selected variables is highly correlated. This would cause multicolinearity which degrades model estimates and its further use. Negative effect could be eliminated by reduction of model (elimination of some variables) or using absolute or relative increments instead of original data.

	y1	y2	y3	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
y1	1,0000	-0,8993	0,8229	-0,9193	0,9452	0,9226	0,8751	0,7488	0,7741	0,8814	0,8833	0,5303	0,8913	0,9557	0,9573	-0,9052	-0,4453
y2	-0,8993	1,0000	-0,4915	0,9988	-0,7881	-0,7074	-0,6179	-0,8344	-0,6342	-0,8127	-0,8355	-0,6762	-0,6575	-0,7862	-0,8374	0,9350	0,2897
y3	0,8229	-0,4915	1,0000	-0,5328	0,8582	0,9180	0,9397	0,4071	0,7175	0,6993	0,6733	0,1774	0,9206	0,8814	0,8183	-0,5877	-0,5104
x2	-0,9193	0,9988	-0,5328	1,0000	-0,8138	-0,7381	-0,6526	-0,8340	-0,6570	-0,8296	-0,8503	-0,6660	-0,6894	-0,8129	-0,8584	0,9400	0,3078
x3	0,9452	-0,7881	0,8582	-0,8138	1,0000	0,9596	0,9313	0,7510	0,8804	0,9424	0,9304	0,5182	0,9073	0,9935	0,9037	-0,7865	-0,3305
x4	0,9226	-0,7074	0,9180	-0,7381	0,9596	1,0000	0,9712	0,6021	0,7667	0,8289	0,8203	0,3714	0,9338	0,9703	0,9279	-0,7782	-0,4458
x5	0,8751	-0,6179	0,9397	-0,6526	0,9313	0,9712	1,0000	0,5084	0,7739	0,7860	0,7654	0,2774	0,9015	0,9381	0,8778	-0,6875	-0,4011
x6	0,7488	-0,8344	0,4071	-0,8340	0,7510	0,6021	0,5084	1,0000	0,5929	0,8271	0,8643	0,6316	0,5787	0,7383	0,6503	-0,7229	-0,2065
x7	0,7741	-0,6342	0,7175	-0,6570	0,8804	0,7667	0,7739	0,5929	1,0000	0,9091	0,8575	0,5864	0,7469	0,8446	0,6947	-0,5642	-0,0498
x8	0,8814	-0,8127	0,6993	-0,8296	0,9424	0,8289	0,7860	0,8271	0,9091	1,0000	0,9939	0,6813	0,8086	0,9164	0,7893	-0,7247	-0,1702
x9	0,8833	-0,8355	0,6733	-0,8503	0,9304	0,8203	0,7654	0,8643	0,8575	0,9939	1,0000	0,6858	0,8005	0,9077	0,7906	-0,7454	-0,1971
x10	0,5303	-0,6762	0,1774	-0,6660	0,5182	0,3714	0,2774	0,6316	0,5864	0,6813	0,6858	1,0000	0,3740	0,4969	0,4410	-0,5656	-0,0561
x11	0,8913	-0,6575	0,9206	-0,6894	0,9073	0,9338	0,9015	0,5787	0,7469	0,8086	0,8005	0,3740	1,0000	0,9263	0,8497	-0,6823	-0,5050
x12	0,9557	-0,7862	0,8814	-0,8129	0,9935	0,9703	0,9381	0,7383	0,8446	0,9164	0,9077	0,4969	0,9263	1,0000	0,9162	-0,8004	-0,4348
x13	0,9573	-0,8374	0,8183	-0,8584	0,9037	0,9279	0,8778	0,6503	0,6947	0,7893	0,7906	0,4410	0,8497	0,9162	1,0000	-0,9276	-0,4276
x14	-0,9052	0,9350	-0,5877	0,9400	-0,7865	-0,7782	-0,6875	-0,7229	-0,5642	-0,7247	-0,7454	-0,5656	-0,6823	-0,8004	-0,9276	1,0000	0,3995
x15	-0,4453	0,2897	-0,5104	0,3078	-0,3305	-0,4458	-0,4011	-0,2065	-0,0498	-0,1702	-0,1971	-0,0561	-0,5050	-0,4348	-0,4276	0,3995	1,0000

Final econometric model:

$$y_{1t} = \beta_{12}\, y_{2t} + \gamma_{11}\, x_{1t} + \gamma_{16}\, x_{6t} + \gamma_{18}\, x_{8t} + \gamma_{1,11}\, x_{11t} + \gamma_{1,13}\, x_{13,t} + \gamma_{1,15}\, x_{15,t} + u_{1t}$$
$$y_{2t} = \beta_{21}\, y_{1t} + \gamma_{21}\, x_{1t} + \gamma_{2,10}\, x_{10,t} + \gamma_{2,11}\, x_{11,t} + \gamma_{2,13}\, x_{13,t} + \gamma_{2,14}\, x_{14,t} + \gamma_{2,15}\, x_{15,t} + u_{2t}$$
$$y_{3t} = y_{1t} + y_{2t}\,. \qquad (4)$$

Currently, there are three endogenous and eight predetermined variables, six of them exogenous and two of them endogenous delayed (x_{13}, x_{14}).

1.9 Estimate by TSLS

Estimate of structural form of simultaneous equations:

$$y_{1t} = -0.426\, y_{2t} + 2{,}896.737\, x_{1t} - 0.179\, x_{6t} + 0.197\, x_{8t} + 7.782\, x_{11,t} + 0.386\, x_{13,t} - 5.116\, x_{15,t} + u_{1t}$$
$$y_{2t} = -0.706\, y_{1t} + 653.565\, x_{1t} - 1.954\, x_{10,t} + 3.119\, x_{11,t} + 0.571\, x_{13,t} + 0.856\, x_{14,t} - 4.690\, x_{15,t} + u_{2t}$$
$$y_{3t} = y_{1t} + y_{2t}\,. \qquad (5)$$

1.10 Verification

Economic Verification

1st equation	Assumption confirmed
If number of unemployed individuals grows by 1 thousand number of employed decreases by 426.	**YES**
If investments grow by 1 bn CZK number of employed decreases by 179.	**NO**
If export grows by 1 bn CZK number of employed increases by 197.	**YES**
If average monthly wage grows by 1 thousands CZK number of employed increases by 7,782.	**YES**
If number of employed individuals in previous quarter would grow hypothetically by 1 thousand then number of employed individuals in actual quarter is higher by 386.	**YES**
If GDP growth rate grows by 1 percentage point number of employed decreases by 5,116.	**NO**

2nd equation	Assumption confirmed
If number of employed individuals grows by 1 thousand number of unemployed decreases by 706.	**YES**
If inflation grows by 1 percentage point number of unemployed decreases by 1,954.	**YES**
If average monthly wage grows by 1 thousand CZK number of unemployed increases by 3,119.	**NO**
If number of employed individuals in previous quarter would grow hypothetically by 1 thousand then number of unemployed individuals in actual quarter is higher by 571.	**NO**
If number of unemployed individuals in previous quarter would grow hypothetically by 1 thousand then number of unemployed individuals in actual quarter is higher by 856.	**YES**
If GDP growth rate grows by 1 percentage point number of unemployed decreases by 4,690.	**YES**

Inversely proportional relation between both dependent variable was confirmed (number of employed and unemployed individuals have inverse trend).
1st equation: effect of investments and GDP growth rate was not confirmed and both variables are statistically insignificant for trend of employed individuals. Weak impact of investments could be explained by delay of reaction of labor market on investments in economy and creation of new job positions. GDP growth rate (QoQ) is insignificant as well, similarly to one-equation model: real data show that correlation coefficient between number of employed and QoQ GDP growth rate based on GDP in real prices and seasonally adjusted is –0.445 although correlation between number of employed and GDP itself is +0.945 as expected from theory. This could be explained by long period that is analyzed here (28 quarters) and fact that GDP was already modified – seasonally adjusted, in real prices whereas labor market shows high level of seasonality.
2nd equation: number of unemployed individuals is not affected by wages and number of employed individuals in previous quarter. There are situations in analyzed 28 quarters when wages grow and unemployment grows as well. This was already explained above by low elasticity of wages. This variable is also statistically insignificant. Delayed number of employed individuals is statistically significant but shows opposite relation than was expected and what suggested correlation coefficient (–0.837). This is probably explainable by multicolinearity that is present in model in spite of substantial effort of its elimination.

Statistical Verification

Individual t-tests in 1st equation show that constant, number of unemployed, export, wage and number of employed individuals in previous quarter are statistically significant variables compared to investments and GDP growth rate which are statistically insignificant.
In 2nd equation there are statistically significant number of employed individuals (both for previous and actual quarter), number of unemployed individuals (for previous quarter) and GDP growth rate, statistically insignificant are constant, inflation and wage. Compared to one-equation linear regression model these results do not correspond.

variable	estimate	standard error	t-value	critical value	
y_2	-0.426	0.138	3.098	2.080	*
x_1	2896.737	432.496	6.698		**
x_6	-0.179	0.299	0.600		
x_8	0.197	0.077	2.555		*

x_{11}	7.782	3.176	2.451		*
x_{13}	0.386	0.082	4.697		**
x_{15}	-5.116	2.566	1.993		

variable	estimate	standard error	t-value	critical value	
y_1	-0.706	0.184	3.834	2.080	*
x_1	653.565	924.769	0.707		
x_{10}	-1.954	2.032	0.962		
x_{11}	3.119	4.236	0.736		
x_{13}	0.571	0.129	4.430		**
x_{14}	0.856	0.190	4.510		**
x_{15}	-4.690	2.100	2.233		*

Index of determination: 1st equation: $I^2 = 98.60\%$; 2nd equation: $I^2 = 97.16\%$
Very high indices of determination confirm that estimated econometric model predicts real trend and changes in endogenous variables very well.

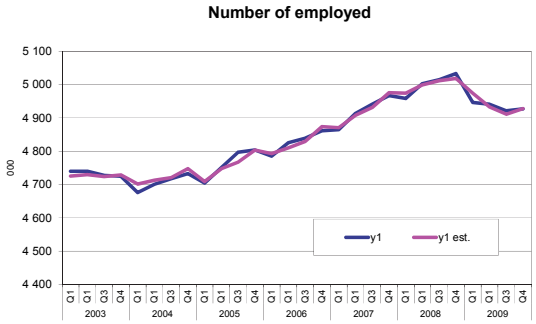


Figure 3: Number of employed – real and predicted time series (CR, 2003/Q1 – 2009/Q4)

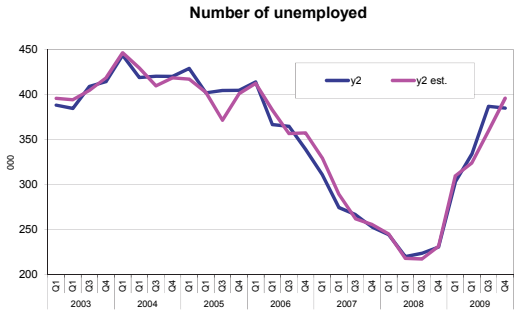


Figure 4: Number of unemployed – real and predicted time series (CR, 2003/Q1 – 2009/Q4)

1.11 Calculation of Reduced Form

Reduced form represents endogenous variables as linear combination of all predetermined variables included in simultaneous model without dependencies on other endogenous variables.

$$y_{1t} = 3{,}745.331 - 0.256\, x_{6t} + 0.282\, x_{8t} + 1.192\, x_{10,t} + 9.231\, x_{11,t} + 0.204\, x_{13,t} - 0.522\, x_{14,t} - 4.457\, x_{15,t} + v_{1t}$$
$$y_{2t} = -1{,}989.839 + 0.181\, x_{6t} - 0.199\, x_{8t} - 2.795\, x_{10,t} - 3.396\, x_{11,t} + 0.428\, x_{13,t} + 1.225\, x_{14,t} - 1.545\, x_{15,t} + v_{2t}$$

$$y_{3t} = 1{,}755.492 - 0.075\, x_{6t} + 0.083\, x_{8t} - 1.603\, x_{10,t} + 5.835\, x_{11,t} + 0.631\, x_{13,t} + 0.703\, x_{14,t} - 6.002\, x_{15,t} + v_{3t} \quad (6)$$

Comparison of estimate of structural and reduced forms of simultaneous equations show that mostly are these estimates consistent, i.e. both direction (sign) and intensity correspond. Differences were found for average wage in 2nd equation only.
There are very similar coefficients in third equation as in the first equation because number of employed individuals form approximately 93% of economic active people.

1.12 Coefficients of Elasticity

Coefficient of elasticity is calculated as percentage change of endogenous variable evoked by 1-percent increase of selected variable. Sensitive reaction, i.e. elasticity above 1 or under –1 was found in second equation for relation between number of unemployed and number of employed (both previous and actual value).
For example: if number of employed grows by 1% number of unemployed decreases by 9.704%.

1st equation	elasticity	2nd equation	elasticity
number of unemployed	-0.031	number of employed	-9.704
investments	-0.008	inflation	-0.014
export	0.024	wages	0.176
wages	0.032	number of employed (t-1)	7.850
number of employed (t-1)	0.386	number of unemployed (t-1)	0.855
GDP growth rate	-0.001	GDP growth rate	-0.019

1.13 Application of Simultaneous Equations – prediction for 2010 and 2011

One of possible application of reduced form of simultaneous equations is to predict values of endogenous variables based on selected scenario for predetermined variables. It could be theoretical what-if analysis or prediction of future values.
Here, real information for 2010/Q1 – 2010/Q3 and prediction for 2010/Q4 – 2011/Q4 were used to obtain prediction for E, U and EA for next two years. Quality of such a prediction could be evaluated by comparison with real values.

		prediction			observations		
		number of employed	number of unemployed	number of economic active	number of employed	number of unemployed	number of economic active
		000	000	000	000	000	000
		y1	y2	y3	y1	y2	y3
2010	Q1	4,886.1	405.1	5,291.3	4,829.2	422.7	5,251.9
	Q2	4,887.3	410.9	5,298.2	4,880.9	374.7	5,255.6
	Q3	4,893.0	410.3	5,303.4	4,911.5	374.7	5,286.2
	Q4	4,897.0	415.4	5,312.4			
2011	Q1	4,891,6	421.3	5,312.9			
	Q2	4,884.9	428.0	5,312.8			

	Q3	4,882.6	431.5	5,314.1			
	Q4	4,895.7	430.2	5,325.9			

Calculation were not able to predict dramatic changes in first quarter of 2010 when number of unemployed was the highest and labor market faced serious impacts of crisis.

Conclusion

Introduced statistical methods contribute to analysis of labor market, employment and unemployment. There were explored various economic indicators as possible effects on number of employed and unemployed individuals but only limited amount of them was finally used to estimate unknown coefficients in linear regression model and three-equations simultaneous model.
Results are not consistent between linear regression and TSLS estimates of simultaneous equations. Inflation and wage were confirmed as statistically significant in one-equation regression model whereas were statistically insignificant in more complex simultaneous econometric model for dependent variable number of unemployed. This could be cause by the fact that other variable(s) included into simultaneous model were 'more significant' (number of employed in previous and actual quarter) which confirms extremely high coefficient of elasticity. This also verifies assumption that number of employed (E) and unemployed (U) are closely related variables.
Results of statistical calculations also suggest that some level of multicolinearity still remain in model although it was carefully examined. This show how complex and internally related are economical variables and reflects difficulties when using real economic data to demonstrate economical theories. Further analysis could use other statistical methods (such as factor analysis) that could eliminate effect of multicolinearity.

Acknowledgement

The paper was supported by research project 2D06026 with title "Reprodukce lidského kapitálu" – "Reproduction of Human Capital" financed by Ministry of Education, Youth and Sports within National program of research II.

References

[1.] ARLTOVÁ, M., LANGHAMROVÁ, J.: *Reprodukce lidského kapitálu*. In Acta Oeconomica Pragensia, Praha, Vol. 18, No. 2, pp. 96-98, 2010. ISSN 0572-3043.
[2.] BÍLKOVÁ, D.: *Application of Pareto Distribution in Modelling of Wage Distributions*. In 12th International scientific conference applications of mathematics and statistics in economy, Oeconomica, Praha, 2009, pp. 13-28. ISBN 978-80-245-1600-4.
[3.] FISCHER, J., MAZOUCH, P.: *Souvislosti vzdělanosti, nezaměstnanosti a ekonomického růstu z regionálního hlediska*. In Demografie [CD-ROM], 2007, Vol. 49, No. 4, pp. 182-188. ISSN 0011-8265.

[4.] FROYEN, R. T.: *Macroeconomics*. Theories and Policies. 2nd ed. Macmillan Publishing Company, New York, 1986. ISBN 0-02-339410-2.
[5.] HOLMAN, R.: *Ekonomie*. 2nd ed. C. H. Beck, Praha, 2001. ISBN 80-7179-255-1.
[6.] JÍROVÁ, H.: *Trh práce a politika nezaměstnanosti*. VŠE Praha, Praha, 2006. ISBN 80-7079-635-9.
[7.] KOTÝNKOVÁ, M.: *Trh práce na přelomu tisíciletí*. Oeconomica, Praha, 2006. ISBN 80-245-1149-5.
[8.] LANGHAMROVÁ, J., Fiala, T., Hulík, V., Miskolczi, M., Kačerová, E.: *Prognóza lidského kapitálu obyvatelstva České republiky do roku 2050*. In Demografie, 2010, Vol. 52, No. 3, pp. 181-196, 2010. ISSN 0011-8265.
[9.] LAW No. 435/2004 Sb., about employment
[10.] MACH, M.: *Makroekonomie II pro magisterské (inženýrské) studium 1. a 2.část*. Melandrium, Praha, 1998. ISBN: 80-86175-18-9.
[11.] MISKOLCZI, M.: *Trends in Unemployment in the Czech Republic and Regions*. In IDIMT-2010 Information Technology – Human Values, Innovation and Economy, Vol. 32, pp. 219-228. Trauner Verlag, Linz, 2010. ISBN 978-3-85499-760-3.
[12.] MISKOLCZI, M., LANGHAMROVÁ, J.: *Analysis of Unemployment of Males and Females*. In PEFnet 2010. Vydavatelství MU Brno, Brno, 2010. ISBN 978-80-7375-450-1.

Current address

Martina Miskolczi, Ing. Mgr. MBA
Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra demografie
nám. W. Churchilla 4
130 67 Praha 3, Česká republika
e-mail:martina.miskolczi@vse.cz

Jitka Langhamrová, Doc. Ing. CSc.
Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra demografie
nám. W. Churchilla 4
130 67 Praha 3, Česká republika
e-mail: langhamj@vse.cz

Appendix A: Input Data

		number of employed	number of unemployed	number of economic active	unit vector	rate of unemployment	GDP (seasonally adjusted)	C	G	I	Ex-Im	export	import	inflation	average gross monthly wage	GDP (t-1)	number of emplyed (t-1)	number of unemployed (t-1)	GDP growth rate (QoQ)
		000	000	000		%	bn CZK	bn CZK	bn CZK	bn CZK	bn CZK	bn CZK	bn CZK	%	000 CZK	bn CZK	000	000	%
		y1	y2	y3	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
2003	**Q1**	4 739,9	388,3	5 128,2	1	7,6	629,8	325,5	147,2	163,0	-9,6	393,9	403,5	-0,4	15,0	604,9	4 791,7	374,9	4,12
	Q2	4 740,0	384,7	5 124,7	1	7,5	637,9	328,5	145,9	177,8	-17,9	388,0	405,9	0,3	16,5	629,8	4 739,9	388,3	1,29
	Q3	4 727,8	409,1	5 136,9	1	8,0	638,5	333,4	147,1	172,0	-17,8	392,0	409,8	0,0	16,1	637,9	4 740,0	384,7	0,08
	Q4	4 724,9	414,5	5 139,4	1	8,1	647,5	335,2	148,5	181,9	-22,0	418,0	440,1	1,0	18,1	638,5	4 727,8	409,1	1,42
2004	**Q1**	4 675,9	443,8	5 119,7	1	8,7	658,4	333,7	145,4	183,8	-8,5	430,7	439,2	2,5	16,2	647,5	4 724,9	414,5	1,68
	Q2	4 700,6	419,1	5 119,7	1	8,2	667,6	335,9	148,5	192,9	-13,7	492,0	505,7	2,9	17,2	658,4	4 675,9	443,8	1,40
	Q3	4 717,4	420,4	5 137,8	1	8,2	678,0	342,3	144,5	191,1	-4,1	490,0	494,1	3,0	17,2	667,6	4 700,6	419,1	1,55
	Q4	4 732,7	420,2	5 152,9	1	8,2	685,1	342,7	143,5	192,5	2,3	503,9	501,6	2,8	19,2	678,0	4 717,4	420,4	1,05
2005	**Q1**	4 704,5	429,1	5 133,6	1	8,4	730,0	352,5	155,5	186,5	31,3	517,3	486,1	1,5	17,1	685,1	4 732,7	420,2	6,56
	Q2	4 750,7	402,1	5 152,8	1	7,8	740,9	355,6	162,6	188,3	30,1	536,8	506,8	1,8	18,1	730,0	4 704,5	429,1	1,50
	Q3	4 797,2	404,6	5 201,8	1	7,8	753,4	360,7	164,3	195,4	28,4	559,7	531,3	2,2	18,2	740,9	4 750,7	402,1	1,68
	Q4	4 803,7	404,8	5 208,5	1	7,8	766,2	365,5	157,1	196,3	42,8	586,4	543,6	2,2	20,0	753,4	4 797,2	404,6	1,70
2006	**Q1**	4 785,2	414,1	5 199,3	1	8,0	777,6	369,8	167,0	198,4	36,5	602,4	565,9	2,8	18,3	766,2	4 803,7	404,8	1,49
	Q2	4 825,9	366,8	5 192,7	1	7,1	791,8	376,7	166,6	211,9	30,7	604,1	573,4	2,8	19,3	777,6	4 785,2	414,1	1,83
	Q3	4 839,4	364,9	5 204,4	1	7,0	805,5	381,6	167,6	218,5	31,6	621,8	590,3	2,7	19,3	791,8	4 825,9	366,8	1,73
	Q4	4 861,7	339,3	5 201,0	1	6,5	815,5	388,4	165,1	214,0	41,6	671,1	629,5	1,7	21,3	805,5	4 839,4	364,9	1,25
2007	**Q1**	4 865,0	311,2	5 176,2	1	6,0	844,7	396,8	172,9	234,2	34,0	693,4	659,4	1,9	19,7	815,5	4 861,7	339,3	3,58
	Q2	4 913,9	274,6	5 188,5	1	5,3	849,5	403,2	169,7	242,1	27,6	682,5	654,9	2,5	20,7	844,7	4 865,0	311,2	0,56
	Q3	4 941,9	266,7	5 208,6	1	5,1	860,3	405,9	170,3	246,8	30,3	712,6	682,3	2,8	20,7	849,5	4 913,9	274,6	1,27
	Q4	4 967,2	252,8	5 220,0	1	4,8	868,9	407,6	177,5	224,1	52,7	749,3	696,5	5,4	22,6	860,3	4 941,9	266,7	1,01
2008	**Q1**	4 958,4	244,5	5 202,9	1	4,7	901,4	426,9	178,5	231,1	57,5	792,4	734,8	7,1	21,6	868,9	4 967,2	252,8	3,73
	Q2	5 003,3	220,1	5 223,4	1	4,2	908,1	430,3	180,5	223,3	66,5	773,3	706,8	6,7	22,2	901,4	4 958,4	244,5	0,74
	Q3	5 014,8	223,9	5 238,7	1	4,3	909,5	430,0	182,7	225,9	63,5	747,1	683,6	6,6	22,2	908,1	5 003,3	220,1	0,16
	Q4	5 033,4	230,7	5 264,2	1	4,4	901,5	430,7	183,0	247,0	33,3	683,5	650,2	3,6	24,3	909,5	5 014,8	223,9	-0,89
2009	**Q1**	4 946,8	302,8	5 249,6	1	5,8	885,0	450,8	188,0	218,4	20,1	608,7	588,6	2,3	22,3	901,5	5 033,4	230,7	-1,82
	Q2	4 941,3	333,9	5 275,2	1	6,3	880,5	451,1	190,4	193,5	37,4	619,1	581,7	1,2	23,0	885,0	4 946,8	302,8	-0,51
	Q3	4 921,7	387,0	5 308,7	1	7,3	885,1	449,2	197,3	193,7	36,6	651,4	614,8	0,0	23,2	880,5	4 941,3	333,9	0,52
	Q4	4 927,3	385,0	5 312,3	1	7,2	888,4	448,0	196,6	182,5	52,8	663,8	611,0	1,0	25,6	885,1	4 921,7	387,0	0,38
average		**4 841,5**	**352,1**	**5 193,6**	**1,0**	**6,8**	**786,0**	**384,2**	**166,6**	**204,5**	**24,8**	**592,3**	**567,6**	**2,5**	**19,8**	**775,8**	**4 836,7**	**351,8**	**1,39**

Journal of Applied Mathematics

volume IV (2011), number III

EXACT AND QUASI-EXACT CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO BINOMIAL PROPORTIONS

POBOČÍKOVÁ Ivana, (SK)

Abstract. 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 consider two exact methods: the Chan-Zhang interval, the Agresti-Min interval and quasi-exact method: the Chen interval. We compare the performance of the confidence intervals in terms of the coverage probability and the interval length.

Key words and phrases. binomial distribution, difference of two binomial proportions, confidence interval, coverage probability, interval length, Chan-Zhang interval, Agresti-Min interval, Chen interval.

Mathematics Subject Classification. Primary 60A05, 62F25.

1 Introduction

The confidence intervals for the difference of two independent binomial proportions are important problem in a biomedical research. Confidence intervals are often used in clinical trials to compare a new treatment with a standard treatment or placebo. We consider a clinical trial in which we want to compare the efficacy of a new treatment with a standard treatment. This situation can be represent with 2×2 contingency table

Table 1.

	new treatment	standard treatment
number of successes	X	Y
number of failures	$n_1 - X$	$n_2 - Y$
total	n_1	n_2

where $X \sim Bi(n_1, \pi_1)$ and $Y \sim Bi(n_2, \pi_2)$ be two independent binomial random variables. X denotes number of successes in n_1 independent trials, with the probability of success on a single

trial π_1 (new treatment) and Y denotes number of successes in n_2 independent trials, with the probability of success on a single trial π_2 (standard treatment).

Let $\delta = \pi_1 - \pi_2$ is the difference of two independent binomial proportions, $-1 < \delta < 1$. Let $\pi = \pi_1$ and substitute $\pi_2 = \pi - \delta$. The joint probability mass function can be expressed as

$$P(X = x,\, Y = y) = \left(\begin{array}{c} n_1 \\ x \end{array} \right) \pi^x (1-\pi)^{n_1 - x} \left(\begin{array}{c} n_2 \\ y \end{array} \right) (\pi - \delta)^y (1 - \pi + \delta)^{n_2 - y}, \tag{1}$$

for $x = 0, 1, ..., n_1,\ y = 0, 1, ..., n_2,\ \pi_i \in (0,\, 1),\ n_i \in \mathcal{N},\ i = 1,\, 2.$

For any given δ the domain of π is

$$D(\delta) = \{\pi :\ \max\{0,\, \delta\} \leq \pi \leq \min\{1,\, 1 + \delta\}\}. \tag{2}$$

It is known that π is a nuisance parameter for the inference on δ.

We want to find the $100 \times (1-\alpha)\,\%$ two-sided confidence interval $\langle \delta_L,\, \delta_U \rangle$ for the difference of two independent binomial proportions $\delta = \pi_1 - \pi_2$, where $(1 - \alpha)$ is the desired confidence level and $\alpha \in (0,\, 1)$.

The literature contains several methods for constructing confidence intervals for the difference of two independent binomial proportions. The most commonly used confidence interval is the Wald interval, which is based on the standard normal approximation. The lower and upper bounds of the $100 \times\ (1 - \alpha)\,\%$ Wald confidence interval are

$$\begin{aligned} \delta_L &= (p_1 - p_2) - k_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1\,(1-p_1)}{n_1} + \frac{p_2\,(1-p_2)}{n_2}}, \\ \delta_U &= (p_1 - p_2) + k_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1\,(1-p_1)}{n_1} + \frac{p_2\,(1-p_2)}{n_2}}, \end{aligned} \tag{3}$$

where $p_1 = \dfrac{X}{n_1}$, $p_2 = \dfrac{Y}{n_2}$ are maximum likelihood estimates of parameters π_1, π_2 and k_α is the $\alpha-$quantile of standard normal distribution $N(0,\, 1)$. It is known that this interval performs poorly (Agresti, Caffo, 2000, Newcombe, 1998).

In this paper we compare the two exact confidence intervals for the difference of two independent binomial proportions: the Chan-Zhang interval (Chan and Zhang, 1999), the Agresti-Min interval (Agresti and Min, 2001) and quasi-exact confidence interval: the Chen interval (Chen, 2002). All these confidence intervals are test based. They are constructed by inverting a hypothesis test under an appropriate alternative hypothesis.

We compare the performance of the 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 the Confidence Intervals

In this section we describe the three alternatives of the confidence intervals for the difference of two independent binomial proportions compared in this paper.

2.1 Chan - Zhang interval

The exact Chan-Zhang confidence interval (Chan and Zhang, 1999) is based on inverting two one-sided tests

$$H_0 : \delta = \delta_0 \qquad \text{versus} \qquad H_1 : \delta < \delta_0. \tag{4}$$

$$H_0 : \delta = \delta_0 \qquad \text{versus} \qquad H_1 : \delta > \delta_0. \tag{5}$$

Chan and Zhang (1999) used for testing (4) and (5) the score test statistic

$$Z(X,\, Y,\, \delta_0) = \frac{p_1 - p_2 - \delta_0}{\sqrt{\dfrac{\widetilde{p_1}\,(1-\widetilde{p_1})}{n_1} + \dfrac{\widetilde{p_2}\,(1-\widetilde{p_2})}{n_2}}}, \tag{6}$$

where $p_1 = \dfrac{X}{n_1}$, $p_2 = \dfrac{Y}{n_2}$ are maximum likelihood estimates of parameters π_1, π_2 and $\widetilde{p_1}$, $\widetilde{p_2}$ are maximum likelihood estimates of parameters π_1, π_2 under the restriction that $\widetilde{p_1} - \widetilde{p_2} = \delta_0$.

Miettinen, Nurminen (1985) showed that the unique restricted maximum likelihood estimates $\widetilde{p_1}$, $\widetilde{p_2}$ can be obtained for given $X = x$, $Y = y$ by solving the cubic equation

$$a_0\pi^3 + a_1\pi^2 + a_2\pi + a_3 = 0 \tag{7}$$

for $\pi \in \langle \max\{0,\, \delta_0\},\, \min\{1,\, 1+\delta_0\}\rangle$, where $a_0 = 1+\dfrac{n_2}{n_1}$, $a_1 = -\delta_0\left(2+\dfrac{n_2}{n_1}\right) - \dfrac{x}{n_1} - \dfrac{y}{n_1} - \dfrac{n_2}{n_1} - 1$, $a_2 = \delta_0^2 + \delta_0\left(2\,\dfrac{x}{n_1} + \dfrac{n_2}{n_1} + 1\right) + \dfrac{x}{n_1} + \dfrac{y}{n_1}$ and $a_3 = -\delta_0^2\,\dfrac{x}{n_1} - \delta_0\,\dfrac{x}{n_1}$.

Thus the restricted maximum likelihood estimates are

$$\widetilde{p_1} = 2\,u\,\cos(w) - \frac{a_1}{3\,a_0}, \qquad \widetilde{p_2} = \widetilde{p_1} - \delta_0, \tag{8}$$

where $v = \left(\dfrac{a_1}{3\,a_0}\right)^3 - \dfrac{a_1\,a_2}{6\,a_0^2} + \dfrac{a_3}{2\,a_0}$, $u = \operatorname{sgn}(v)\sqrt{\left(\dfrac{a_1}{3\,a_0}\right)^2 - \dfrac{a_2}{3\,a_0}}$ and $w = \dfrac{1}{3}\left(\pi + \arcsin\left(\dfrac{v}{u^3}\right)\right)$.

Chan and Zhang (1999) used the maximization method to eliminate the effect of the nuisance parameter π. The exact $p-$value is maximizing over all possible values of the nuisance parameters π. At first we invert the hypothesis (4).

For given $X = x$, $Y = y$ is the exact one-sided $p-$value for δ_0 defined by

$$\beta_{CZU}\,(x,\, y|Z,\, \delta_0) = \max_{\pi \in D(\delta_0)} \left\{ \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} P(X = i,\, Y = j|\delta_0,\, \pi)\, I\,(Z(i,\, j,\, \delta_0) \le Z(x,\, y,\, \delta_0)) \right\} \tag{9}$$

where $I(A \le B) = \begin{cases} 1 & \text{if } A \le B \\ 0 & \text{otherwise} \end{cases}$ is an indicator function.

The upper bound of the $100 \times \left(1 - \frac{\alpha}{2}\right)$ % confidence interval $(-1,\, \delta_U\rangle$ for δ is given by

$$\delta_U = \sup_{\delta}\{\beta_{CZU}\,(x,\, y|Z,\, \delta_0) \geq \frac{\alpha}{2}\}. \tag{10}$$

We obtain similarly the lower bound of the $100 \times \left(1 - \frac{\alpha}{2}\right)$ % confidence interval for δ. We invert the hypothesis (5). For given $X = x$, $Y = y$ is the exact one-sided $p-$value for δ_0 defined by

$$\beta_{CZL}\,(x,\, y|Z,\, \delta_0) = \max_{\pi \in D(\delta_0)} \left\{ \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} P(X = i,\, Y = j|\delta_0,\, \pi)\, I\left(Z(i,\, j,\, \delta_0) \geq Z(x,\, y,\, \delta_0)\right) \right\} \tag{11}$$

where $I(A \geq B) = \begin{cases} 1 & \text{if } A \geq B \\ 0 & \text{otherwise} \end{cases}$ is an indicator function.

The lower bound of the $100 \times \left(1 - \frac{\alpha}{2}\right)$ % confidence interval $\langle \delta_L,\, 1)$ for δ is given by

$$\delta_L = \inf_{\delta}\{\beta_{CZL}\,(x,\, y|Z,\, \delta_0) \geq \frac{\alpha}{2}\}. \tag{12}$$

Therefore the $100 \times\, (1 - \alpha)\,\%$ Chan-Zhang confidence interval for δ is $\langle \delta_L,\, \delta_U\rangle$.
This exact confidence interval satisfies

$$P(\delta_L \leq \delta \leq \delta_U) = 1 - P(\delta > \delta_U) - P(\delta < \delta_L) \geq 1 - \alpha. \tag{13}$$

2.2 Agresti - Min interval

The exact Agresti-Min interval (Agresti and Min, 2001) is similar to the Chan-Zhang interval, but is based on inverting a two-sided test

$$H_0 :\, \delta = \delta_0 \quad \text{versus} \quad H_1 :\, \delta \neq \delta_0. \tag{14}$$

Agresti and Min (2001) used for testing (14) the score statistic (6).

For given $X = x$, $Y = y$ is the exact Agresti-Min two-sided $p-$value for δ_0 defined by

$$\beta_{AM}\,(x,\, y|Z,\, \delta_0) = \max_{\pi \in D(\delta_0)} \left\{ \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} P(X = i,\, Y = j|\delta_0,\, \pi)\, I\left(|Z(i,\, j,\, \delta_0)| \geq |Z(x,\, y,\, \delta_0)|\right) \right\} \tag{15}$$

where $I(|A| \geq |B|) = \begin{cases} 1 & \text{if } |A| \geq |B| \\ 0 & \text{otherwise} \end{cases}$ is an indicator function.

The lower and upper bounds of the $100 \times\, (1 - \alpha)\,\%$ Agresti-Min confidence interval $\langle \delta_L,\, \delta_U\rangle$ for δ are given by

$$\delta_L \;= \inf_{\delta}\{\beta_{AM}\,(x,\, y|Z,\, \delta_0) \geq \alpha\}, \quad \delta_U \;= \sup_{\delta}\{\beta_{AM}\,(x,\, y|Z,\, \delta_0) \geq \alpha\}. \tag{16}$$

2.3 Chen interval

The Chen quasi-exact confidence interval (Chen, 2002) is based on the approximate $p-$value. Chen (2002) approximate the exact $p-$value (15) by the significance level at one single point, the maximum likelihood estimates $\widetilde{p}_1$, $\widetilde{p}_2$ under the restriction that $\widetilde{p}_1 - \widetilde{p}_2 = \delta_0$. Thus he obtained interval, which is simpler to compute. The approximate Chen $p-$value is then

$$\beta_{CQ}\left(x,\, y|Z,\, \delta_0\right) = \sum_{i=0}^{n_1}\sum_{j=0}^{n_2} P(X=i,\, Y=j|\delta_0,\, \widetilde{p}_1)\, I\left(|Z(i,\, j,\, \delta_0)| \geq |Z(x,\, y,\, \delta_0)|\right), \qquad (17)$$

where $I(|A| \geq |B|) = \left\{ \begin{array}{cc} 1 & \text{if } |A| \geq |B| \\ 0 & \text{otherwise} \end{array} \right.$ is an indicator function.

The lower and upper bounds of the $100 \times (1-\alpha)\,\%$ Chen quasi-exact confidence interval $\langle \delta_L,\, \delta_U \rangle$ for δ are given by

$$\delta_L \;\; = \inf_{\delta}\{\beta_{CQ}\left(x,\, y|Z,\, \delta_0\right) \geq \alpha\}, \quad \delta_U \;\; = \sup_{\delta}\{\beta_{CQ}\left(x,\, y|Z,\, \delta_0\right) \geq \alpha\}. \qquad (18)$$

Similarly we can obtain quasi-exact confidence interval from the Chan-Zhang interval.

3 Comparsion of the 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. At first we introduce the criteria for comparing of the confidence intervals.

3.1 Criteria for Comparing of Confidence Intervals

Coverage probability

The coverage probability of the confidence interval $\langle \delta_L,\, \delta_U \rangle$ is for fixed n_1, n_2 and π_1, π_2 defined by

$$C_{n_1,\, n_2}(\pi_1,\, \pi_2) = \sum_{x=0}^{n_1}\sum_{y=0}^{n_2} I(x,\, y,\, \pi_1,\, \pi_2) \left(\begin{array}{c} n_1 \\ x \end{array} \right) {\pi_1}^x \left(1-\pi_1\right)^{n_1-x} \left(\begin{array}{c} n_2 \\ y \end{array} \right) {\pi_2}^y \left(1-\pi_2\right)^{n_2-y} \qquad (19)$$

where $I(x,\, y,\, \pi_1,\, \pi_2) = \left\{ \begin{array}{cc} 1 & \text{if } \delta \in \langle \delta_U(x,y), \delta_L(x,y) \rangle \\ 0 & \text{otherwise} \end{array} \right.$ is an indicator function and $\delta = \pi_1 - \pi_2$.

The confidence interval is strict conservative if $C_{n_1,\, n_2}(\pi_1,\, \pi_2) \geq 1-\alpha$ for all $\pi_1,\, \pi_2 \in (0,\, 1)$ and $n_1,\, n_2 \in \mathcal{N}$.

Expected length

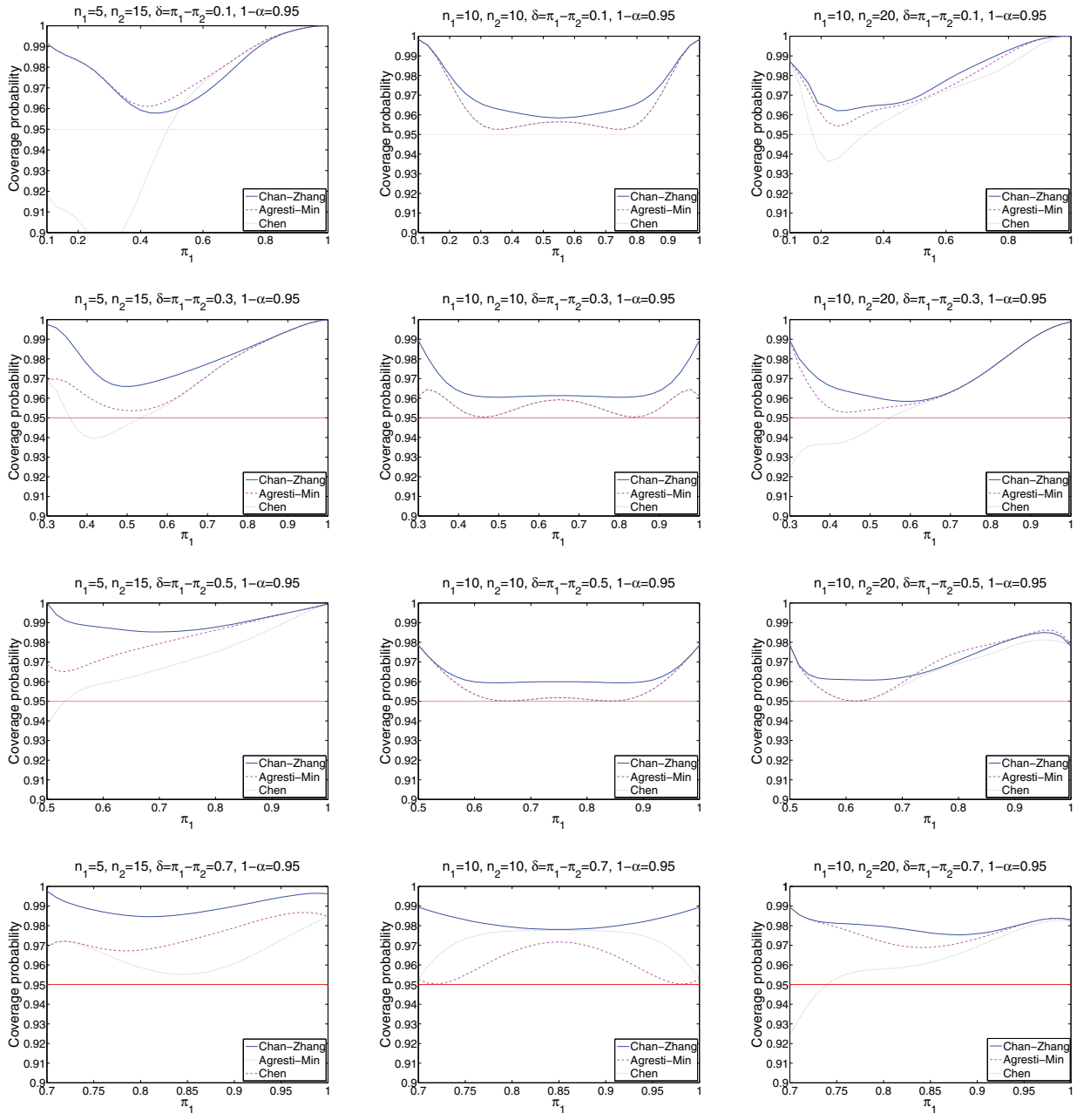


Figure 1: Coverage probability of the 95 % confidence intervals for $\delta = \pi_1 - \pi_2 = 0.1$, 0.3, 0.5, 0.7 and $(n_1,\, n_2) = (5,\, 15)$, $(10,\, 10)$, $(10,\, 20)$

The expected length of the confidence interval is defined by

$$EL_{n_1,n_2}(\pi_1,\, \pi_2) = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} [\delta_U(x,\, y) - \delta_L(x,\, y)] \left(\begin{array}{c} n_1 \\ x \end{array} \right) {\pi_1}^x \, (1-\pi_1)^{n_1-x} \left(\begin{array}{c} n_2 \\ y \end{array} \right) {\pi_2}^y \, (1-\pi_2)^{n_2-y} \quad (20)$$

where $\delta_L(x, y)$, $\delta_U(x, y)$ are lower and upper bounds of a particular confidence interval.

Average expected length

The average expected length (AVEL) of the confidence interval is defined by

$$AVEL(n_1,\, n_2) = \int_0^1 \int_0^1 EL_{n_1,n_2}(\pi_1,\, \pi_2)\, d\pi_1\, d\pi_2. \tag{21}$$

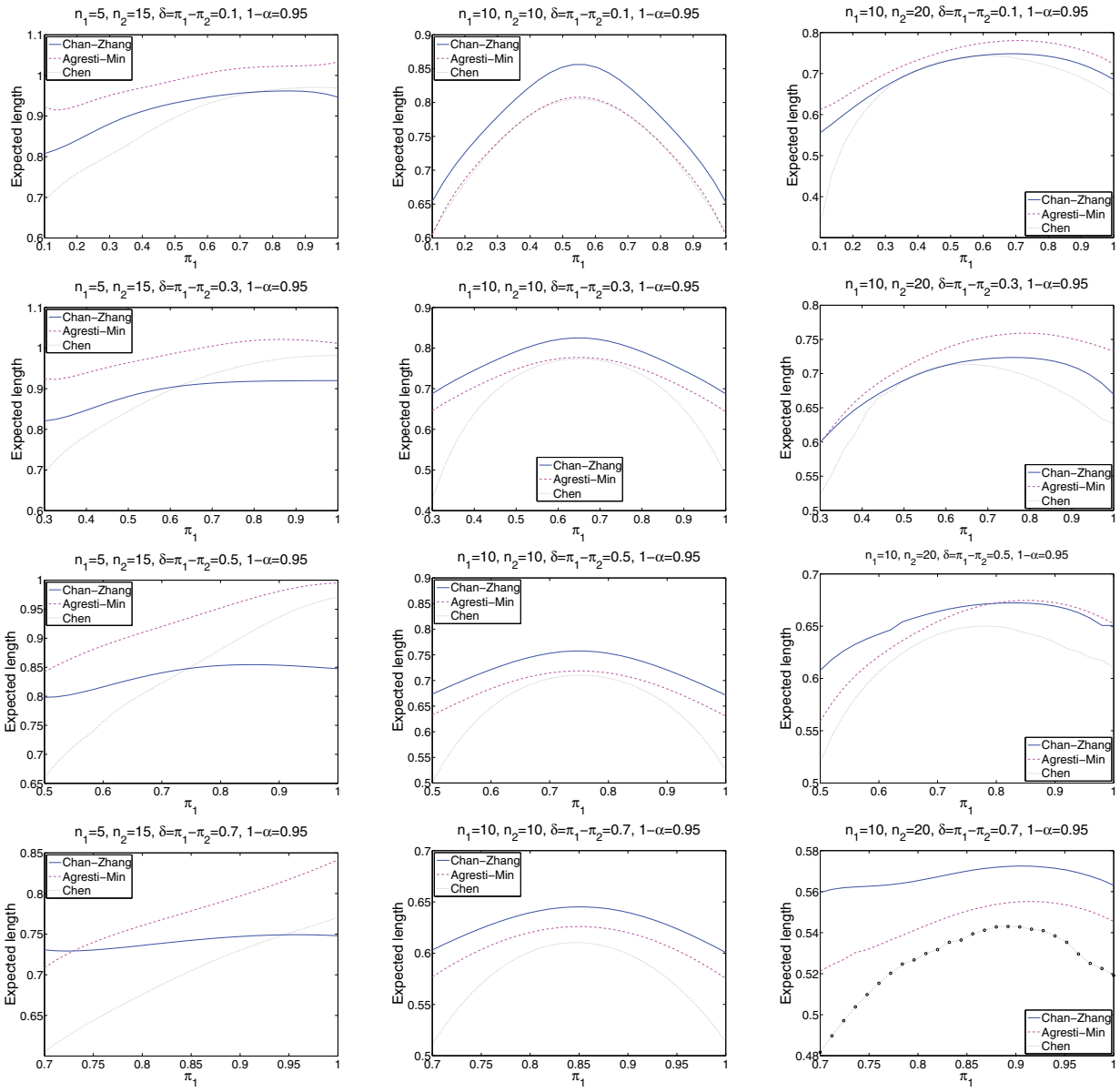


Figure 2: Expected lengths of the 95 % confidence intervals as a function of π_1 for $\delta = \pi_1 - \pi_2 =$ 0.1, 0.3, 0.5, 0.7 and $(n_1,\, n_2) = (5,\, 15)$, $(10,\, 10)$ and $(10,\, 20)$

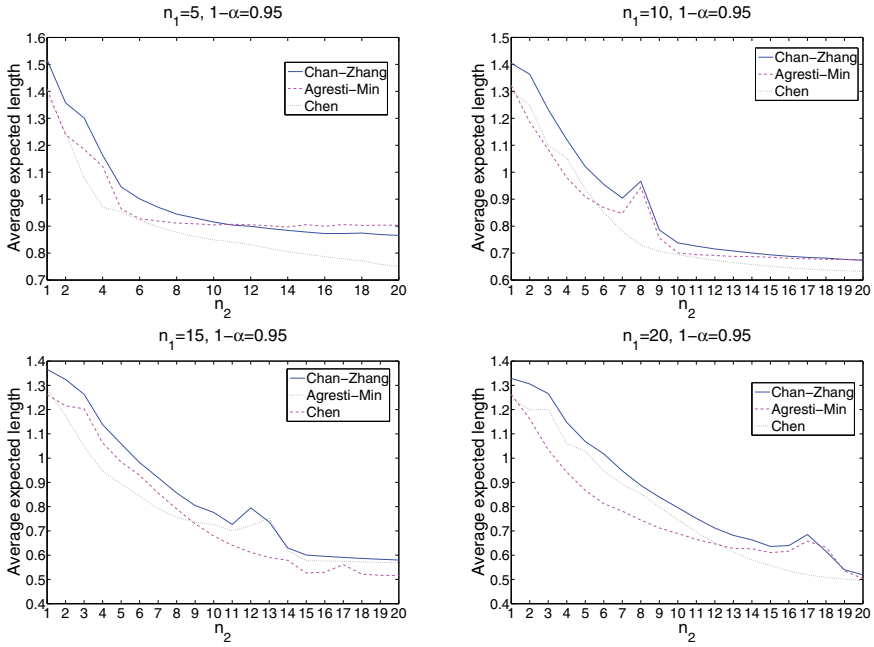


Figure 3: Average expected length of the 95 % confidence intervals for $n_1 = 5$, 10, 15 and 20

3.2 Comparsion of the Performances

To compare the performances of the confidence intervals the coverage probability has been computed. We consider ten small sample sizes (n_1, n_2) =(5, 5), (5, 10), (5,15), (5,20), (10, 15), (10, 20), (15, 15), (15, 20), (20, 20) and $\alpha = 0.05$. We use a fixed δ approach. We consider fixed $\delta = \pi_1 - \pi_2 = 0$, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. For each pair (n_1, n_2) the coverage probability and the expected length has been computed. The calculations were performed in Matlab.

Figure 1. shows the examples of the coverage probabilities of the 95 % confidence intervals as a function of π_1 for $(n_1, n_2) = (5, 15)$, (10, 10) and (10, 20) when $\delta = 0.1$, 0.3, 0.5 and 0.7.

The Chan-Zhang interval and the Agresti-Min interval are strict conservative methods. They guarantees the coverage probability above or equal to the nominal level. The Agresti-Min interval is less conservative. The coverage probability of the Chen interval fall below the nominal level, but is close to the nominal level.

The Chen interval is the shortest interval for most cases, the Chan-Zhang interval is the widest for most cases. In some cases the Agresti-Min interval become wider than the Chan-Zhang interval, but the Agresti-Min interval is generally shorter than the Chan-Zhang interval.

The average expected length of the Chan-Zhang interval is larger than others interval.

Figure 2. shows the expected lengths of the 95 % confidence intervals as a function of π_1 for $\delta = \pi_1 - \pi_2 = 0.1$, 0.3, 0.5, 0.7 and $(n_1, n_2) = (5, 15)$, (10, 10) and (10, 20). Figure 3. shows the average expected lengths of the 95 % confidence intervals for $n_1 = 5$, $n_1 = 10$, $n_1 = 15$ and $n_1 = 20$.

4 Concluding Remarks

The better confidence interval is such confidence interval, which coverage probability is close to the nominal level. The shorter interval and the smaller average expected length are preferred.

When the strict conservatism is a major criterion the Agresti-Min and the Chan-Zhang interval are a good choice.

The Chen quasi-exact interval is computationally simpler than the exact Chan-Zhang interval and the exact Agresti-Min interval. This interval is the reasonable altervative of the exact intervals in situation, when a strict conservatism is not required.

References

[1] 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.

[2] CHEN, X.: *A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases.* Statistics in Medicine 21, p. 943-956, 2002.

[3] MIETTINEN, O. S., NURMINEN, M.: *Comparative analysis of two rates.* Statistics in Medicine 4, p. 213-226, 1985.

[4] NEWCOMBE, R. G.: *Interval estimate for the difference between independent proportions: comparison of eleven methods.* In Statistics in Medicine, Vol. 17, p. 873-890, 1998.

[5] ZHOU, X.H., TSAO, M., QIN, G.: *New intervals for the difference between two independent binomial proportions.* Journal os Statistical Planning and Inference 123, p. 97-115, 2004.

[6] SANTNER, T. J., PRADHAN, V., SENCHAUDRI, P., MEHTA, C. R., TAMHANE, A.: *Small-sample comparsion of confidence intervals for the difference of two independent binomial proportions.* Computational Statistics and Data Analysis 51, p. 5791-5799, 2007.

[7] SANTNER, T. J., SNELL M. K.: *Small-sample confidence intervals for* $p_1 - p_2$ *and* p_1/p_2 *in* 2×2 *contingency tables.* Journal of the American Statistical Association 75, p. 336-394, 1980.

Current address

Ivana Pobočíková (Mgr.),
Univerzitná 1, Department of Applied Mathematics, Faculty of Mechanical Engineering, University of Žilina, 010 26 Žilina, ivana.pobocikova@fstroj.uniza.sk.

Journal of Applied Mathematics

volume IV (2011), number III

ESTIMATING INFORMATION VALUE FOR CREDIT SCORING MODELS

ŘEZÁČ Martin, (CZ)

Abstract. Assessing the predictive power of credit scoring models is an important question to financial institutions. Because it is impossible to use a scoring model effectively without knowing how good it is, quality indexes like Gini, Kolmogorov-Smirnov statisic and Information value are used to adress this problem. The paper deals with the Information value, which enjoy high popularity in the industry. Commonly it is computed by discretisation of data into intervals using deciles. One constraint is required to be met in this case. Number of cases have to be nonzero for all intervals. If this constraint is not fulfilled there are some issues to solve for preserving reasonable results. To avoid these computational issues, I proposed an alternative algorithm for estimating the Information value, named the empirical estimate with supervised interval selection. This advanced estimate is based on requirement to have at least k, where k is a positive integer, observations of scores of both good and bad clients in each considered interval. Simulation study with normally distributed scores shows high dependency on choise of the parameter k. If we choose too small value, we get overestimated value of the Information value, and vice versa. The quality of the estimate was assessed using MSE. According to this criteria, adjusted square root of number of bad clients seems to be a reasonable compromise.

Keywords: Credit scoring, Quality indexes, Information value, Empirical estimate, Normaly distributed scores

Mathematics Subject Classification: Primary 62G05, 62P05; Secondary 65C60.

1 Introduction

Credit scoring is a set of statistical techniques used to determine whether to extend credit (and if so, how much) to a borrower. When performing credit scoring, a creditor will analyze a relevant data sample to see what factors have the most effect on credit worthiness. Once these factors and their importances are known, a model is developed to calculate a credit score for new applicants.

Methodology of credit scoring models and some measures of their quality were discussed in works like Hand and Henley (1997) or Crook at al. (2007) and books like Anderson (2007), Siddiqi

(2006), Thomas et al. (2002) and Thomas (2009). Further remarks connected to credit scoring issues can be found there as well.

Once a scoring model is available, it is natural to ask how good it is. To measure the partial processes of a financial institution, especially their components like scoring models or other predictive models, it is possible to use quantitative indexes such as Gini index, K-S statisic, Lift, Information value and so forth. They can be used for comparison of several developed models at the moment of development. It is possible to use them for monitoring the quality of models after the deployment into real business as well. See Wilkie (2004) or Siddiqi (2006) for more details.

The paper deals primarily with the Information value. Commonly it is computed by discretisation of data into bins using deciles with requirement on the nonzero number of cases for all bins. As an alternative method to the empirical estimates one can use the kernel smoothing theory, which allows to estimate unknown densities and consequently, using some numerical method for integration, to estimate value of the Information value. See Koláček and Řezáč (2010) for more details.

The main objective of this paper is a description of the empirical estimate with supervised interval selection. This advanced estimate is based on requirement to have at least k, where k is a positive integer, observations of scores of both good and bad clients in each considered interval. Simulation study with normally distributed scores shows high dependency on choise of the parameter k. If we choose too small value, we get overestimated value of the Information value, and vice versa. The quality of the estimate is assessed using MSE. According to this criteria, I proposed a rule for choice of k, which seems to be a reasonable compromise.

2 Basic notations

Consider the realization $s \in \mathbb{R}$ of random value S (score) is available for each client. Let D be the indicator of good and bad client

$$D = \begin{cases} 1, & client\ is\ good \\ 0, & client\ is\ bad \end{cases} \tag{1}$$

and let F_0, F_1 denote cumulative distribution functions of score of bad and good clients, i.e.

$$\begin{aligned} F_0(a) &= P(S \le a \mid D = 0), \\ F_1(a) &= P(S \le a \mid D = 1),\ a \in \mathbb{R}. \end{aligned} \tag{2}$$

Aassume F_0, F_1 and their corresponding densities f_0, f_1 are continuous on $\mathbb{R}$.

In practice, the empirical distribution functions are used

$$\begin{aligned} \hat{F}_0(a) &= \frac{1}{m}\sum_{i=1}^{N} I(s_i \le a \wedge D = 0) \\ \hat{F}_1(a) &= \frac{1}{n}\sum_{i=1}^{N} I(s_i \le a \wedge D = 1),\ a \in [L, H], \end{aligned} \tag{3}$$

where s_i is the score of i-th client, $n,\ m$ are the number of good and bad clients, respectively and $N = n + m$. L is the minimum value of given score, H is the maximum value. Finally, we denote $p_B = \frac{m}{N}$ the proportion of bad clients.

3 The Information value

Very popular quality index, which is based on densities of scores of good and bad clients, is the Information value (statisic) defined as

$$I_{val} = \int_{-\infty}^{\infty} f_{IV}(x)dx, \tag{4}$$

where

$$f_{IV}(x) = \big(f_1(x) - f_0(x)\big)\ln\left(\frac{f_1(x)}{f_0(x)}\right). \tag{5}$$

Note that the Information value is also called Divergence. See Wilkie (2004), Hand and Henley (1997) or Thomas (2009) for more details. The example of $f_{IV}(x)$ for 10% of bad clients with $f_0: N(0,1)$ and 90% of good clients with $f_1: N(4,2)$ is illustrated in Figure 1.

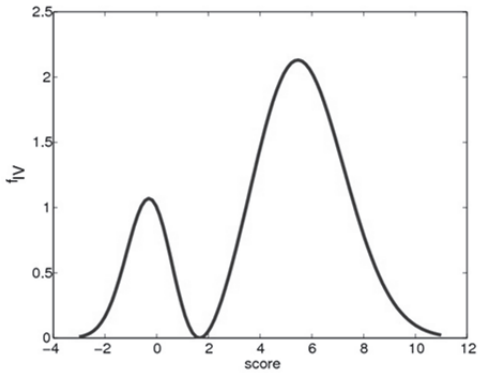


Figure 1: Contribution to Information value.

However, in practice, the procedure of computation of the Information value can be a little bit complicated. Firsty, we don't know the right form of densities f_0, f_1 generally and as the second, mostly we don't know how to compute the integral. I show some approaches to solve these computational problems.

3.1 Estimates for normally distributed data

In case of normally distributed data, we know everything what is needed. We just have to discriminate between two cases. Firstly, we consider that scores of good and bad clients have common variance. In this case we have

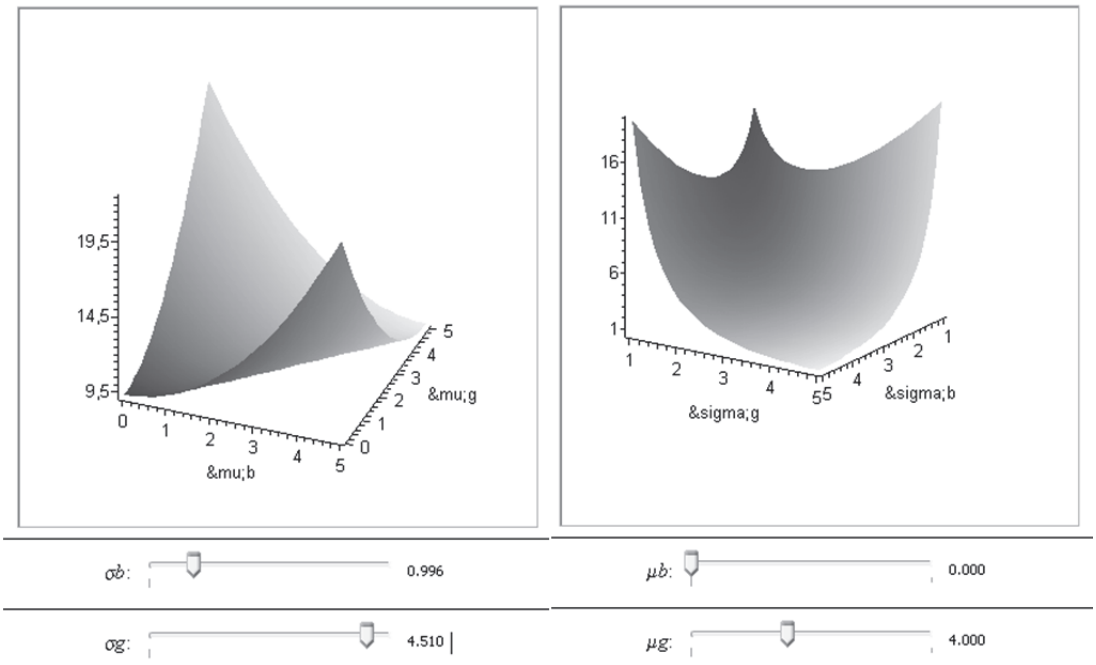
$$I_{val} = D^2\ , \tag{6}$$

where $D = \frac{\mu_g - \mu_b}{\sigma}$, μ_g and μ_b are expectations of scores of good and bad clients and σ is common standard deviation, see Wilkie (2004) for more details. When equality of variances is not considered, then in Řezáč (2009) one can find generalized form of I_{val} given by

$$I_{val} = (A+1)D^{*^2} + A - 1, \tag{7}$$

where $D^* = \frac{\mu_g - \mu_b}{\sqrt{\sigma_g^2 + \sigma_b^2}}$, $A = \frac{1}{2}\left(\frac{\sigma_g^2}{\sigma_b^2} + \frac{\sigma_b^2}{\sigma_g^2}\right)$, σ_g^2 and σ_b^2 are variances of scores of good and bad clients. The similar formula can be found in Thomas (2009). For a given data, estimation of I_{val} is done by replacing theoretical means and variances in (6) or (7) by their appropriate empirical expressions.

To explore behaviour of the expression (7) it is possible to use tools offered by system Maple. See Hřebíček and Řezáč (2008) for more details. An example of usage of the Exploration Assistant is given in Figure 2. We can see a quadratic dependence on difference of means in part (a). Furthermore, it is clear from (7) that I_{val} takes quite high values when both variances are approximately equal and smaller or equal to 1, and that it grows to infinity if ratio of the variances tends to infinity or is nearby zero. These properties of I_{val} are illustrated in Figure 2, part (b).



(a) **(b)**

Figure 2: Maple Exploration Assistant for 3D-plot of I_{val}. Dependence of I_{val} (a) on μ_g and μ_b for fixed σ_g^2 and σ_b^2, (b) on σ_g^2 and σ_b^2 for fixed μ_g and μ_b.

3.2 Empirical estimates

The main idea of this chapter is to replace unknown densities by their empirical estimates. Let's have m score values $s_{0_i}, i = 1, \dots, m$ for bad clients and n score values $s_{1_j}, j = 1, \dots, n$ for good clients and denote L (resp. H) as the minimum (resp. maximum) of all values. Let's divide the interval $[L, H]$ up to r subintervals $[q_0, q_1], (q_1, q_2], \dots, (q_{r-1}, q_r]$, where $q_0 = L - 1, q_r = H$ and $q_i, i = 1, \dots, r-1$ are appropriate quantiles of score of all clients. Set

$$\begin{aligned} n_{0_j} &= \sum_{i=1}^{m} I\left(s_{0_i} \in \left(q_{j-1}, q_j\right]\right) \\ n_{1_j} &= \sum_{i=1}^{n} I\left(s_{1_i} \in \left(q_{j-1}, q_j\right]\right), \quad j = 1, \dots, r \end{aligned} \tag{8}$$

observed counts of bad or good clients in each interval. Denote $\hat{f}_{IV}(j)$ the contribution to the information value on j[th] interval, calculated by

$$\hat{f}_{IV}(j) = \left(\frac{n_{1_j}}{n} - \frac{n_{0_j}}{m}\right) \ln\left(\frac{n_{1_j} m}{n_{0_j} n}\right), \quad j = 1, \dots, r. \tag{9}$$

Then the empirical information value is given by

$$\hat{I}_{val} = \sum_{j=1}^{r} \hat{f}_{IV}(j). \tag{10}$$

However in practice, there could occur computational problems. The Information value index becomes infinite in cases when some of n_{0_j} or n_{1_j} are equal to 0. When this arises there are numerous practical procedures for preserving finite results. For example one can replace the zero entry of numbers of goods or bads by a minimum constant of say 0.0001. Choosing of the number of bins is also very important. In the literature and also in many applications in credit scoring, the value $r = 10$ is preferred.

3.3 Empirical estimates with supervised interval selection

This approach follows ideas in the previous chapter. Estimation of information value is given again by formulas (8) to (10). The main difference lies in construction of the intervals. Because we want to avoid zero values of n_{0_j} and n_{1_j}, I simply looked for such selection of intervals, which provides such values n_{0_j} and n_{1_j}, which are all positive. This will lead to situation when all fractions and logarithms in (9) are defined and finite.

More generally, I propose to require to have at least k, where k is a positive integer, observations of socres of both good and bad client in each interval, i.e. $n_{0_j} \geq k$ and $n_{1_j} \geq k$ for j= $1, \dots, r$. Set

$$\begin{gathered} q_0 = \mathrm{L} - 1 \\ q_{\mathrm{i}} = \widehat{\mathrm{F}_0^{-1}}\left(\frac{\mathrm{k}\cdot\mathrm{i}}{\mathrm{m}}\right), i = 1, \dots, \left\lfloor \frac{m}{k} \right\rfloor \\ q_{\left\lfloor \frac{m}{k} \right\rfloor + 1} = \mathrm{H}, \end{gathered} \tag{11}$$

where $\widehat{\mathrm{F}_0^{-1}}(\cdot)$ is the empirical quantile function appropriate to the empirical cumulative distribution function of scores of bad clients. $\lfloor x \rfloor$ means lower integer part of number x. Usage of quantile function of scores of bad clients is motivated by the assumption, that number of bad clients is less than number of good clients, which is quite natural assumption. If m is not divisible by k, it is necessary to adjust our intervals, because we obtain number of scores of bad clients in the last interval, which is less than k. In this case, we have to merge the last two intervals. This will lead to situation, when it holds $n_{0_j} \geq k$ for all computed intervals of scores.

Furthermore we need to ensure, that the number of scores of good clients is as required in each interval. To do so, we compute n_{1_j} for all actual intervals. If we obtain $n_{1_j} < k$ for j[th] interval, we merge this interval with its neighbor on the right side. This is equivalent with the removal of $q_{\mathrm{j}+1}$ from the sequence of borders of the intervals. This can be done for all intervals except the last one. If we have $n_{1_j} < k$ for the last interval, than we have to merge it with its neighbor on the left side, i.e. we merge the last two intervals. However, this situation is not very probable. If we have a

reasonable scoring model, we can assume that good clients have higher scores than bad clients. It means that we can expect that the number of scores of good clients is higher than number of scores of bad clients in the last interval. Due to construction of the intervals, number of scores of bad clients in the last interval is greater than k. Thus, it is natural to expect that number of scores of good clients in the last interval is also greater than k. After all, we obtain $n_{0_j} \geq k$ and $n_{1_j} \geq k$ for all created intervals.

Very important is the choice of k. If we choose too small value, we get overestimated value of the Information value, and vice versa. As a reasonable compromise seems to be adjusted square root of number of bad clients given by

$$k = \lceil \sqrt{m} \rceil, \tag{12}$$

where $\lceil x \rceil$ means upper integer part of number x.

Denote $\hat{\hat{f}}_{IV}(j)$ the contribution to the information value on j[th] interval, calculated by

$$\hat{\hat{f}}_{IV}(j) = \left(\frac{n_{1_j}}{n} - \frac{n_{0_j}}{m}\right) \ln\left(\frac{n_{1_j} m}{n_{0_j} n}\right), \quad j = 1, \dots, r. \tag{13}$$

where n_{1_j} and n_{0_j} correspond to observed counts of good and bad clients in intervals created according to the procedure described in this chapter. The empirical information value with supervised interval selection is now given by

$$\hat{\hat{I}}_{val} = \sum_{j=1}^{r} \hat{\hat{f}}_{IV}(j). \tag{14}$$

4 Simulation results

It is clear, and it is easy to show that $\hat{\hat{I}}_{val}$ outperformes $\hat{I}_{val}$. However, this chapter is focused on properties of $\hat{\hat{I}}_{val}$ depending on choice of parameter *k* and depending on proportion of bad clients p_B and difference of means of scores of bad and good clients $\mu_g - \mu_b$. Consider 10000 clients, $100 \cdot p_B\%$ of bad clients with $f_0: N(\mu_b, 1)$ and $100 \cdot (1 - p_b)\%$ of good clients with $f_1: N(\mu_g, 1)$. Set $\mu_b = 0$ and consider $\mu_b = 0.5, 1$ and 1.5, $p_B = 0.02, 0.05, 0.1$ and 0.2. The case $\mu_g - \mu_b = 0.5$, i.e. $I_{val} = 0.25$ in our settings, represents weak, $\mu_g - \mu_b = 1$ means high and $\mu_g - \mu_b = 1.5$ very high performance of given scoring model. 2% bad rate ($p_B = 0.02$) represents low risk portfolio, e.g. mortgages (before current crises). 20% bad represents very high risk portfolio, e.g. subprime cash loans.

Appropriate data sets for simulation was randomly generated 1000 times. Quality of $\hat{\hat{I}}_{val}$ was assessed using mean squared error given by

$$MSE = E\left(\left(\hat{\hat{I}}_{val} - I_{val}\right)^2\right). \tag{15}$$

Given this measure, denote

$$k_{MSE}=\underset{k}{argmin}\, MSE. \tag{16}$$

Following Table 1 consists of k_{MSE} for all considered values of p_B and $\mu_g - \mu_b$. Proposed values of k, $k = \lceil\sqrt{m}\rceil$, are presented in the last row of the table.

Table 1: k_{MSE} depending on p_B and $\mu_g - \mu_b$.

k_{MSE}		p_B			
		0.02	0.05	0.1	0.2
$\mu_g - \mu_b$	0.5	29	42	62	84
	1	12	18	23	32
	1.5	6	9	8	9
$k = \lceil\sqrt{m}\rceil$		15	23	32	45

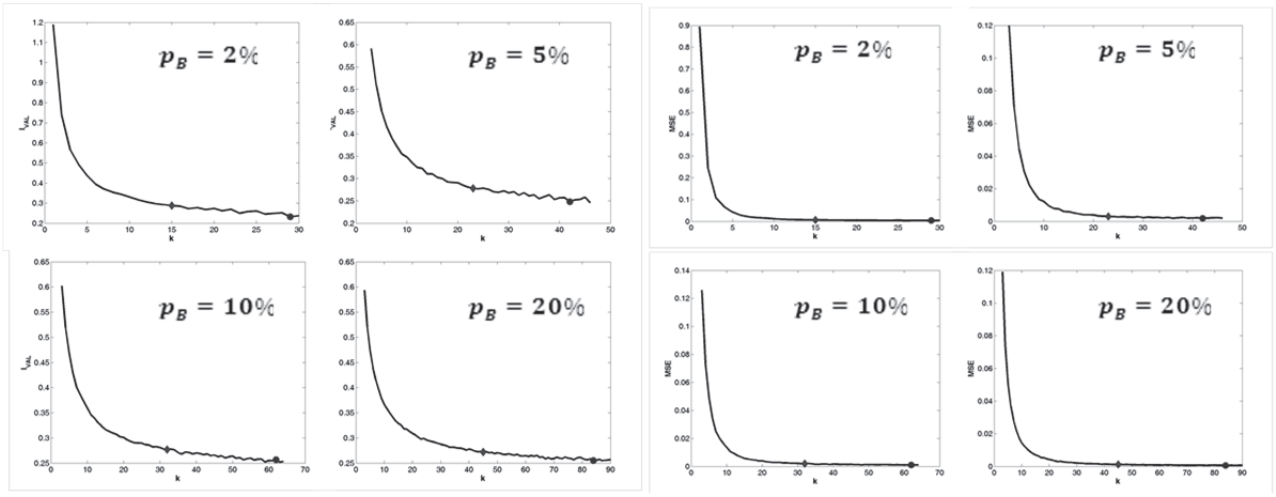
We can see that k_{MSE} is increasing according to p_B. This is maybe somewhat surprising, but it is quite natural. The increasing p_B means increasing number of bad clients, becasuse the number of all clients was fixed to 10000. If we have enough of bad clients, then too small k leads to too many bins and consequently to overestimated results. But what is surpricing, it is the dependence on $\mu_g - \mu_b$. While for weak models it is optimal to take very high number of observation in each bin, the contrary holds for high perfoming models. Overall, $k = \lceil\sqrt{m}\rceil$ seems to be a reasonbale comprimise.

For completeness, Table 2 consists of average numbers of bins for all considered values of p_B and $\mu_g - \mu_b$. We can see that they took values from 8 to 127.

Table 2: Average number of bins depending on p_B and $\mu_g - \mu_b$.

avg. # of bins		p_B			
		0.02	0.05	0.1	0.2
$\mu_g - \mu_b$	0.5	8,00	13,00	18,00	24,90
	1	18,00	28,80	42,76	51,88
	1.5	33,62	50,20	95,96	127,67

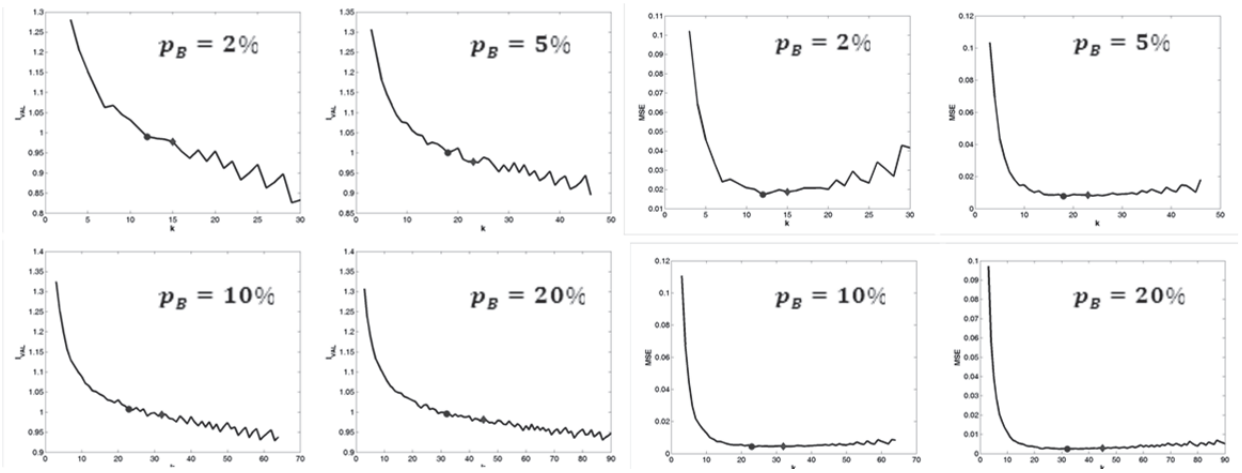
The dependece of $\hat{\hat{I}}_{val}$ on k is illustrated in Figure 3 to 5. The highlighted circles correspond to values of k, where minimal value of the *MSE* is obtained. The diamonds correspond to values of k given by (12).



(a) **(b)**

Figure 3: Dependence of (a) $\hat{\hat{I}}_{val}$ and (b) MSE on k, 100000 clients, $\mu_g - \mu_b = 0.5$.

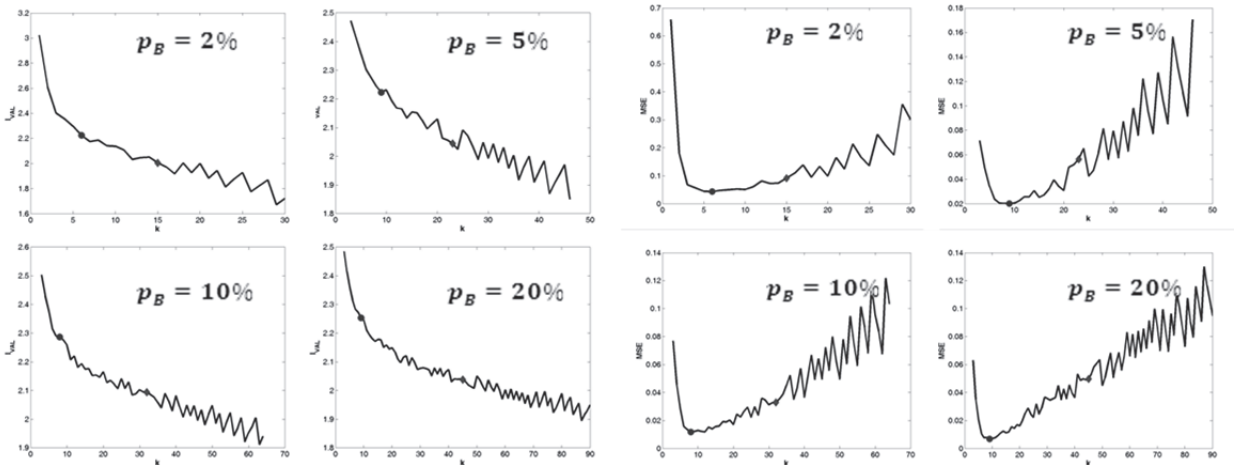
We can see that $\hat{\hat{I}}_{val}$ is decreasing when k is increasing. In case of $\mu_g - \mu_b = 0.5$, speed of this decrease is very high for small values of k, while it is nearly negligible for values of k higher than some critical value. The similar holds for *MSE*.



(a) **(b)**

Firuge 4: Dependence of (a) $\hat{\hat{I}}_{val}$ and (b) MSE on k, 100000 clients, $\mu_g - \mu_b = 1$.

When $\mu_g - \mu_b = 1$, the speed of the decrease is lower compared to the previous case. Furthermore *MSE* is not so flat, especially for $p_B = 2\%$. But what is interesting and important here, our choice of k is nearly optimal according to *MSE*. Moreover, it is valid for all considered values of p_B.



(a) **(b)**

Figure 5: Dependence of (a) $\hat{\hat{I}}_{val}$ and (b) MSE on k, 100000 clients, $\mu_g - \mu_b = 1.5$.

Tha last considered difference of means of scores of good and bad clients was $\mu_g - \mu_b = 1.5$. In this case, the speed of the decrease of $\hat{\hat{I}}_{val}$ is the lowest compared to the previous two cases. The novelty, relative to the previous two cases, is the shape of *MSE*. Especially for the highest considered value of proportion of bad clients, i.e. $p_B = 20\%$, we can see that *MSE* has really sharp minimum.

Overall, Figure 3 and Figure 4 show that curves of *MSE* are quite flat nearby its minimum . It means that a small deviation of k from k_{MSE} cause a small change in *MSE*. On the other hand Figure 5 shows the strong dependence on choice of k.

5 Conclusions

I focused on the Information value and described difficulties of its estimation. The most popular method is the empirical estimator using deciles of given score. But it can lead to infinite values of I_{val} and so a remedy is necessary. To avoid these difficulties I proposed the adjustment for the empirical estimate, called the empirical estimate with supervised interval selection. It is based on the assumption that we have at least some positive number of observed scores in each interval. This directly leads to situation when all fractions and all logarithms are defined and finite. Consequently, I_{val} is defined and finite.

The simulation study was focused on properties of $\hat{\hat{I}}_{val}$ depending on choice of parameter k and depending on proportion of bad clients and difference of means of scores of bad and good clients. Quality of $\hat{\hat{I}}_{val}$ was assessed using mean squared error, which is easy to compute for normally distributed scores. Moreover, the optimal value of k_{MSE} was computed.

It was shown that k_{MSE} was increasing according to p_B. This is maybe somewhat surprising, but it is quite natural. The increasing p_B means increasing number of bad clients, becasuse the number of all clients was fixed in our case. If we have enough of bad clients, then too small k leads to too many bins and consequently to overestimated results. But what was surpricing, it was the dependence on $\mu_g - \mu_b$. While for weak models it is optimal to take very high number of observation in each bin, the contrary holds for high perfoming models. Overall, $k = \lceil \sqrt{m} \rceil$ seems to be a reasonbale comprimise.

On the other hand, the obtained results open additional possibilities for research. Especially, it seems that inclusion of $\mu_g - \mu_b$, represented by appropriate estimates, to the rule of choise of k could lead to significantly better estimates of I_{val} when using the proposed empirical estimate with supervised interval selection.

References

[1] ANDERSON, R.: *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford : Oxford University Press, 2007.

[2] CROOK, J.N., EDELMAN, D.B., THOMAS, L.C.: Recent developments in consumer credit risk assessment. *European Journal of Operational Research,* 183 (3), 1447-1465, 2007.

[3] HAND, D.J. and HENLEY, W.E.: Statistical Classification Methods in Consumer Credit Scoring: a review. *Journal of the Royal Statistical Society, Series A.* 160 (3), 523-541, 1997.

[4] HŘEBÍČEK, J., ŘEZÁČ, M.: Modelling with Maple and MapleSim In: *22nd European Conference on Modelling nad Simulation ECMS 2008 Proceedings,* Dudweiler, 60-66, 2008.

[5] KOLÁČEK, J., ŘEZÁČ, M.: Assessment of Scoring Models Using Information Value. In: *Compstat' 2010 proceedings*. Paris, 1191-1198, 2010.

[6] ŘEZÁČ, M.: Indexy kvality normálně rozložených skóre. *Forum Statisticum Slovacum*, Bratislava, 2009.

[7] SIDDIQI, N.: *Credit Risk Scorecards: developing and implementing intelligent credit scoring*. New Jersey: Wiley, 2006.

[8] TERRELL, G.R.: The Maximal Smoothing Principle in Density Estimation. *Journal of the American Statistical Association,* 85, 470-477, 1990.

[9] THOMAS, L.C.: *Consumer Credit Models: Pricing, Profit, and Portfolio*. Oxford: Oxford University Press, 2009.

[10] THOMAS, L.C., EDELMAN, D.B., CROOK, J.N.: *Credit Scoring and Its Applications*. Philadelphia: SIAM Monographs on Mathematical Modeling and Computation, 2002.

[11] WAND, M.P. and JONES, M.C.: *Kernel smoothing*. London: Chapman and Hall, 1995.

[12] WILKIE, A.D.: Measures for comparing scoring systems. In: Thomas, L.C., Edelman, D.B., Crook, J.N. (Eds.), *Readings in Credit Scoring*. Oxford: Oxford University Press, s. 51-62, 2004.

Current address

Martin Řezáč, Mgr., Ph.D.,

Department of Mathematics and Statistics,
Faculty of Science, Masaryk University,
611 37 Brno, Czech Republic, tel. +420 549 493 919,
e-mail: mrezac@math.muni.cz

Journal of Applied Mathematics

volume IV (2011), number III

ANALYSIS OF THE DEPENDENCE OF THE HOUSING CHARACTERISTICS ON THE HOUSEHOLD TYPE IN THE CZECH REPUBLIC

ŘEZANKOVÁ Hana, (CZ), LÖSTER Tomáš, (CZ)

Abstract. The paper focuses on the application of coefficients of asymmetric dependence for nominal and ordinal variables. Data obtained on the basis of the EU-SILC (European Union – Statistics on Income and Living Conditions) survey in 2008 were analyzed. Dependence is investigated for pairs of nominal variables (equipment of households with durables and household type) and for pairs of ordinal variables (number of rooms, total floor area, financial burden of housing costs, and households type where the categories form an ordinal scale). The highest values of coefficients were achieved in the dependence of equipment with PC on the household type from the viewpoint of the number of working persons and according to the status of head of household; higher values were also found in the dependence of equipment with a car on the basic household type, the number of rooms and the total floor area on the household type from the viewpoint of the number of working persons and according to education level.

Key words. categorical data analysis, contingency tables, nominal variables, ordinal variables

Mathematics Subject Classification: Primary 62H17, 62H20, 62P20; Secondary 91C99.

1 Introduction

One of the series of surveys carried out by the statistical offices of European countries is the survey of the statistics on income and living conditions. In 2003 a regulation of the European Parliament and the EU Council was passed, which regulates the common framework for this survey in European countries. In the Czech Republic the EU-SILC (European Union – Statistics on Income and Living Conditions) survey has been carried out since 2005. The aim is to gain an overview of the state and development of the social situation of the population.

The survey focuses mainly on the income distribution of individual types of households, on the manner, quality and financial burden of housing, equipment of households with durables and it also provides data on the working, material and health conditions of adults, see [1]. In this way information valuable for the creation and evaluation of the state social policy is obtained.

Through the given survey a great quantity of information is acquired. On the basis of them it is possible to investigate various associations and dependences among indicators and either test assumed ones or seek out new ones. Some analyses can be also realized on the basis of the tables published by the Czech Statistical Office on the Internet (analyses of contingency tables with conditional relative frequencies, see [2]). With regard to the fact that the studied indicators are of different types it is possible to apply a wide range of statistical methods.

In this contribution we are focusing on the analysis of the data obtained from the EU-SILC survey in 2008. The file purchased from the Czech Statistical Office contains data on 11,294 households. Our aim was to discover or verify various dependences concerning the characteristics of a dwelling and its equipment on the one hand and types of household according to various viewpoints on the other hand.

In some cases it can be expected where the dependence will be weaker and where stronger. With regard to the large size of the data file it is also possible to assume that in statistical tests the hypothesis of independence will be rejected. For this reason, in the further text we shall deal only with investigation of the intensity of dependence. The coefficients calculated on the basis of frequencies in a contingency table (or absolute values of these coefficients) usually have values from the interval from 0 to 1, however from the dependence aspect even values smaller than 0.5 are interesting.

Some indicators were recoded before the analysis. We were inspired by tables presented on the Internet, which in some cases are also created on the basis of recoded indicators – for instance in the indicator "number of working persons" it is useful to combine values of 3 or more into one category.

2 Methods Used for Analyses

From the nature of the type of defined problems it derives that this is an asymmetric dependence where the characteristics of the dwelling depend on the household type. If we consider a contingency table for two variables then we will suppose that the column variable is dependent. We will denote this variable with the letter Y and the explanatory row variable with the letter X.

The variable X influences statistically on Y if the statistical properties of the variable Y change with the changes in categories x_i. From the method called analysis of variance it derives that we can express the variability of the dependent variable as a sum of two components: the variability explained by the variable X (between-group variability) and the residual (unexplained) variability (within-group variability). In mathematical notation this association can be expressed as

$$var(Y) = var(Y, X) + var(Y|X), \tag{1}$$

where for the expressing of variability $var(.)$ it is necessary to use a measure suitable for the given type of dependent variable.

In the analysis of variance we can calculate the intensity of dependence as a ratio of between-group variability and total variability which has values in the interval from 0 to 1. This measure is usually called as R-square. It can be written as

$$S_{Y|X} = \frac{var(Y,X)}{var(Y)} = \frac{var(Y) - var(Y|X)}{var(Y)}. \tag{2}$$

If we calculate this ratio on the basis of a contingency table, then

$$S_{Y|X} = \frac{var(Y) - \sum_{i=1}^{R} p_{i+} var(Y|x_i)}{var(Y)}, \tag{3}$$

where p_{i+} is the row marginal frequency and R is the number of rows (i.e. of categories of the row variable X).

In practice measures are used for quantitative and nominal dependent variables. In the first case variance is used and the root of R-square is also given; in the second case it is possible to use one of the following measures for the variable Y (see also [6]):

a) *variation ratio V*, which we calculate according to the formula

$$V(Y) = 1 - p_{+\mathrm{Mo}}, \tag{4}$$

where $p_{+\mathrm{Mo}}$ is the relative frequency of the modal category of the column variable Y,

b) *nominal variance G* (Gini's coefficient, measure of mutability), see [3], calculated according to the formula

$$G(Y) = 1 - \sum_{j=1}^{C} p_{+j}^2, \tag{5}$$

where C is the number of columns (i.e. of categories of the column variable Y) and p_{+j} is the column marginal relative frequency,

c) *entropy H*, which is for all $p_{+j} \neq 0$ given by the formula

$$H(Y) = -\sum_{j=1}^{C} p_{+j} \ln p_{+j} \tag{6}$$

(in the case that $p_{+j} = 0$, the corresponding value for the given j is equal to zero).

By the application of the formula (4) we get *Goodman and Kruskal lambda*, see [4], i.e.

$$\lambda_{Y|X} = \frac{V(Y) - \sum_{i=1}^{R} p_{i+} V(Y|x_i)}{V(Y)} = \frac{\sum_{i=1}^{R} p_{i\mathrm{Mo}} - p_{+\mathrm{Mo}}}{1 - p_{+\mathrm{Mo}}}, \tag{7}$$

where $p_{i\mathrm{Mo}}$ is the relative frequency of the modal category in the ith row. This coefficient reflects only the change of the column category with the greatest frequency in individual rows with regard to the category with the highest marginal frequency. If in all rows the highest frequency is in the same category as is the modal category for the whole data set, then the coefficient is equal to zero and therefore the test for the zero of the coefficient is not carried out.

By the application of the formula (5) we get *Goodman and Kruskal tau*, see [4], i.e.

$$\tau_{Y|X} = \frac{G(Y) - \sum_{i=1}^{R} p_{i+} G(Y|x_i)}{G(Y)} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}}}{1 - \sum_{j=1}^{S} p_{+j}^2}. \tag{8}$$

By the application of the formula (6) we get the *uncertainty coefficient*, see [5], i.e.

$$\begin{aligned} U_{Y|X} &= \frac{H(Y) - \sum_{i=1}^{R} p_{i+} H(Y|x_i)}{H(Y)} = \\ &= \frac{-\sum_{i=1}^{R} p_{i+} \ln p_{i+} - \sum_{j=1}^{C} p_{+j} \ln p_{+j} + \sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij} \ln p_{ij}}{-\sum_{j=1}^{C} p_{+j} \ln p_{+j}}. \end{aligned} \tag{9}$$

In investigating the asymmetric dependence of two ordinal variables we use Somers' d with values from –1 to 1. For its calculation it is necessary to know the number of concordant, discordant and tied pairs. If in a pair of objects for one object the values in both variables are smaller (or larger) than for the second object, then we denote such a pair as *concordant*. If in one variable the value is smaller and in the other variable bigger, then this is a *discordant* pair. In other cases (the value in one variable or the values in both variables are the same) we speak of *tied* pairs.

For the simplification of notations of formulae the following symbols are used:

Γ – the number of concordant pairs,

Δ – the number of discordant pairs,

T_Y – the number of pairs, which contain the same value of the variable Y, but a different value for X.

Mathematically these numbers can be expressed according to the following formulae:

$$\Gamma = \sum_{i=2}^{R}\sum_{j=2}^{C}\left(n_{ij} \sum_{h<i}\sum_{k<j} n_{hk} \right), \quad \Delta = \sum_{i=2}^{R}\sum_{j=1}^{C-1}\left(n_{ij} \sum_{h<i}\sum_{k>j} n_{hk} \right), \quad T_Y = \sum_{i=2}^{R}\sum_{j=1}^{C}\left(n_{ij} \sum_{h<i}\sum_{k=j} n_{hk} \right),$$

where n_{ij} are associated absolute frequencies in the contingency table. *Somers' d* is calculated according to the formula

$$d_{Y|X} = \frac{\Gamma - \Delta}{\Gamma + \Delta + T_Y}. \tag{10}$$

For two dichotomous variables a contingency table has four cells. It is possible to apply the coefficients mentioned above. Moreover, the odds ratio can be also used. If we denote the absolute frequencies in a table with letters *a*, *b*, *c* and *d*, then the odds ratio can be expressed as

$$\psi = \frac{ad}{bc}. \tag{11}$$

3 Application to Living Condition Survey

We investigated the asymmetric dependences in a series of indicators. With regard to the large size of the set, in the statistical tests the hypothesis of independence was rejected in all cases (in accordance with the expectation), but the intensity of dependence is generally low. In this part the results are given for some analyses in which the values of at least some coefficients are higher than 0.2, and eventually for comparison also the results of some further analyses.

We applied measures of dependence for two nominal variables (equipment of households with durables and household type) and for two ordinal variables (number of rooms, total floor area, financial burden of housing costs, and households type where the categories form an ordinal scale).

In investigating the dependence of the equipment of the household on the household type we focused on durables, i.e. washing machine, color TV, PC, phone (fixed, mobile) and car. The indicator has the categories 1 (household has own), 2 (household does not have – cannot afford to buy) and 3 (household has not for other reasons/does not want). Because this is a nominal variable, for the expression of the intensity of dependence we used the coefficients lambda, tau and the uncertainty coefficient, see formulae (7), (8), and (9).

First we selected the household type according to the number of working persons. In the indicator "number of working persons" we combined values of 3 and more in a single category, which means that the recoded variable has categories from 0 (no working person) to 3 (3 or more working persons). The values of the coefficients appropriate for this type of asymmetric dependence are given in Table 1.

Table 1. Dependence of the household equipment on the number of working persons

Durables	lambda	tau	U
washing machine	0	0.018	0.076
color TV	0	0.002	0.035
PC	0.478	0.290	0.218
phone (fixed, mobile)	0	0.030	0.113
car	0.144	0.159	0.130

In the case of the washing machine, color TV and phone, for all households the highest frequency is for category 1 (subsequently: 96.5 %, 98.8 % and 96.3 %) and also for the individual types of household according to working persons the highest frequency is always for this category. For this reason coefficient lambda equals zero and the values of coefficient tau and the uncertainty coefficient are also very low.

As far as concerns ownership of PC, then for the whole data set the greatest frequency is for category 1 (50%), followed by category 3 (42%). The first category is most frequent for households with at least one working person. For households without working persons the highest frequency is in category 3 (76%). Of all the items of durables the strongest dependence was found for the dependence of equipment with PC on the household type according to working persons.

A weaker dependence was found for ownership of a car. For the whole set the greatest frequency is for category 1 (61.4%), followed by category 3 (27%). The first category is most frequent for households with at least one working person. For households without working persons the highest frequency is in category 3 (50%).

We achieved similar results in investigating the equipment with durables in relation to the household type according to the status of head of household. Seven categories of household are

distinguished, these being 1 (employee with lower education), 2 (self-employed), 3 (employee with higher education), 6 (pensioner with working persons), 7 (pensioner without working persons), 8 (unemployed) and 9 (other household). For the reason of the insignificant dependence of equipment with washing machine, color TV and phone is (with regard to the overall majority of the first category throughout the data set), we will focus only on ownership of PC and a car. The resulting values of the appropriate coefficients are given in Table 2.

Table 2. Dependence of the household equipment on the status of head of household

Durables	lambda	tau	U
PC	0.487	0.322	0.249
car	0.153	0.156	0.129

In the case of equipment with PC the first category is not the most frequent only for households headed by a pensioner without working persons. In this case the most frequent category is 3 (79.6%). In the case of ownership of a car the situation is similar. In this case in households headed by a pensioner without working persons the relative row frequencies in category 3 is 52%.

Similar or higher values of coefficients were further found in investigating the dependence of equipment with PC and a car on the household type according to working activity and intensity and according to the classification of the EU and OECD.

A stronger dependence of equipment with a car than with PC was found on the basic household type with categories 1 (two-parent nuclear family), 2 (two-parent family with other relatives), 3 (lone-parent family with children), 4 (lone-parent family with other relatives with children), 7 (non-family household), 8 (individual – male) and 9 (individual – female). The resulting values of the appropriate coefficients are given in Table 3.

Table 3. Dependence of the household equipment on the basic household type

Durables	lambda	tau	U
PC	0.285	0.154	0.123
car	0.286	0.239	0.194

In the case of the type 7 household in equipment with PC the same frequency was found for categories 1 and 3 and the highest frequency for the 3rd category was in households of type 8 (54%) and 9 (79%). For the car indicator the highest frequency for the 3rd category was found only in the type 9 households (72%).

Lower values of coefficients were obtained in the case of dependence on the household type according to education level with the categories 1 (low level), 2 (medium level – at least one partner with secondary education) and 3 (high level – at least one partner with university education), see Table 4.

Table 4. Dependence of the household equipment on the household type according to education level

Durables	lambda	tau	U
PC	0.145	0.097	0.083
car	0.114	0.074	0.059

Both in the case of the PC and of the car the highest frequency for the 3rd category was in the household type with a low level of education (76.2% and 60.1%).

We also investigated the dependence of the number of rooms on the ordinal indicators concerning the household. We recoded the indicator for the number of rooms to the values 1, 2, 3 and 4 (meaning 4 or more). In investigating the dependence on the number of working persons it was found that in households with no or with one working person there were mainly 3-room apartment, whereas for households with two, three and more working persons mainly had four or more rooms. The value of Somers' d in this case is 0.259, which, although it is a lower dependence, is one that is worthy of notice. In investigating the dependence of the number of rooms on the household type from the viewpoint of education level it was found that the households with the lowest education level most often have two rooms, households with medium education have three rooms and the households with the highest level of education then have 4 or more rooms. In this case the value of Somers' d is 0.206.

In investigating the dependence of the total floor area, very similar results were obtained according to the above-mentioned ordinal indicators concerning households. The indicator of floor area was recoded for these purposes to the values 1 (40 m^2 or less), 2 (40–60 m^2), 3 (60–80 m^2) and 4 (more than 80 m^2). In investigating the dependence of this indicator on the number of working persons it was found that in households with no working person the prevalent floor area is 40–60 m^2, in households with one working person the usual floor area is 60–80 m^2 and in households with three, four or more such members the floor area is greater. The value of Somers' d in this case is 0.252, which, although it is again a weaker dependence, is one that is worthy of notice. In investigating the dependence of the floor area on the household type from the viewpoint of education level it was found that the households with the lowest level of education most often have 40–60 m^2, households with medium education have 60–80 m^2 and the households with the highest education level then have more than 80 m^2. In this case the value of Somers' d is 0.209.

An example of another ordinal variable is the indicator "housing costs in terms of financial burden" with categories 1 (large burden), 2 (a certain burden) and 3 (no burden). Overall category 2 is prevalent, being the most frequent category also in the individual categories of household. The value of Somers' d in the case of the dependence on households according to the number of working persons is 0.077 and in the case of the dependence on households according to education level it is 0.17.

For the purpose of comparison, we also investigated the dependence of quantitative continuous variables (original value of floor area in m^2, housing costs and their individual items) on the household type with the use of R-square, see formula (2) or its root (coefficient eta). The values of coefficient eta indicate a weaker dependence. The highest values were found in the dependence on the household type according to the number of working persons, these being for total housing costs (0.336), followed by the cost of electricity (0.289), floor area (0.282) and the cost of water supply (0.274). There are somewhat lower costs when investigating these indicators for the household type according to education level: for total housing costs 0.221, followed by costs of other utilities with 0.19 and floor area with 0.176.

We obtained some interesting results by using the odds ratio, concerning dependences on the sex of the household head. In the case of independence the value of the odds ratio is one. This measure expresses a chance for a certain event (a category of a column variable) in dependence on categories of a row variable. If a man is a head of household, then the household has more than 2.5 times greater chance to go for an annual one-week holiday in comparison with a household with a woman as a head. In a case of unexpected outlay of CZK 7500 this chance is more than 3 times greater in comparison of men and women as heads of household.

4 Conclusion

Through analysis of the data from the EU-SILC survey in 2008 interesting dependences were found concerning the dependence of equipment with PC on the household type from the viewpoint of the number of working persons and according to status of head of household, ownership of a car on the basic household type, further the number of rooms and the total floor area on the household type from the viewpoint of the number of working persons and according to education level.

In our further research we plan to focus on modeling the dependences of categorical indicators with the use of logistic regression, see [7] for modeling the dependences on the basis of the survey in 2006.

Acknowledgement

This paper was prepared with the support for the long-term conceptual development of science and research at the Faculty of Informatics and Statistics of the University of Economics, Prague.

References

[1.] BARTOŠOVÁ, J.: *Analysis and modelling of financial power of Czech households*. In APLIMAT 2009 (8th International Conference on Applied Mathematics), pp. 717-722, Slovak University of Technology, Bratislava, 2009.

[2.] ČAPKOVÁ, K.: *Zkoumání závislosti charakteristik a vybavení bytu na typu domácnosti*. Diplomová práce, VŠE, Praha, 2010.

[3.] GINI, C. W.: *Variability and Mutability. Contribution to the study of statistical distributions and relations*. Studi Economico-Giuridici della R. Universita de Cagliari. 1912. Reviewed in: Light, R. J., Margolin, B. H.: An Analysis of Variance for Categorical Data. In J. American Statistical Association, Vol. 66, pp. 534-544, 1971.

[4.] GOODMAN, L. A., Kruskal, W. H.: *Measures of association for crossclassification*. In Journal of the American Statistical Association, Vol. 49, pp. 732-764, 1954.

[5.] PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A., VETTERLING, W. T.: *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press. Cambridge, p. 634, 1988.

[6.] ŘEZANKOVÁ, H.: *Analýza dat z dotazníkových šetření*. Professional Publishing, Praha, 2010.

[7.] STANKOVIČOVÁ, I. PASTOREK, L.: *Analýza činiteľov monetárnej chudoby v českých a slovenských domácnostiach v roku 2006.* In Finanční potenciál domácností 2009 [CD-ROM]. VŠE, Praha, 2009.

Current address

Hana Řezanková, prof. Ing., CSc.
University of Economics, Prague, nám. W. Churchilla 4, Prague, 13067, Czech Republic,
tel.: +420 224 095 483, e-mail: hana.rezankova@vse.cz

Tomáš Löster, Ing.
University of Economics, Prague, nám. W. Churchilla 4, Prague, 13067, Czech Republic,
tel. +420 224 095 484, e-mail: tomas.loster@vse.cz

Journal of Applied Mathematics

volume IV (2011), number III

ON COMPARISON OF UNIVARIATE FORECASTING METHODS: THE CASE OF LATVIAN RESIDENTIAL PROPERTY PRICES

SINENKO Nadezhda, (LV) VALEINIS Janis, (LV)

Abstract. This paper investigates the forecasting ability of different univariate forecasting techniques (local regressions, unobserved component model), compared with the standard ARIMA approach. A forecasting exercise is carried out with each method, using monthly price time series on residential property prices in Latvia. The accuracy of the different methods is assessed by comparing several measures of forecasting performance based on the out-of-sample predictions for various horizons.

Key words and phrases. Unobserved components model; ARIMA models; Forecasting comparison; local linear regression.

Mathematics Subject Classification. Primary 91B84, 62M10; Secondary 62G08.

1 Introduction

The presence of a turning point at the end of a sample introduces a degree of uncertainty in econometric forecasting. Deterministic trend-based forecasting strategies are not relevant in this case. Cointegration relationship also may seem to be broken thus not providing a solid basis for making projections of future paths of economic variables. Uncertainty about the future behavior of exogenous variables due to turn in economic cycle makes it difficult to use traditional multivariate forecasting methods. In this case univariate forecasting techniques may serve as an additional work aid. The purpose of the article is to investigate the forecasting performance of several univariate modelling methods (unobserved component model, local regressions and traditional ARIMA models) applied to Latvian residential property prices.

The development of real estate prices serves as an important economic indicator, closely connected to economic and credit cycles. As the behavior of the property prices influences

the performance of the whole financial system, the forecasting of its future developments is an important task. Over the last decade Latvian residential property prices experienced dramatic changes. Joining the European Union in 2004 provided access to cheap credits. The economic boom that followed the EU accession led to growing wages and therefore the demand for real estate and affordability of the loans for house purchase increased ([1]). Credit and property prices grew in mutually reinforcing manner, producing a speculative real estate price bubble, which burst in April 2007. The price drop on the real estate market coincided with a severe downturn in economic activity in Latvia, which was reinforced by the world-wide financial and economic crises. During 2007 - 2009 Latvian residential property prices fell by more than 70%. After hitting the bottom in the summer of 2009, the prices exhibited moderate growth. The idea of the exercise appeared at the end of 2009, when future prospects of the developments of the property prices were highly uncertain and the deep recession made the forecasting of the fundamental determinants of property prices comlicated for relatively long time horizons. Different classes of univariate models were fitted to residential property price time series on the estimation sample from January 1999 to December 2009 and 12 months ahead forecasts were produced. Now, when 10 of 12 actual values of property prices time series for 2010 are already known, the precision of forecasts could be accessed.

The paper is organised as follows: Section 2 describes the theoretical UC model based on simple Integrated Random Walk model for the trend and results, obtained employing this model, including some modifications. Section 3 and Section 4 describe the nonparametric regression methods and benchmark ARIMA class models, respectively. Section 5 analyses the predictive performance of our models for Latvian residential property prices. Section 6 concludes.

2 Forecasts based on structural time series model

One of the popular ways of modelling time series is a structural time series model, which is set in terms of components having a direct economic interpretation. In the most common form additive structural model has the folowing form

$$y_t = T_t + C_t + S_t + \epsilon_t, \quad t = 1, \ldots, T, \tag{1}$$

where T_t - trend, C_t - cyclical, S_t - seasonal, ϵ_t - irregular component. Typically the irregular component is assumed to be white noise with zero mean and variance σ_ϵ^2. In some cases (like in [3]) the cyclical component is dropped and cyclical movements are incorporated into trend component. The new framework elaborated for structural models by Harvey (see e.g. [4], [5]) made the models more flexible, in particular, by letting the level and slope parameters of trend to change over time. Harvey and Jaeger in [5] proposed to fit trend as

$$T_t = T_{t-1} + D_{t-1} + \zeta_t, \quad \zeta_t \sim NID(0, \sigma_\zeta^2), \tag{2}$$

$$D_t = D_{t-1} + \eta_t, \quad \eta_t \sim NID(0, \sigma_\eta^2), \tag{3}$$

where D_t is the slope and the normal white noise disturbances ζ and η are independent of each other. In case $\sigma_\eta^2 = 0$, the formula (2) reduces to random walk with drift and if, in

addition, $\sigma_\zeta^2 = 0$, it reduces to deterministic linear time trend; the case when $\sigma_\zeta^2 = 0$ and $\sigma_\eta^2 \neq 0$ corresponds to integrated random walk (IRW). So, the proposed way of modelling time trend allows for time-varying parameters and incorporates possibilities of random walk and linear trend as limiting cases. The key idea of Harvey (see [4]) was to handle structural models in the state space form with the state of the system representing the various unobserved components such as trends and cycles. The forecast in this type of structural time series model are constructed automatically by the Kalman filter. The trend and other unobservables are extracted by a smoothing algorithm. The parameters, which govern the evolution of the observed series, are estimated by maximum likelihood, again using the Kalman filter. Thus the whole model is handled within a unified statistical framework, which produces optimal estimates with well defined properties.

3 Empirical results

As Latvian residential property prices time series exhibits no seasonal variation, the seasonality term in equation (1) was omitted and a cycle was incorporated into the trend-cycle component (2) - (3), for which integrated random walk specification was chosen. IRW model is known to be particularly useful for describing large smooth changes in the trend ([2]).

In classical Kalman filter framework our model has the following representation: the state equation

$$\xi_{t+1} = F\xi_t + \nu_{t+1} \tag{4}$$

and the observation equation

$$y_t = H'\xi_t + \epsilon_t,$$

where

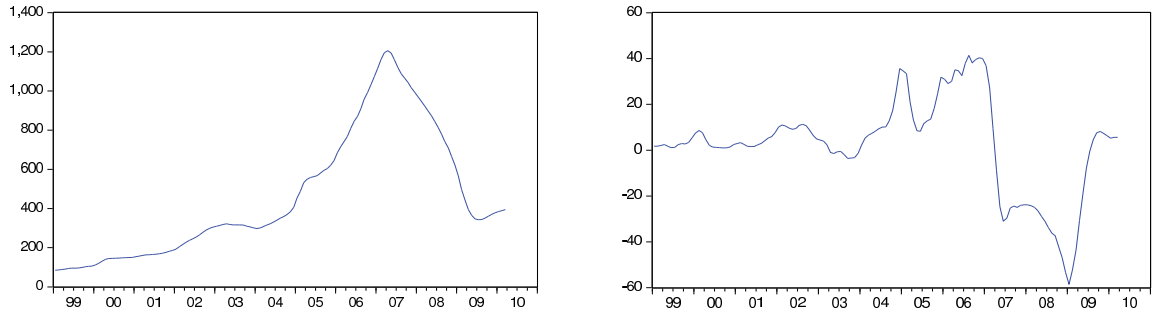
$$\xi_t = \begin{pmatrix} T_t \\ D_t \end{pmatrix}, \; F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \; H = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \nu_t = \begin{pmatrix} 0 \\ \eta_t \end{pmatrix}.$$

The model appears to be very simple and parsimonious with the only unknown parameter - σ_η^2. Following Garcia-Ferrer and Queralt [2], we will call the slope component D_t a trend derivative.

Estimation of the parameters and extracting of the components was carried out by the means of econometric package EViews7. The estimated trend and the trend derivative are shown on Figure 1.

Garcia-Ferre and Queralt (see [2]) argues, that the trend derivative can be used as a device for anticipating peaks and troughs, in particular, when the derivative reaches its maximum value, the recession is to be expected, and the recession is confirmed, when the derivative becomes negative. Indeed, also in our case the peaks in derivative precedes the turning points in the levels of time series for some 2-3 months and thus can serve to improve quantitative forecasts in the vicinity of turning points.

The trend prediction in the IRW trend model is a straight line with a constant slope equal to the last value of the derivative. This seems to be a rather restrictive and conservative assumption given the evolution of the derivative through time. Two alternatives seem to be open: (1) propose different (more flexible) trend model, like Smooth Random Walk (SRW) or Double Integrated Autoregressive model (DIAR); and (2) direct modelling of IRW trend derivative and obtaining forecasts from its univariate model, as proposed in [2]. We have left



(a) (b)

Figure 1: Estimated trend (a) and trend derivative (b) for Latvian residential property prices during January 1999 - October 2010.

the first alternative for further research and followed Garcia-Ferre identifying and estimating corresponding ARIMA models for the trend derivative. The two alternatives, suggested by the inspection of ACF, PACF and the analysis of unit root tests are AR(3) model with one root close to unity ("quasi-stationary") and non-stationary ARI(2,1) model. Therefore both competing models are rather close. Despite this fact, the forecasts, produced by the above models are rather different (see Figure 2 and Tables 2 and 3). In the following, we will call both model "modified structural models" and refer in tables as UC_AR(3) and UC_ARI(2,1) (unobserved component model with trend derivative forecasted by AR(3) and ARI(2,1) models correspondingly).

4 Forecasts based on the local linear smoothing method

For the time series observations $y_1, y_2, \ldots, y_T$ consider the commonly used kernel regression estimator introduced by Nadaraya [6] and Watson [7] defined by

$$\hat{T}_{t_0} = \frac{\sum_{t=1}^{T} y_t K(\frac{t-t_0}{h})}{\sum_{i=1}^{T} K(\frac{t-t_0}{h})},$$

where T_{t_0} denotes the trend function at a fixed timepoint t_0, K is a kernel function and h - the bandwidth parameter. It is well known that this method has a big drawback - it has the so called design bias (see, for example, [8] or [9]). This issue is more pronounced at the boundary regions, therefore this method is not quite suitable for the forecasting purposes.

The local linear smoothing is another nonparametric regression method, which minimizes the following expression

$$\sum_{i=1}^{T} \{y_i - a - b(i-t)\}^2 K_h(i-t),$$

with respect to a and b. Denote by $\hat{a}_t$ and $\hat{b}_t$ the least-squares solutions, where the subscript t is used to indicate that the solution depends on the given timepoint t. Then T_t is estimated

by the local intercept $\hat{a}_t$, which admits the explicit expression

$$\hat{T}_t = \hat{a} = \frac{\sum_{i=1}^{T} w_{t,i} y_i}{\sum_{i=1}^{T} w_{t,i}},$$

where

$$w_{t,i} = K_h(i-t)\{S_{T,2}(t) - (i-t)S_{T,i}(t)\}$$

and

$$S_{T,j} = \sum_{i=1}^{T} K_h(i-t)(i-t)^j.$$

It can be shown that using the local linear smoothing method the design bias vanishes, thus it improves over the usual Nadaraya-Watson kernel regression estimator. It is also possible to consider more general local polynomial fitting methods described by [9] in details. However, practically the local linear smoothers are used most commonly.

We have implemented the nonparametric regression smoothers in program **R**. It is well known that the kernel choice is not essential, thus we choose the standard Gaussian kernel. However, the smoothing parameter or the bandwidth choice is crucial. There exist many methods for the bandwidth selection which roughly can be divided into cross-validation and plug-in methods. Moreover, for the time-domain smoothing due to the local dependence the bandwidth selectors for independent samples do not work well (see, for example, Section 6 in [8]). Therefore for the comparison we examined several built-in automatic methods for bandwidth selection such as 1) cross-validation (`sm.regression` command choosing the method `cv`); 2) plug-in method (command `dpill`) and 3) iterative method based on autoregressive regresion errors (command `sm.regression.autocor`). In all cases we obtain similar results, that is, $h = \{2.39; 1.23; 1.14\}$, respectively. The reason may be very simple: the time series data are already quite smooth. Therefore we present here (see Section 6) only the forecasts using the plug-in method (command `dpill` which works under the package `KernSmooth`).

5 Forecasting based on ARIMA models

It has become common to use the best-fitting ARIMA class model (see [10]) as a benchmark, investigating the performance of different class models. The graph of Latvian residential property prices time series and the correlogram of levels are suggestive of non-stationary process, which was supported by the results of unit root tests (ADF and KPSS tests results both confirm I(1) process). That's why we were looking for the best model among the class of Integrated models of first order (ARMA models for first differences). ACF and PACF properties of differenced series were used to select the orders of autoregressive and moving average polinomials of the models, which were then estimated by maximum likelihood. Taking into account Akaike (AIC) and Schwartz Bayesian (BIC) information criteria, the best models from this class on the considered observation sample from January 1999 to December 2009 appeared to be ARI(2,1) and ARIMA(1,1,1) (both fitted to logs of dependant variable) with the impulse dummies for periods 2005M1 and 2009M2. Estimated equations have the following form:

$$dln(y_t) = 0.14d2005m1 - 0.11d2009m2 + 0.37(dln(y_{t-1}) + dln(y_{t-2})), \tag{5}$$

$$dln(y_t) = 0.12d2005m1 - 0.08d2009m2 + 0.87dln(y_{t-1}) - 0.37\hat{\epsilon}_{t-1}. \quad (6)$$

All estimated parameters are significant on 1% significance level. Both dummies serve to fix outliers in residuals. Positive shock in January of 2005 is connected with the change of the peg of Latvian currency - Lats from SDR currency basket to euro. Prior to 2005 property prices were denominated in USD, and the change to euro was used by property owners to rise the prices. The negative shock in February of 2009 was caused by the negative trends in real economy, which accelerated fall in property prices.

According to the both information criteria, ARI(2,1) is slightly superior to ARIMA(1,1,1) due to more parsimonious specification (see Table 1).The residual tests (for autocorrelation, normality and heteroscedasticity, not shown to save the space) revealed good statistical properties of both models ((5) and (6)).

Table 1: The values of Information criteria for the ARI(2,1) and ARIMA(1,1,1) models.

Model	ARI(2,1)	ARIMA(1,1,1)
AIC	-4.48	-4.38
SBC	-4.40	-4.29

Fitted ARIMA models (5) and (6) were employed to forecast the values of property prices series for a year ahead (January to December 2010). The forecasting results are shown and discussed in the next section.

6 Forecasting results

To compare the forecasting performance of different models, all the models were estimated, using data only from estimation set (from January 1999 to December 2009). Then the estimated models were used to predict next 12 values (from January 2010 to December 2010). Figure 2 shows forecasts produced by the models and actual values of data series, which are already known (from January 2010 to October 2010).

To access accuracy of forecasts, we used two measures MSE and MAPE, based on prediction errors ϵ_t, that is

$$e_t = y_t - \hat{y}_t,$$

where y_t - the observed value from the test set and $\hat{y}_t$ - the forecast for the time moment t, based on the values from estimation set. The **predictive mean squared error (MSE)** uses squared residuals,

$$MSE = \frac{\sum_{t=1}^{n} e_t^2}{n},$$

where n is a number of forecasts. The **mean absolute percentage error (MAPE)** considers the relative absolute error of each forecast,

$$MAPE = \frac{\sum_{t=1}^{n} \left|\frac{e_t}{y_t}\right|}{n}.$$

Tables 2 and 3 show the calculated measures for different models.

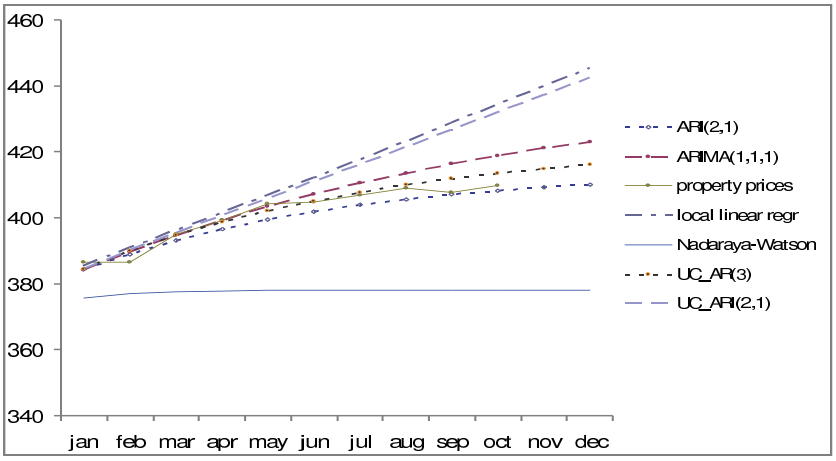


Figure 2: Different forecasts based on previously described methods together with the actual values of the data series.

Table 2: Forecast MSE for horizons 1-10.

period	lin.local regr	Kernel regr.	ARI(2,1)	ARIMA(1,1,1)	UC_AR(3)	UC_ARI(2,1)
2010M01	**1.25**	116.41	4.57	5.50	4.34	4.34
2010M02	20.85	91.93	**6.11**	10.18	11.54	13.03
2010M03	1.98	304.22	3.08	**0.05**	0.11	0.26
2010M04	5.74	456.86	6.69	**0.02**	0.25	2.18
2010M05	7.59	683.90	21.12	**0.44**	3.78	2.93
2010M06	55.21	717.28	8.35	5.59	**0.11**	38.22
2010M07	116.31	832.54	8.81	13.17	**0.62**	86.08
2010M08	201.62	957.49	11.45	20.85	**0.79**	153.71
2010M09	448.05	871.90	**0.32**	76.01	17.59	362.82
2010M10	607.41	1000.55	**2.21**	82.55	13.86	492.44
MSE	146.60	603.31	7.27	21.44	**5.30**	115.60

Table 3: Forecast MAPE (in %) for horizons 1-10.

period	lin.local regr	Kernel regr.	ARI(2,1)	ARIMA(1,1,1)	UC_AR(3)	UC_ARI(2,1)
2010M01	**0.29**	2.79	0.55	0.61	0.54	0.54
2010M02	1.18	2.48	**0.64**	0.83	0.88	0.93
2010M03	0.36	4.42	0.44	**0.05**	0.08	0.13
2010M04	0.60	5.35	0.65	**0.03**	0.12	0.37
2010M05	0.68	6.47	1.14	**0.16**	0.48	0.42
2010M06	1.84	6.62	0.71	0.58	**0.08**	1.53
2010M07	2.65	7.09	0.73	0.89	**0.19**	2.28
2010M08	3.47	7.57	0.83	1.12	**0.22**	3.03
2010M09	5.19	7.24	**0.14**	2.14	1.03	4.67
2010M10	6.02	7.72	**0.36**	2.22	0.91	5.42
MAPE	2.23	5.77	0.62	0.86	**0.45**	1.93

Tables 2 and 3 show that the overall modified structural model with AR(3) process for trend derivative performs the best according to the both criteria. The forecasts from ARIMA models also are fairly close to actual data; ARI(2) seems to be the second best. The linear local regression smoother captured well the slope for the beginning of the test sample: forecasts for the first 5 months are accurate, but the deviation is rather big for the rest of the test sample. Surprisingly, but the forecasts from the modified structural model with ARI(2,1) differ very much from the first structural model, despite the fact that actually the two models for the slope are rather close. However, those results show the improvement compared to the structural model with a pure random walk slope (not shown due to big deviations from actual series and other forecasts: forecasted value for October 2010 was 475, much worse than local linear regression). Nadaraya-Watson method also was not a success. So, forecasting fairly smooth time series after the turning point, usage of the local linear regression can be recommended only for short time horizons. ARIMA models performed fairly well for all horizons (1-10) with MAPE not exceeding 0.62%. The performance of the modified structural model crucially depends on the choice of the model for forecasting trend derivative.

7 Conclusions

In this paper we investigated the forecasting ability of different univariate forecasting techniques (local regressions, modified structural model, standard ARIMA approach). A forecasting example was carried out with each method, using monthly price time series on residential property prices in Latvia, which has recently experienced turn in the trend. The accuracy of the different methods was assessed by comparing the forecasts MSE and MAPE based on the out-of-sample predictions for 10 horizons. The modified structural model with AR(3) process for trend derivative performed the best according to the both criteria, with MAPE less than 0.5%. ARIMA models performed fairly well for all horizons (1-10) with MAPE not exceeding 0.62% and the local linear regression gave accurate forecasts for short horizons (up to 5). Nadaraya-Watson method didn't look suitable for this example. But it should be mentioned that the results from the modified structural model should be interpreted with caution, as another specification for trend derivative yielded much worse results. This feature of the model calls for the further investigation. In particular, more flexible trend characterisations (like smoothed random walk or dependences of orders higher than one) seems an important area for future research.

References

[1] ARINS, M.: *Decade in Latvia's housing Market (2001 - 2010)*, Národná Banka Slovenska, Biatec (Banking journal), Vol. 26, pp. 28-29, 2010.

[2] GARCIA-FERRER, A., QUERALT, R.A.: *Can univariate models forecast turning points in seasonal economic time series?*. International Journal of Forecasting, Vol. 14, pp. 433446, 1998.

[3] YOUNG, P.C.: *Time variable parameter and trend estimation in non-stationary economic time series*. Journal of Forecasting, Vol. 13, No. 2, pp. 179210, 1994.

[4] HARVEY A.C.: *Forecasting, Structural Time series Models and the Kalman Filter*. Cambridge: Cambridge University Press, 1989.

[5] HARVEY A.C., JAEGER, A.: *Detrending, stylized facts and the business cycle*. Journal of applied Econometrics, Vol. 8, pp. 31-47, 1993.

[6] NADARAYA, E.A.: *On estimating regression*. Theory of Probability and Its Applications. Vol. 9, No. 1, pp. 141–142, 1964.

[7] WATSON, G.S.: *Smooth regression analysis*. Sankhyā: The Indian Journal of Statistics, Series A. Vol. 26, No. 4, pp. 359-372, 1964.

[8] FAN, J. and YAO, Q.: *Nonlinear time series: nonparametric and parametric methods*. Springer Verlag, 2003.

[9] FAN, J. and GIJBELS, I.: *Local polynomial modelling and its applications*. Chapman & Hall/CRC, 1996.

[10] BOX, G.E.P., JENKINS, G.M.: *Time Series Analysis: the Forecasting and Control*. Holden Day, San Francisco, 1970.

Current address

Nadezhda Sinenko, Dr. math.
Department of Mathematics,
Faculty of Physics and Mathematics,
University of Latvia,
Zellu 8, LV-1002, Riga, Latvia.
e-mail: sinenko@latnet.lv

Janis Valeinis, Dr. math.
Department of Mathematics,
Faculty of Physics and Mathematics,
University of Latvia,
Zellu 8, LV-1002, Riga, Latvia.
e-mail: valeinis@lu.lv

Journal of Applied Mathematics

volume IV (2011), number III

CLASSIFICATION OF INDIVIDUALS: WILLINGNESS TO START THEIR OWN BUSINESS BASED ON FRANCHISE SYSTEM

ŽAMBOCHOVÁ Marta, (CZ), TIŠLEROVÁ Kamila, (CZ)

Abstract. Encouraging and supporting people to start their own business are one of the important government measurements towards to the GDP growth. It is necessary for the success of this policy to identify the types of people who are willing to start a business as independent entrepreneurs, or run it in the form of license and establish a business based on a franchise system. The research was conducted in order to clearly identify and classify these people and to outline their main characteristics important for future effective franchise systems promotion. A survey on the willingness to start a business in the form of a franchise license was conducted with the usage of questionnaires and different methods of classification were chosen. Two of the algorithms for classification trees – CART and CHAID were used for its evaluation from the group of supervised learning methods. From the group of unsupervised learning methods the two–step cluster analysis was chosen. This paper analyzes the results obtained using these methods. It aims to create classes of respondents with similar opinions on entrepreneurship and franchising, as well as to classify respondents in terms of their willingness to start a business and opinions on investment possibilities in a business context. The survey showed that franchising might be a matter of change and new hope for many people.

Key words. classification, decision tree, cluster analysis, starting business, franchising

Mathematics Subject Classification: 91B06, 90B50.

1 Introduction

The franchise model of doing business is regarded as the less risky option of starting a business. The research on willingness to start one´s own business, both with franchise and regular concept, was conducted in the form of questionnaire surveys realized in the Czech Republic by the team of the Faculty of Social and Economic Studies (FSE) in Ústí nad Labem in 2008 – 2009. The aim of the research was to identify the segment of the population which is most likely to start their own business. According to the description of characteristics for this group the proper measures for investment incentives can be taken. The questionnaire research was conducted in order to determine public awareness of the franchise; find out the views and attitudes towards the franchise system;

explore the willingness to start a business; and collect enough information to be able to suggest some measures for effective communication on the franchise system – the presentation of the franchise as a promising way of doing business.
Unfortunately, the traditional basic statistic methods led to unsatisfactory outputs so that the methods of classification by the support of cluster analysis and decision tree methods were chosen. This contribution deals with the analysis results gained by means of those two methods.

2 Methodology

This contribution deals with the classification of two basic ways of learning – namely supervised learning and unsupervised learning. In the first case the decision rules for assigning objects to the groups are created according to the training set. The decision trees are representatives of this group of methods. In the second case the selected objective function due to its minimization divides objects into categories, so that the objects belonging to one category are more similar to each other than data from different categories. Cluster analysis belongs to this group.

2.1 Cluster Analysis

Cluster analysis, see [1], [3], deals with data objects similarity. It solves the set of objects splitting into several previously non-specified groups (clusters) so that the objects in the single clusters are the most similar to each other as possible and the objects outside of the different clusters should be the least similar as less as possible. Cluster analysis can be realized by many different methods.
The statistical program systems usually include both the hierarchical algorithm result which is usually displayed in the form of a dendogram and non hierarchical iterative algorithm *k*-means and very often also a two-way joining. In the statistical system SPSS there is a two-step method implemented starting at the 11.5 version.
The choice of a hierarchical method was not suitable for this survey due to the relatively large number of subjects. The algorithm *k*-means is designed for a clustering of objects which are described with the use of quantitative variables and it was not the case on this research. The usage of this method would require pre-proceeding the data with the help of binarization; it means that each variable transfers into several binary variables (the variable of the value 0 and 1). The most suitable method for data proceeding in this survey seems to be the two-step method.
The principles of the two-step cluster method are described for example in [2]. This method uses the algorithm of BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), which is explained in more detail in [6], or [7]. The algorithm creates a so-called CF-tree, which is progressively fulfilled by incoming data. The advantage of this principle is that it goes through the data file only once. The disadvantage is the sensitivity for the entry data ordering.
CF-trees work with so-called CF-characteristic – Clustering Feature of the cluster. Data collected in CF-characteristic are sufficient for the calculation of centroids, inter-group proximity measures and compactness of clusters. This characteristic creates an organized triad of $CF = (N, LS, SS)$, where N means the number of objects in a cluster, LS represents a vector sum of all cluster objects and SS states these objects Square coordinates, e.g.

$$LS = \sum_{i=1}^{N} X_i \ , \ SS = \sum_{i=1}^{N} X_i^2 \ . \tag{1}$$

CF-trees are highly balanced trees of two parameters. The first parameter is the threshold P and the second one is the branching factor (F, L). Each internal node of CF-tree applies in that it contains maximally F descents. The task of the internal nodes is to allow the finding of the proper leaf for new subject categorisation. Each leaf contains maximally L entries. Every leaf node represents a cluster created by all the sub-clusters constituted by the single entries of the leaf. However, the threshold rule has to be valid for every leaf entry that the entry radius is smaller than thereshold P.
The clustering algorithm is realized in three main phases. In the first phase the CF-tree is created and the entering objects are progressively organized. In the second phase the CF-tree is condensated and optimalized due to its threshold adjustment and with the help of the proper tree re-designing the outliers is eliminated. In the third phase the impact of entry data order sensitivity is minimized. The algorithm clusters together with the leave´s tops using the agglomerative hierarchy cluster algorithm.

2.2 Decision Trees

Various types of decision trees are widely used in data models. The decision trees can be regarded as the structures which recursively separate surveyed data according to certain decision criteria. The root represents all of population file. The inner nodes demonstrate the sub-systems of the population set. The values of dependent variable are explained in the tree leaves. Two types of decision trees have been used: the classification trees (every leaf contains a category) and regression trees (every leaf contains a constant – the estimation of dependent variable).
The decision tree has been recursively created by space division of independent variable values and has been based on searching the question (splitting condition), which is the best of all for dividing the surveyed data space into sub-sets, it means which one maximizes the splitting criterium. The splitting procedure is finished as soon as the cessation rule is reached. There are two possible ways to set up the quality of generated tree: the system of training and test data and the other way is the cross validation.
A large number of algorithms was developed for the decision trees creation. CART, ID3, C4.5, AID, CHAID and QUEST algorithms are the most frequent ones, see [8]. This contribution treats with two algorithm types implemented in the statistical system – CART and CHAID.

2.2.1 Algorithm CART

This algorithm was originally described by its authors Breiman, Freidman, Olshen and Stone in 1984 in the article „Classification and Regression trees“. The algorithm (see [4], [5]) can be applicable in the case that there are one or more independent variables. These variables can be continuous or categorical (both ordinal and nominal). There is also one dependent variable, which can be also categorical (both ordinal and nominal) or continuous.
Because only YES/NO questions (condition of splitting) are permitted, the algorithm result can be composed only in the form of the binary tree (it means that every node is divided into two child nodes). In every algorithm step the algorithm goes through all potential splitting with the help of all permissible values of all variables and the best solution is searched for. The increasing of data purity serves as the measurement. It means that one splitting is better than the other one if two more homogenous (according to independent variables) data files are acquired compared to another way of splitting. Algorithm splitting differs for classification trees and for regression trees.

The child node homogeneity is in the case of the classification trees measured by the impurity function $i(t)$. The maximal homogeneity of two newly built child nodes is constructed as the maximal purity reduction $\Delta i(t)$.

$$\Delta i(t) = i(t_r) - E(i(t_d)), \tag{2}$$

where t_r represents parent node, t_d is the child node. In order to set up the child node t_p, the probability of child node P_p and the left child node t_l, the probability of the left child node P_l the expected value formula should be supplied as follows:

$$\Delta i(t) = i(t_r) - P_l \cdot i(t_l) - P_p \cdot i(t_p). \tag{3}$$

For each node the CART algorithm solves the maximization problem for the $\Delta i(t)$ function going through all the potential splitting. The $\Delta i(t)$ function can be defined in different ways. The most frequent is the Gini index method.
The regression trees are used in the case when the dependent variable is not categorical. The algorithm searches for the best splitting based on the sum of variance minimizing in the terms of two newly built child nodes in this case. This algorithm works on the basis of the algorithm of minimizing the sum of squares.

2.2.2 CHAID Method

The method of CHAID (Chi-squared Automatic Interaction Detector) was developed in 1980 by G.V. Kass. This method, see [5], has arisen by the modification of the AID method for categorical dependent variable. The non-binary trees can be regarded as a result of this modification. The method uses the χ^2-test. The splitting algorithm is realized as follows: In the terms of one leaf node the contingency table (sized $m \times k$) values of independent variable (m categories) is created. After that the pair of the category of independent variable predictor is found and the sub-table sized $2 \times k$ has the less important value of χ^2- test. These two categories are merged. By this operation the new contingency table is created – sized $(m - 1) \times k$. The merge procedure is repeated until the significance of χ^2-test declines under the pre-scribed value. Reaching this the splitting procedure of one parent node to several child nodes has been finished. The process continues in this way for each of the leaf node until the insignificant result of χ^2-test is reached.

3 The Results and Evaluation

The questionnaire survey was carried out on 658 respondents. It was a random sample of the Usti region and only people over 18 years were asked to take part in it. The idea was to keep the structure of respondents as close to the natural structure of population situation as possible. That is the reason why three control points were considered as follows: age, education and residence.

3.1 Basic Information about the Structure of the Respondents

54% of respondents answered “yes” to the question “Have you ever heard of a franchise before?”. The overall response shows that 30% of respondents have heard of it at school, 28% on the internet, 18% in the press, 17% by friends and 7% from TV. Wide differences, however, were in the age structure of respondents. The results of the survey shows that nearly half of the population under 35

is equipped with the knowledge of the franchise from school (i.e. passive reception), while in the age group of 36-55 years it was only 13%. Yung people have heard of the franchise from the press only at 11%, older people at 28%. For ages above 56 years information from the press predominates (35%), followed by information by friends (25%).
Perceptions of the advantages and the disadvantages of franchising were examined. Almost half of the population already perceives the franchise system as an easy way suitable even for less experienced people to start. There should not be a major problem to convince the "rest of the population" of this fact (using suitable means). 30% of respondents consider a hindrance to business in the first place "risks of their failure". It is necessary to work intensively with the fact that people generally perceive the franchise as a solution for those who still defer starting a business (72%), but only a half of them (56%) would recommend the franchise to their friend.
A branch of business that is seen as the best for starting a business was also studied. A third of people could be in the field of the gastronomy, 13% in education, 10% in construction, 8% in information technology, 7% in real estate.
Their investment would have saved most respondents in real estate (57%), in own business (12%) over other forms of investment have slightly higher securities (10%).

3.2 Formulation of Respondent Categories with the Similar Opinion on Doing Business and Franchising

At first the cluster analysis classification was done on the basis of 23 variables containing answers to questions regarding respondents´ opinions on doing business and franchising. Because there is the combination of different variables types, the two-steps cluster method and dissimilarity measure (of the type of distance likelihood) were used. The procedure in SPSS system evaluated as optimal just two clusters. The result was regarded by the procedure as a good result, see Figure 1.

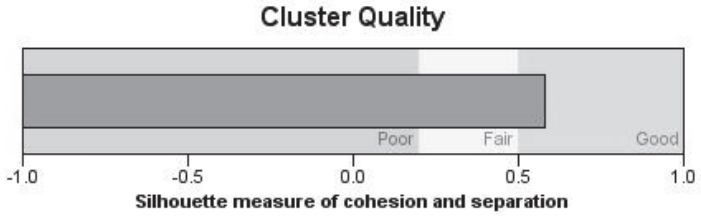


Figure 1: Cluster Quality

The cluster merge was introduced as another variable. In the next step the classification tree for the newly built variable "cluster No." As the independent variable the identification data of respondents was chosen. The classification tree was created with the help of two methods which are included in the SPSS system, namely the CHAID method and the CRT (CART) method. In both cases the cross validation was selected in order to confirm the tree quality. The tree of higher quality was generated by the CRT method – see Figure 2. The level of risk estimate resulted in the value of 0.253. So the „risk" of misclassifying is 25.3%, the model classifies 74.7% cases correctly.
From the tree structure it is possible to acquire the information that a similar opinion on doing business and franchising occurs with higher educated people, managers, respondents aged more then 35 living in big cities and aged less than 35 living in the villages which are larger then 10 000 inhabitants working in their job position from one up to five years. The second group of similar opinion people consists of persons without higher education aged more then 35 years living in the cities of up to 50 000 inhabitants and also persons under 35 years living in small villages and also

young people from middle – and bigger sized cities working in their positions for the period of one to five years.

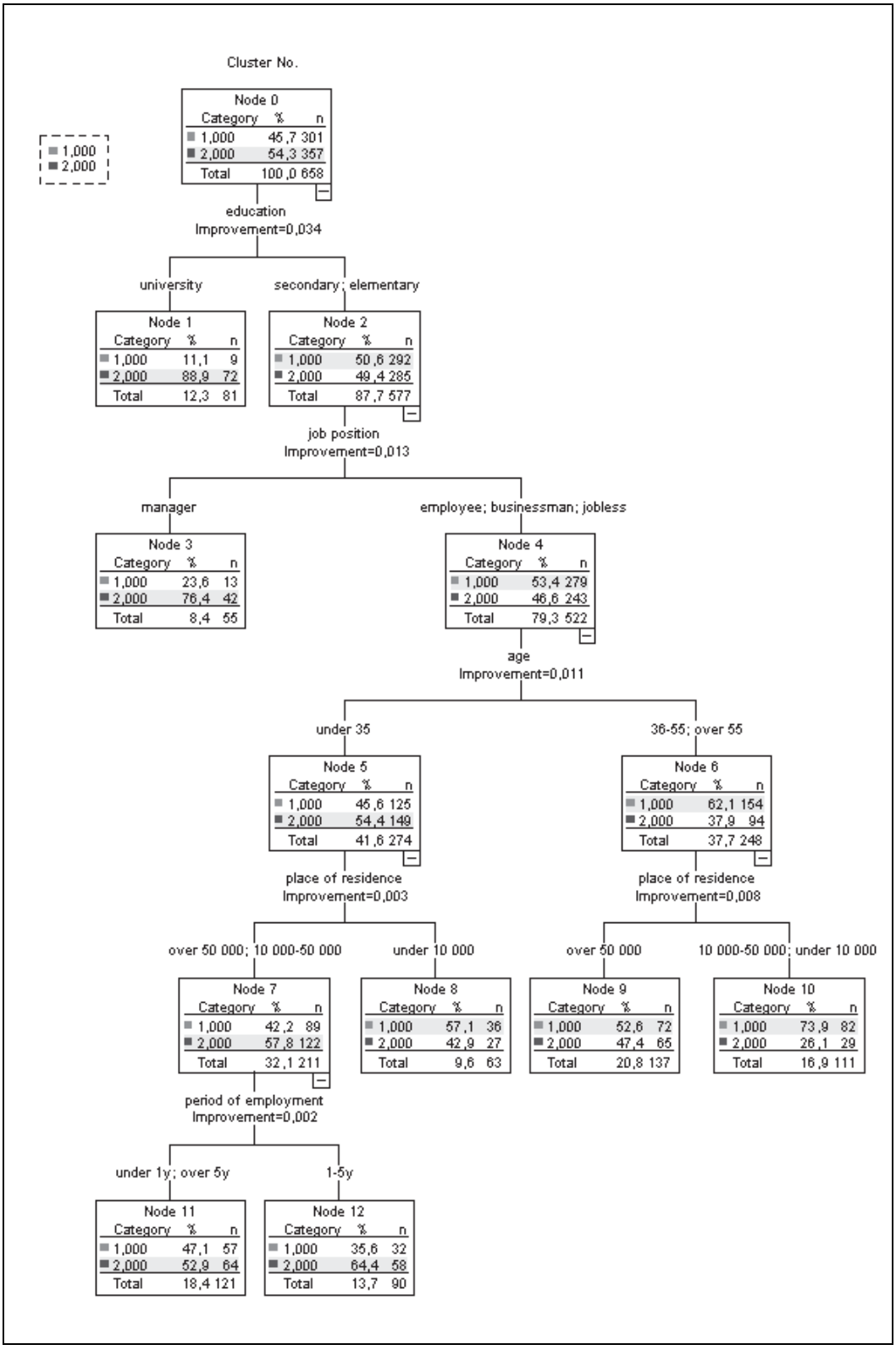


Figure 2: CRT – the opinion on doing business and franchising

3.3 The Respondent Classification from the Point of View of Their Willingness to Start Their Own Business

The next step was focused on the respondent classification from the point of view of their willingness to start their own business. The classification tree dependent variable "Willingness to start own business" which takes the value Yes – No. The respondent identification data was chosen as the independent variables. Both of the above mentioned methods were used again to create the tree. In this case the better result was generated by the CHAID method – see Figure 3. The risk estimate value resulted in 0.224. So the „risk" of misclassifying is 22.4%, the model classifies 77.4% cases correctly.

Due to the tree structure it was found that the highest level of willingness to start and run their own business show the unemployed people, the ordinary employees aged less than 35 years and a substantial part of these persons is composed by managers and self-employed people. The significant disinclination to run one´s own business was identified among ordinary employees aged more than 35 years.

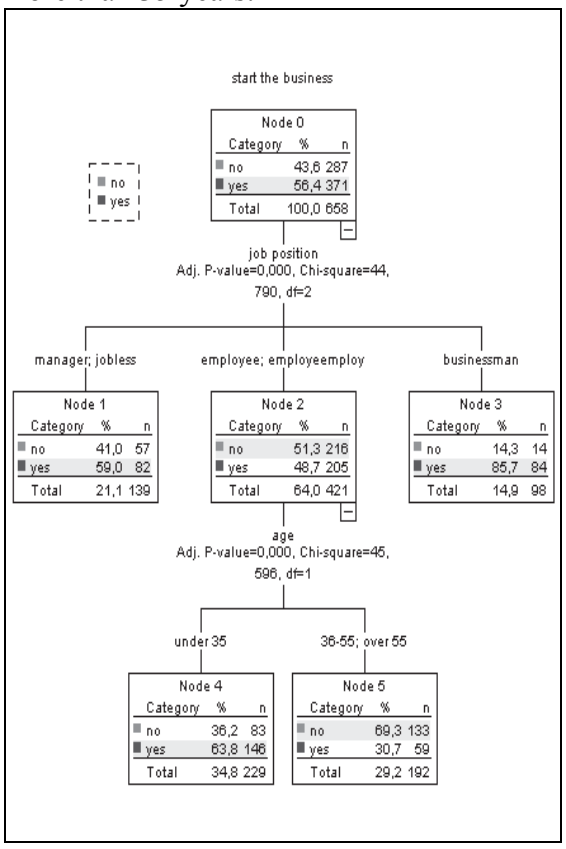


Figure 3: CHAID – willingness (business)

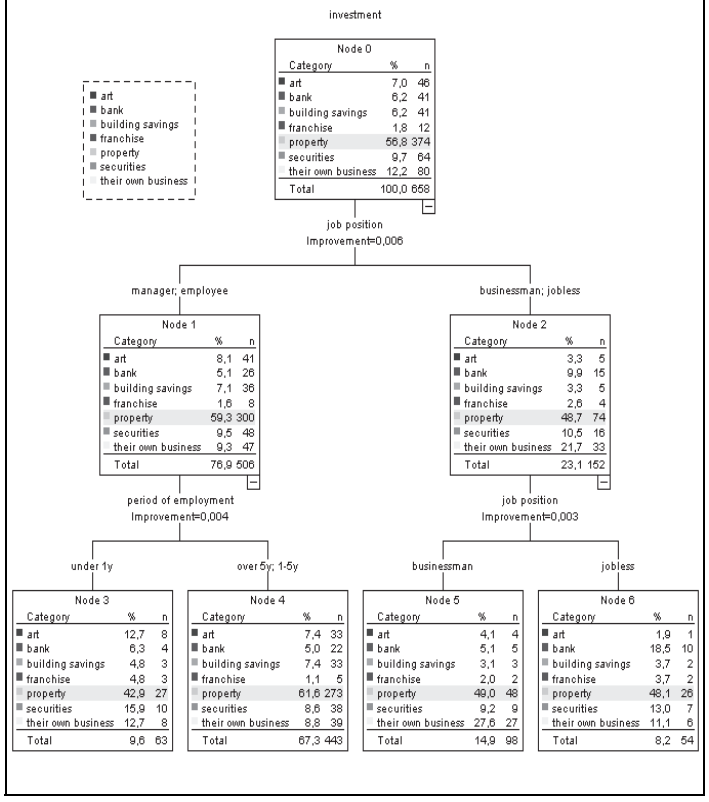


Figure 4: CRT – investments

3.4 The Respondent Classification according to Their Investment Criteria

In the last step, attention to respondent classification according to their criteria for surplus funds allocation was paid. In this case the quality of the two trees generated by both the methods was the same, unfortunately worse than in the previous cases. CRT tree is demonstrated in Figure 4. The risk estimate value is 0.332. So the „risk" of misclassifying is 33.2%, the model classifies 66.8% cases correctly.

The results of the above mentioned classification can be summarized as follows: Managers and ordinary employees holding their position less than one year think about the investment allocation in a similar way. For above the average, they prefer investment in franchise systems and the investment allocation in the bonds and subjects of art is subordinated to franchise. Conversely, they have a negative approach to real estate investments.
The next group having similar consideration towards investments consists of managers and ordinary employees holding their position for more than one year. They prefer funds allocation in the field of real estate and savings accounts designed for future construction. In contrast they avoid investment into both their own businesses and businesses under the franchise system.
The third group contains entrepreneurs who (as was expected) highly prefer business on their own account and refuse investments in the field of real estate, savings accounts designed for future construction and objects of art.
The last group consists of unemployed persons who give significant priority to deposits. They also prefer bond purchase and franchise investment at an above average. They are representatives of refusing objects of art purchases and under the average they would like to place their funds in the real estate and construction savings accounts.

4 Conclusion

The results of the survey show that the majority of the population is equipped with the knowledge of franchise from school (i.e. passive reception) and that is why it would be worth considering an involvement in further education at schools (lectures, competitions, eventually the usage of student media). Half of the respondents who would eventually start their business recruits from ordinary employees who in many cases occupy their position for more then five years. Therefore, franchising for them must be a matter of change and new hope. Also the franchise seems to be a chance for unemployed people and it is therefore appropriate to focus the Labour Office courses on franchising opportunities.
The following reasons can be (almost equally) regarded as the greatest obstacle to starting a business: the risk of failure, lack of funds and administrative barriers. As noted in part above, it is necessary to develop such a franchise system promotion which is based on the elimination of fear, mainly the fear of the risk of failure and fear of administrative complications.

References

[1] EVERIT, B.S., LANDAU, S., LEESE, M.: *Cluster Analysis*, 4. edition, Hodder Arnold, London, 2001.
[2] ŘEZANKOVÁ, H.: *Shlukování a velké soubory dat*, Lázně Bohdaneč 29.11.2004 – 01.12.2004. In: KUPKA, Karel (ed). Analýza dat 2004/II Pardubice: TriloByte Statistical Software, 2005, pp. 7–19.
[3] ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V.: *Shluková analýza dat, 2. vydání*, Professional Publishing, Praha, 2009.
[4] TIMOFEEV R.: *Classification and Regression Trees* (*CART*) *Theory and Applications*, CASE–Center of Applied Statistics and Economics, Humboldt University, Berlin, 2004.

[5] WILKINSON, L.: *Tree Structured Data Analysis: AID, CHAID and CART*, Sun Valley, ID, Sawtooth/SYSTAT Joint Software Conference, 1992.
[6] ZHANG, T., RAMAKRISHNAN, R., LIVNY, M.: *BIRCH: An Efficient Data Clustering Method for Very Large Databases*, ACM SIGMOD Record, Vol. 25. No. 2, 1996, pp. 103–114.
[7] ZHANG, T., RAMAKRISHNAN, R., LIVNY, M.: *BIRCH: A New Data Clustering Algorithms and Its Applications*, Journal of Data Mining and Knowledge Discovery, Vol. 1, No. 2, 1997, pp. 141–182.
[8] ŽAMBOCHOVÁ, M.: *Jak na rozhodovací stromy*, Informační Bulletin, Praha, Vol. 19. No. 3, 2008, pp. 1–12.

Current address

Marta Žambochová, RNDr., Ph.D.
University of Jan Evangelista Purkyne, Faculty of Social and Economic Studies, Department of Mathematics and Statistics, Moskevska 96, Usti nad Labem, 400 96, tel.: 475 284 808, e-mail: Marta.Zambochova@ujep.cz

Kamila Tišlerová, Ing.
University of Jan Evangelista Purkyne, Faculty of Social and Economic Studies, Department of Business Administration, Moskevska 96, Usti nad Labem, 400 96, tel.: 475 284 730, e-mail: Kamila.Tislerova@ujep.cz

Journal of Applied Mathematics

volume IV (2011), number III

INDICATORS OF TURNING POINTS IN CZECH FINANCIAL TIME SERIES

ŽIŽKA David, (CZ)

Abstract. The development of financial time series was analysed separately before and after depression (2008-2009) in Czech financial markets. In the first part financial time series of stocks and exchange rates were modeled. The volatility models were used mainly. In the second part the turning points in specific time series were tested and the deviations of parameters in different parts of time series were analysed. Final part was focused on finding indicators which predicted the turning points.

Key words. Turning points, financial time series, volatility models

Mathematics Subject Classification: 62M10

1 Introduction

Czech financial market was strongly hit by the financial crisis during 2008-2009. Behavior of financial time series was heavily predictable. This study is focused on analyzing the behavior of these series.

Financial time series often exhibit characteristics that allow them to model using classical methods. In particular, the outliers, variance is depend on time, coefficient of kurtosis is high. Models of volatility can cover these properties.

The input data for analysis are three time series. Daily values of ČEZ a.s. (Czech Power Company), index PX (Prague market index), exchange rates of CZK/EUR.

Reference period is during 2002- June 2010 and involves *CEZ*, *PX*, *CZK/EUR* close daily values. Logarithmic returns expressed as percentages will be used.

First part is focused on estimating the best volatility models during 2002-2010.

Second aim of this study is model time series during depression 2008-2009 and compares with parameters whole time series. For this purpose is necessary to determine the turning points in the time series. For detection the turning points can be used technical analysis but for purpose of this study are chosen local maximum (minimum) for each time series before (after) market fall.

Final part is focus on finding indicator which predicts the turning points. Analyse short term before turning points by volatility models. Forecasts by volatility models are compared with real values.

1.1 Models of volatility

GARCH model

GARCH means Generalized Autoregressive Conditional Heteroskedasticity model. Bollerslev (1986) proposes a useful extension of ARCH model known as the generalized ARCH (GARCH) model. Bollerslev extended ARCH model of delayed conditional variance.

GARCH (1,1)

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{1.1}$$

Positive conditional variance ensures the conditions: $\omega > 0, \alpha_1 > 0, \beta_1 \geq 0$.

IGARCH model

IGARCH means Integrated GARCH and is a special form of the more general GARCH model.
It looks exactly like a regular GARCH model. In order for this model to be an IGARCH model, it has to fulfill the following condition: *alpha*+*beta*=1.
Hence, the conditional variance of the IGARCH model is clearly non-stationary. This has important implications for interpreting the volatility of such a time series. If *alpha* and *beta* indeed sum up to 1, the volatility of the model is not mean-reverting. External shocks leading to the change in volatility are permanent.

EGARCH model

Exponential GARCH model by Nelson (1991) describes an asymmetric effect between positive and negative asset returns. Conditional variance is an asymmetric function of past ε_t as defined by

EGARCH (p,q)

$$\log \sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \left(\phi z_{t-i} + \gamma \left(\left| z_{t-i} \right| - E\left| z_{t-i} \right| \right) \right) + \sum_{i=1}^{p} \beta_i \log \sigma_{t-i}^2 \tag{1.2}$$

There are no restrictions on the parameters α_i, β_i to ensure nonnegativity of the conditional variances. If $\alpha_i \phi < 0$, the variance tends to rise (fall) when ε_{t-i} is negative (positive) in accordance with the empirical evidence for stock returns. Assuming $z_t = \varepsilon_t \sigma_t^{-1}$, is i.i.d. normal, it follows that is ε_t covariance stationary provided all the roots of the autoregressive polynomial $\beta(\lambda) = 1$ lie outside the unit circle.

TGARCH

Threshold GARCH model by Zakoian [7] is similar to GJR GARCH model by Glosten, Jagannathan and Runkle (1993). Specification for conditional variance is

$$\sigma_t^2 = \omega + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{k=1}^{r} \gamma_k \varepsilon_{t-k}^2 I_{t-k}^- \tag{1.3}$$

where I_t^- =1 if ε_t <0 and 0 otherwise.

Positive and negative returns have different effects on conditional variance in this model. Positive returns ($\alpha_{t-i} > 0$) affect α_i, but negative returns affect $\alpha_i + \gamma_k$. If $\gamma_k > 0$ negative returns increase volatility and called there are *leverage effect* for the i-th order. If $\gamma_k \neq 0$ the returns impact is asymmetric. GARCH model is a special case of model TARCH if *threshold* expression equals zero.

2 Model description

1 2.1 CEZ

Following graphs show absolute close daily values and returns in percentage of CEZ stocks. The beginning of the market decline in 2008 is evident from the x-axis value 1600.

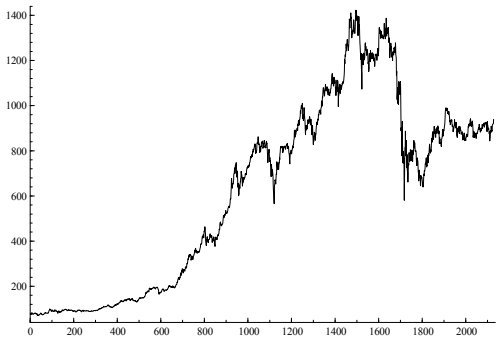


Fig. 2.1 Values of CEZ stocks during 2002-2010

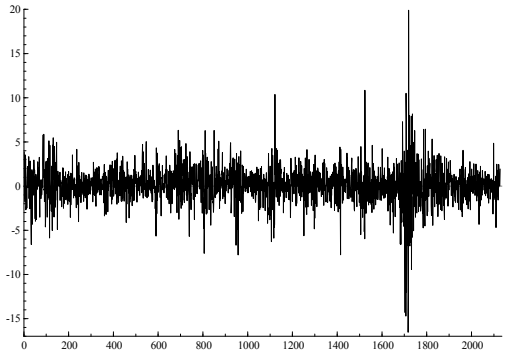


Fig. 2.2 Returns of CEZ stocks during 2002-2010 (in percentage)

Linear models

In GARCH(1,1) model are parameters statistically significant and their sum is close to 1 (alpha+beta=0.9731). For this reason, we can construct an accurate model IGARCH.

Table 2.1 IGARCH (1,1) model parameters

IGARCH	Coefficient	Std.Error	t-value	t-prob
omega	8.6507E-6	1.2132E-6	7.13	0.0000
alpha_1	0.1352	0.007089	19.07	0.0000
beta_1	0.8648	0.007089	122.00	0.0000

IGARCH (1,1): $\sigma_t^2 = 8.65E^{-6} + 0.1352\varepsilon_{t-1}^2 + 0.8648\sigma_{t-1}^2$

Nonlinear models

Sign bias (SB) test indicates impact of positive and negative returns on conditional heteroscedasticity. Negative size bias (NSB) test and Positive size bias (PSB) test indicate impact of positive and negative returns on conditional heteroscedasticity depend on their values. Model TGARCH confirms result of SB, PSB, NSB tests. Parameter treshold is statistically significant at 1% level in the model and indicates nonlinearity.

Table 2.2 TGARCH (1,1) model parameters

TGARCH	Coefficient	Std.Error	t-value	t-prob
omega	0.159391	0.02762	2.58	0.0101
alpha_1	0.062334	0.01163	3.25	0.0010
beta_1	0.859369	0.01285	27.5	0.0000
Threshold	0.093551	0.01988	3.77	0.0000

TGARCH (1,1): $\sigma_t^2 = 0.8594\sigma_{t-1}^2 + 0.0623\varepsilon_{t-1}^2 + 0.0936\varepsilon_{t-1}^2 I(\varepsilon_{t-1} < 0)$

Model checking

Portmanteau test does not confirm autocorrelation for both models. Further, Jarque-Bera test (Jarque, Bera 1987) often called *asymptotic test* does not confirm normal distribution. Result of ARCH test does not confirm conditional heteroscedasticity.

2.2 PX index

Linear models

In GARCH(1,1) model are parameters statistically significant and their sum is close to 1 (alpha+beta=0.9826). For this reason, we can construct an accurate model IGARCH.

Table 2.3 IGARCH (1,1) model parameters:

IGARCH	Coefficient	Std.Error	t-value	t-prob
omega	3.6497E-6	6.111E-7	5.97	0.0000
alpha_1	0.1539	0.0133	11.53	0.0000
beta_1	0.8461	0.0133	63.42	0.0000

IGARCH (1,1): $\sigma_t^2 = 3.65E^{-6} + 0.1539\varepsilon_{t-1}^2 + 0.8461\sigma_{t-1}^2$

Nonlinear models

SB, PSB, NSB tests indicate nonlinearity. Model TGARCH confirms result of SB, PSB, NSB tests. Parameter threshold is statistically significant at 1% level in the model and indicates nonlinearity.

Table 2.4 TGARCH (1,1) model parameters

TGARCH	Coefficient	Std.Error	t-value	t-prob
omega	0.074381	0.01459	3.73	0.0000
alpha_1	0.071807	0.01429	5.04	0.0000
beta_1	0.829651	0.01845	35.6	0.0000
Threshold	0.136707	0.03099	3.01	0.0030

TGARCH(1,1): $\sigma_t^2 = 0.0744 + 0.8297\sigma_{t-1}^2 + 0.0718\varepsilon_{t-1}^2 + 0.13676\varepsilon_{t-1}^2 I(\varepsilon_{t-1} < 0)$

Model checking

Portmanteau test does not confirm autocorrelation for both models. Further, asymptotic test does not confirm normal distribution. Result of ARCH test does not confirm conditional heteroscedasticity.

2.3 CZK/EUR exchange rates

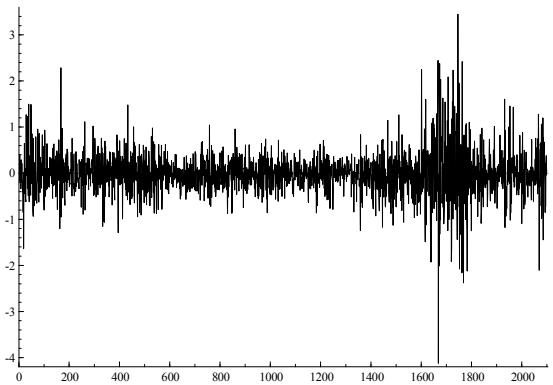


Fig. 2.3 Returns of CZK/EUR exchange rates during 2002-2010 (in percentage)

Linear models

In GARCH(1,1) model are parameters statistically significant and their sum is close to 1 (*alpha*+*beta*=0.9891). For this reason, we can construct an accurate model IGARCH.

Table 2.5 IGARCH (1,1) model parameters

IGARCH	Coefficient	Std.Error	t-value	t-prob
omega	1.7178E-7	3.3086E-8	5.19	0.0000
alpha_1	0.0928	0.00872	10.64	0.0000
beta_1	0.9072	0.00872	104.03	0.0000

$$\text{IGARCH (1,1): } \sigma_t^2 = 1.78E^{-7} + 0.0928\varepsilon_{t-1}^2 + 0.9072\sigma_{t-1}^2$$

Nonlinear models

SB, PSB, NSB tests do not indicate nonlinearity. Model EGARCH and TGARCH confirm result of SB, PSB, NSB tests. Parameters are not statistically significant in the models.

3 Turning points

For analyzing series of economic downturn it is necessary to determine turning points (TP) which define the term. For detection the turning points can be used technical analysis e.g. Chaikin's Volatility, Bollinger Bands. These technical indicators give SELL signals for several days after local maximum. Chaikin's Volatility Index (10-day exponential moving average) for PX gives SELL signal 5 days after reaching local maximum (Fig.3.1). For purpose of this study are chosen local maximum (minimum) for each time series before (after) market fall. Behavior of volatility models will be analyzed in these points.

Table 3.1 Turning points

TP	Negative TP	Positive TP
CEZ	1.9.2008	5.3.2009
PX	1.9.2008	18.2.2009
CZK/EUR	22.7.2008	19.2.2009

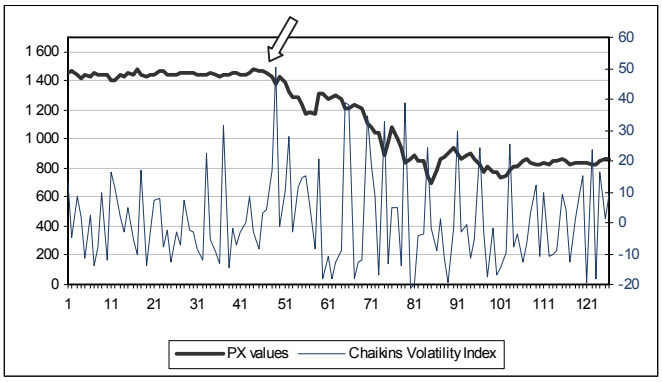


Fig. 3.1 Negative turning point - PX index during 1.7.2008-30.12.2008 and Chaikin's Volatility Index

4 Comparison of models before and during depression

Time series are estimated by volatility models as a whole from 2002 to the "Negative TPs". Further, series are divided by "Positive TPs" and estimated individual parts. For testing structural changes is used Chow test.

Chow test
Chow test introduced by G.Chow [6] determines whether the coefficients in the model are consistent with coefficients in the groups generated by separating the time series of the turning point to 2 (or more) groups. F statistic:

$$F = \frac{\frac{RSSR - SSR_1 - SSR_2}{k}}{\frac{SSR_1 + SSR_2}{T - 2k}} \tag{4.1}$$

RSSS, SSR_1, SSR_2 = sum of squared scaled residuals

$F_{0,95}(k,T-2k)=2.6049$

Table 4.1 F-values

	CEZ	PX	CZK/EUR
F	<0.001	<0.001	<0.001

Values of criteria F are in all cases less than the critical value, do not reject the null hypothesis of stability parameters. Parameters do not change in time.

5 Indicator

Strong market failure may be caused by fundamental information which starts confidence of investors. These turning points are identified in Chapter 3. In the immediate period preceding TP investors can just wait for this information. This assumption verify by analyzing the period before the TP. This expectation should be accompanied by a reduction in the value of the conditional variance model. This theory is supported by a significant reduction in the volume of shares traded before the turning point.

5.1 Conditional variance

Conditional variance reached a local minimum in the turning point. For all three time series was tested whether this local minimum is also a global minimum in moving periods (1 year ago, 2 years ago, etc.).

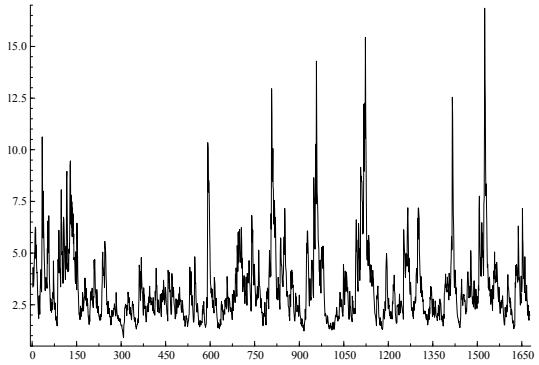


Fig. 5.1 Conditional variance of TGARCH(1,1) model - CEZ stocks during 2002-2008

For the tested models were stored values of conditional variance. Minimum of the conditional variance was found for each year and compared with NTP values (Table 5.1). The assumption was refuted already in the first year for CEZ and CZK/EUR. Both values were identical only for PX index in 2008. Did not prove conclusively that the downturn was preceded by minimums of the conditional variance.

Table 5.1 Minimums of conditional variance

	NTP	2008	2007	2006	2005	2004	2003	2002
CEZ	1.898	1.555	1.558	1.522	1.524	1.551	1.397	1.683
PX	**0.848**	**0.848**	0.591	0.472	0.431	0.478	0.481	0.633
CZK/EUR	0.195	0.110	0.441	0.488	0.622	0.052	0.678	0.091

5.2 Prediction of the conditional variance

Prediction of TGARCH (1,1) model in the negative turning point was made for CEZ stocks. Frances and Dijk [5] suggested forecasts of TGARCH model. Assuming ε_t distribution is symetric around 0, we can construct prediction of conditional variance (horizon h, in time T):

TGARCH (1,1):

$$\sigma_T^2(h) = \omega \sum_{i=0}^{\sigma^2-2} \left((\alpha_1 + \gamma_1)/2 + \beta_1\right)^i + \left((\alpha_1 + \gamma_1)/2 + \beta_1\right)^{\sigma^2-1} \sigma_{T+1}^2 \tag{5.1}$$

Forecast of the conditional variance is shown in Figure 5.2. Estimation of conditional variance from real data is shown in following Figure 5.3. The comparison of graphs 5.2 and 5.3 is noticeable that the prediction of conditional variance TGARCH was distorted in the turning point. For this reason it is appropriate to be cautious when forecasting at points close to global minimum of the conditional variance.

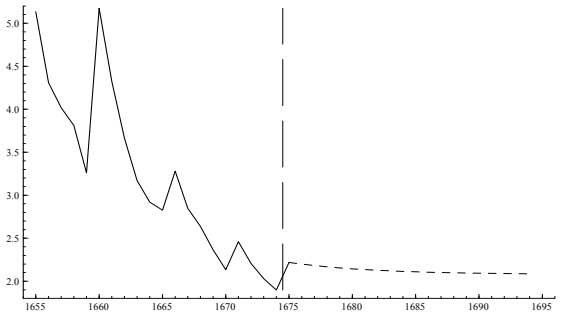


Fig. 5.2 Forecast of the conditional variance of TGARCH(1,1) model - CEZ stocks around NTP (-20;+20)

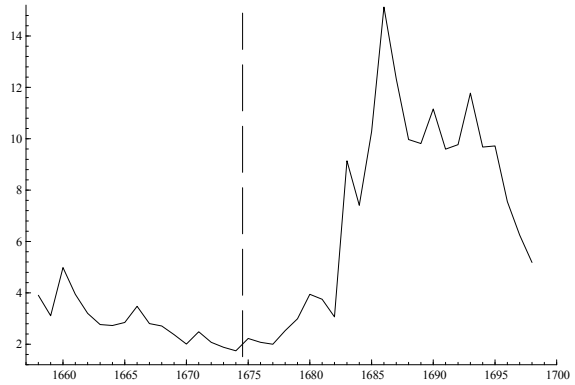


Fig. 5.3 Conditional variance of TGARCH(1,1) model - CEZ stocks – NTP + 20 values

6 Conclusion

First part was focused on the models estimate during 2002-2010. For CEZ and PX were selected as the best non-linear models of volatility. The TGARCH models confirmed different effects of positive and negative returns on the conditional variance. The TGARCH models were statistically better than the EGARCH models. Time series of CZK/EUR were modeled by linear IGARCH. Parameters of non-linear model were not statistically significant for CZK/EUR. These models can by used for forecasts.

In second part were finded turning points and modeled parts of time series. Structural changes were tested. Coefficients in all models are consistent with coefficients in the groups generated by separating the time series by turning points.

In final part was analyzed process of the conditional variance before and after turning points. Time series were estimated by volatility models until the turning points. The indicator minimum of the conditional variance were tested. For selected turning points were not found significant difference in the values of the conditional variance. Assumption was confirmed only for PX index in 2008. Did not prove conclusively that the downturn in 2008 was preceded by minimums of the conditional variance (for selected models).

Subsequently, the prediction based on volatility models was made for next 20 values. Further, the back test was made for these forecasts. It is appropriate to be cautious when forecasting at points close to global minimum (maximum) in the conditional variance.

References

[1] ARLT, J., ARLTOVÁ, M.: *Ekonomické finanční časové řady*, 1. vyd. Praha: Grada 2007
[2] ARLT, J., ARLTOVÁ, M.: *Finanční časové řady*, 1. vyd. Praha: Grada 2003
[3] BOLLERSLEV, T., CHOU, R., KRONER, K.: *ARCH modeling in finance*, Journal of Econometrics 52 (1992)
[4] ENGLE, R.F., NG,V.K.: *Measuring and Testing the Impact of News on Volatility*, Journal of Finance (1993)
[5] FRANCES, P.H., van DIJK, D.: *Non-linear Time Series Models in Empirical Finance*, Cambridge: Cambridge University Press (2000)
[6] GREGORY C. Chow (1960): *Tests of Equality Between Sets of Coefficients in Two Linear Regressions*. Econometrica 28(3): 591–605
[7] ZAKOIAN, M. (1994): *Threshold Heteroscedastic Models*, Journal of Economic Dynamics and Control, 18, 931-955

Current address

Žižka David, Ing.
University of Economics Prague,
W.Churchill Square 4, 130 67 Prague, Czech Republic,
e-mail: xzizd02@vse.cz