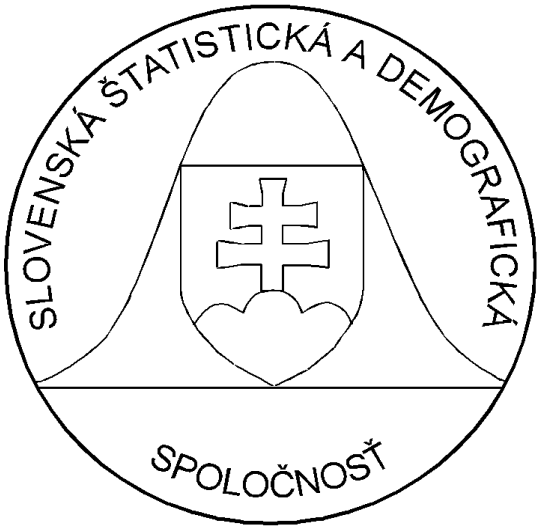
7/2012

FORUM STATISTICUM SLOVACUM



ISSN 1336-7420

Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava
www.ssds.sk

Naše najbližšie akcie:

(tiež na www.ssds.sk, blok Organizované akcie)

Slávnostná konferencia ku 45. výročiu založenia SŠDS
20. marec 2013, Sládkovičovo

14. Slovenská demografická konferencia
20. - 22. marec 2013, Sládkovičovo

IV. MedStat
4. - 6. apríl 2013, Ružomberok

Pohľady na ekonomiku Slovenska 2013
16. apríl 2013, Bratislava, Aula EU

6. Nitrianske štatistické dni
máj 2013, UKF Nitra

EKOMSTAT 2013 – 27. škola štatistiky
19. – 24. máj 2013, Trenčianske Teplice

PRASTAN 2013
27. - 29. máj, 2013

Aplikácie metód na podporu rozhodovania 2013
október 2013, STU Bratislava

FERNSTAT 2013
október 2013, Banská Bystrica

22. Medzinárodný seminár Výpočtová štatistika
5. – 6. december 2013, Bratislava

Prehliadka prác mladých štatistikov a demografov
5. – 6. december 2013, Bratislava

Regionálne akcie
priebežne

INTRODUCTION

Dear colleagues,

the eighth issue of the ninth volume of the scientific peer-reviewed journal published by the Slovak Statistical and Demographic Society (SSDS) is composed of contributions that are in their content compatible with the topics covered by the 21. international seminar Computational statistics 2012 and Rewiev works of Young statisticians and demographers. These actions were held on 6th-7th December 2012. These actions were organized by Slovak Statistical and Demographical Society in collaboration with Faculty of Management Comenius University in Bratislava, Statistical Office of the SR, SAS Slovakia, s. r. o. and Club Dispersus.

The event was organised by Program and Organizational committee: Assoc Prof. Dr. Iveta Stankovičová - president, Dr. Ján Luha. – scientific secretary, Assoc. Prof. Dr. Jozef Chajdiak, Prof. Dr. Beáta Stehlíková, Assoc. Prof. Dr. Bohdan Linda, Assoc. Prof. Dr. Jana Kubanová, Assoc. Prof. Dr. Vladimír Úradníček, Dr. Samuel Koróny, Dr. Tomáš Želinský, Lukáš Pastorek, MA, Dr. Tomáš Löster, Dr. Jitka Bartošová, Dr. Alena Tartaľová.

There was an interdisciplinary set of lectures organised within the Seminar for the Young Statisticians and Demographers. The set of lectures "Insights into Analytics – Analytics perceived by Professionals" was given by the presenters from private, public and academic sector. This event was held on December 7, and was organised in cooperation with Club Dispersus.

Preparation and editing of this FSS issue were performed by: Assoc. Prof. Dr. Iveta Stankovičová, Dr. Tomáš Želinský and Dr. Ján Luha, CSc.

We would also like to thank to the reviewers of papers published in this issue: Assoc. Prof. Dr. Michal Greguš, Dr. Ján Luha, Dr. Martin Řezáč, Assoc. Prof. Dr. Iveta Stankovičová, Dr. Tomáš Želinský, Assoc. Prof. Dr. Vladimír Úradníček.

We are very glad at the participants' interest in Computational Statistics seminar. The Board of SSDS appreciates activity of the young statisticians and demographers, which is and evidence of good work of teachers and their students. We hope that the possibility to present contributes to improvement of scientific level of young statisticians and demographers.

Editorial Board of FSS

ÚVOD

Vážené kolegyne, vážení kolegovia,

siedme číslo ôsmeho ročníka vedeckého recenzovaného časopisu Slovenskej štatistickej a demografickej spoločnosti (SŠDS) je zostavené z príspevkov, ktoré sú obsahovo orientované v súlade s tematikou 21. ročníka medzinárodného seminára Výpočtová štatistika 2012 a Prehliadkou prác mladých štatistikov a demografov. Tieto akcie sa uskutočnili v dňoch 6. a 7. decembra 2012 v kongresovej sále ŠÚ SR na Hanulovej ul. 5/c v Bratislave. Organizátorom seminára a prehliadky bola SŠDS v spolupráci s Fakultou managementu UK v Bratislave, Štatistickým úradom SR, spoločnosťou SAS Slovakia, s. r. o. a klubom Dispersus.

Akcie, z poverenia výboru SŠDS, zorganizoval organizačný a programový výbor v zložení: doc. Ing. Iveta Stankovičová, PhD. - predsedníčka, RNDr. Ján Luha, CSc. – vedecký tajomník, doc. Ing. Jozef Chajdiak, CSc., prof. RNDr. Beáta Stehlíková, CSc., doc. RNDr. Bohdan Linda, CSc., doc. Dr. Jana Kubanová, CSc., doc. Ing. Vladimír Úradníček, PhD., RNDr. Samuel Koróny, PhD., Ing. Tomáš Želinský, PhD., Mgr. Lukáš Pastorek, Ing. Tomáš Löster, Ph.D., RNDr. Jitka Bartošová, Ph.D, Mgr. Alena Tartaľová, PhD.

V spolupráci s klubom Dispersus pri Prehliadke prác mladých štatistikov a demografov bolo zorganizované 7. decembra 2011 aj interdisciplinárne pásmo prednášok prezentátorov zo súkromnej, štátnej i akademickej sféry pod názvom: *„Pohľady do analytiky - Analytika očami profesionálov“*.

Na príprave a zostavení tohto čísla FSS participovali: doc. Ing. Iveta Stankovičová, PhD., Ing. Tomáš Želinský, PhD. a RNDr. Ján Luha, CSc.

Recenziu príspevkov zabezpečili: doc. RNDr. Michal Greguš, PhD., RNDr. Ján Luha, CSc., Mgr. Martin Řezáč, PhD., doc. Ing. Iveta Stankovičová, PhD., Ing. Tomáš Želinský, PhD., doc. Ing. Vladimír Úradníček, PhD.

Veľmi nás teší neustály záujem o seminár Výpočtová štatistika. Výbor SŠDS oceňuje aktivitu mladých v rámci Prehliadky prác mladých štatistikov a demografov, čo svedčí tiež o dobrej práci pedagógov a ich študentov. Dúfame, že možnosť prezentácie príspevkov na tomto podujatí sa podieľa na zvyšovaní odbornej úrovne mladých štatistikov a demografov.

Redakčná rada FSS

Pravdepodobnostné rozdelenie miery rizika chudoby v EÚ pomocou programu EasyFit
Probability distribution of the risk of poverty in EU using EasyFit

Jana Bednáriková, Beáta Stehlíková

Abstract: Knowledge of probability distribution of the phenomenon is statistically important. EasyFit is a small but powerful program that allows to find a suitable probability distribution of empirical data to work with him. Its use is applied to finding the probability the distribution the risk of poverty in the EU at the NUTS 2.

Abstrakt: Poznať pravdepodobnostné rozdelenie skúmaného javu je zo štatistického hľadiska veľmi dôležité. Práve pomocou rozdelenia je náhodná premenná jednoznačne určená. EasyFit je malý, ale výkonný program umožňujúci hľadať vhodné pravdepodobnostné rozdelenie empirických údajov a pracovať s ním. Jeho využite je aplikované na nájdení pravdepodobnostného rozdelenia miery rizika chudoby v EÚ na úrovni NUTS 2.

Key words: EasyFit, risk of poverty, European Union

Kľúčové slová: EasyFit, riziko chudoby, Európska únia

JEL classification: C88, F21, I32, R20

Úvod

Prehlbujúcou sa globalizáciou sa zvýšil aj záujem o skúmanie príjmovej nerovnosti. Napriek tomu, že pojem chudoby je ťažké merať, pretože je založená na subjektívnych pocitoch jedinca, vo svete existuje úsilie chudobu merať. Ekonomické definície majú dva spoločné rysy. Prvým je určenie indikátora blahobytu a druhým určenie hranice, ktorá vymedzuje, že človek, pre ktorého hodnota indikátora nadobúda hodnotu pod touto hranicou je jednotlivec ktorý je, považovaný za chudobného. Svetová banka (2001) zverejnila definíciu chudoby. Zásadnou otázkou však vždy zostáva vzťah medzi nízkymi príjmami a schopnosťou človeka žiť určitý spôsob života. Existuje súbor štandardov, podľa ktorých sa určuje, či sú príjmy a životné podmienky najchudobnejších v spoločnosti sú prijateľné, alebo nie.

1. Materiál a metódy

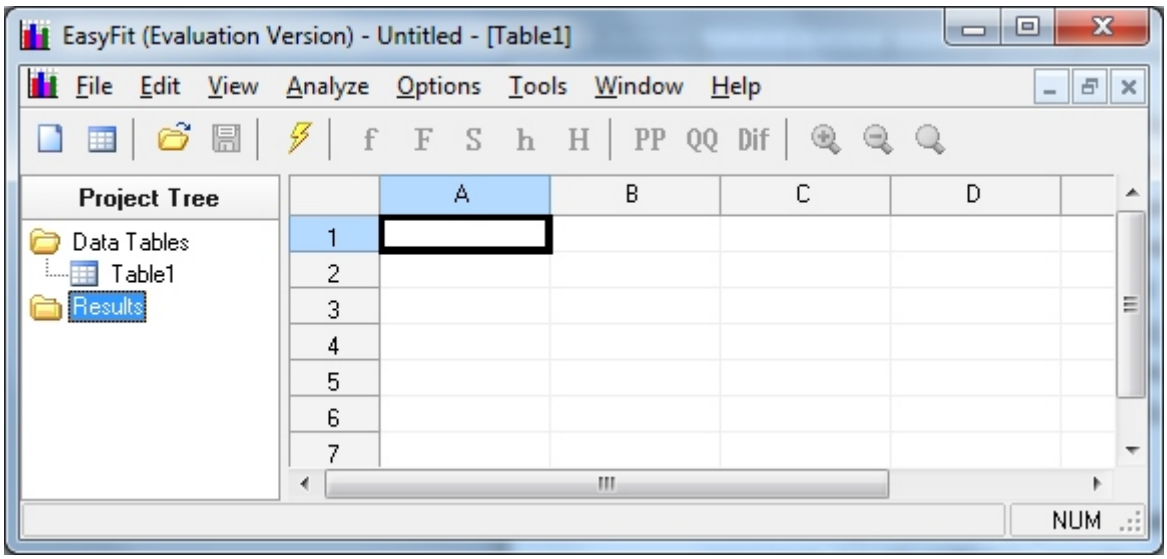
V Slovenskej republike pojem chudoba nie je legislatívne definovaným pojmom. Príspevok je zameraný na identifikáciu pravdepodobnostného rozdelenia medzinárodne akceptovateľného a porovnateľného indikátora chudoby, ktorý vychádza z ekvivalentného disponibilného príjmu definovaného v zisťovaní EU SILC. Ekvivalentný disponibilný príjem domácností, je disponibilný príjem domácnosti pred sociálnymi transfermi vydelený ekvivalentnou veľkosťou domácnosti. Pre výpočet ekvivalentnej veľkosti domácnosti sa v zisťovaní EU SILC požíva tzv. modifikovaná OECD škála, na základe ktorej je každému prvému dospelému členovi domácnosti priradený koeficient 1, každému druhému a ďalšiemu dospelému členovi domácnosti a 14-ročným a starším osobám koeficient 0,5 a každému dieťaťu mladšiemu ako 14 rokov koeficient 0,3. Takto vypočítaný ekvivalentný disponibilný príjem domácnosti je následne priradený každej osobe v rámci domácnosti. Za mieru rizika chudoby budeme v súlade s metodikou SILC považovať podiel osôb s ekvivalentným disponibilným príjmom pod 60% mediánu národného ekvivalentného príjmu.

Údaje za roky 2008 a 2010 na úrovni NUTS 2 boli čerpané z databázy Eurostatu a EU SILC. Pri výpočtoch bol použitý softvér EasyFit.

2. Výsledky a diskusia

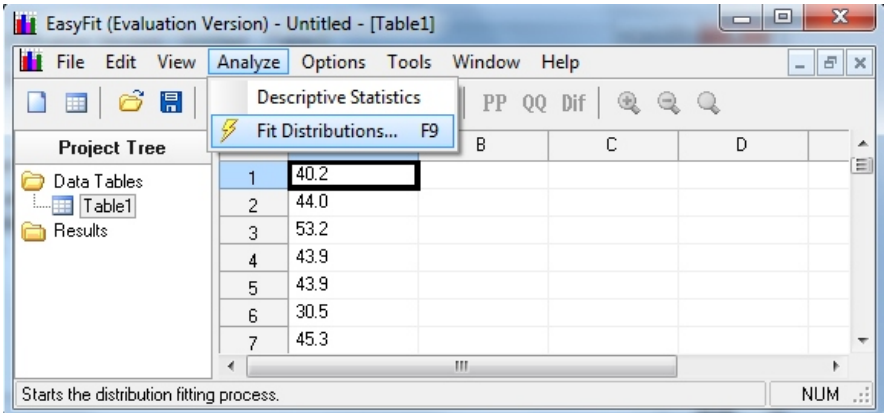
Poznať pravdepodobnostné rozdelenie skúmaného javu je zo štatistického hľadiska veľmi dôležité. Je dôležité poznať odhad priemeru, smerodajnej odchýlky a ďalších charakteristík. Ale oveľa dôležitejšie je poznať pravdepodobnostné rozdelenie. Veď práve pomocou neho je náhodná premenná jednoznačne určená. EasyFit je malý, ale výkonný a komplexný program umožňujúci hľadať vhodné pravdepodobnostné rozdelenie a pracovať s ním. Je užívateľský príjemný. EasyFitXL sa dá navyše pridať do Excelu ako excelovský doplnok.

Úvodná obrazovka programu je na Obrázku 1. Údaje môžeme načítať zo súboru alebo priamo nakopírovať do tabuľky. Počet riadkov v trial verzii je až 5000.



Obr. 1: Úvodná obrazovka programu EasyFit

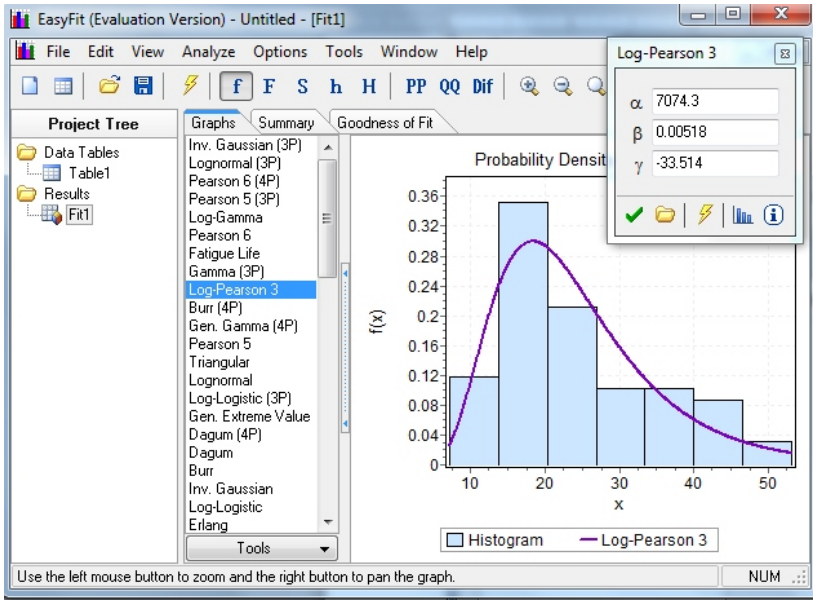
Údaje môžeme načítať zo súboru alebo priamo nakopírovať. Počet riadkov v trial verzii je až 5000. Po zadaní údajov stačí zvoliť príkaz pre hľadanie vhodného rozdelenia (Analyze -> Fit distribution) a ako výsledok tohoto kroku nám program ponúkne výber premennej (v našom prípade Var1) a voľbu medzi spojitými a diskrétnymi premennými. Naša voľba je hľadať medzi spojitými pravdepodobnostnými rozdeleniami (Data Domain: Continuous).



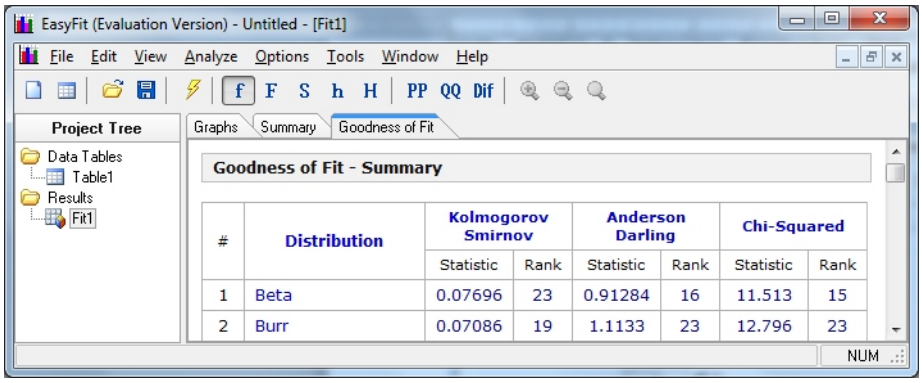
Obr. 2: Výber z ponuky - hľadanie vhodného pravdepodobnostného rozdelenia

Následne sa nám zobrazia v záložke funkcie hustoty až šesťdesiatich (v trial verzii) spojitých pravdepodobnostných rozdelení v spojení s histogramom našich údajov. V záložke

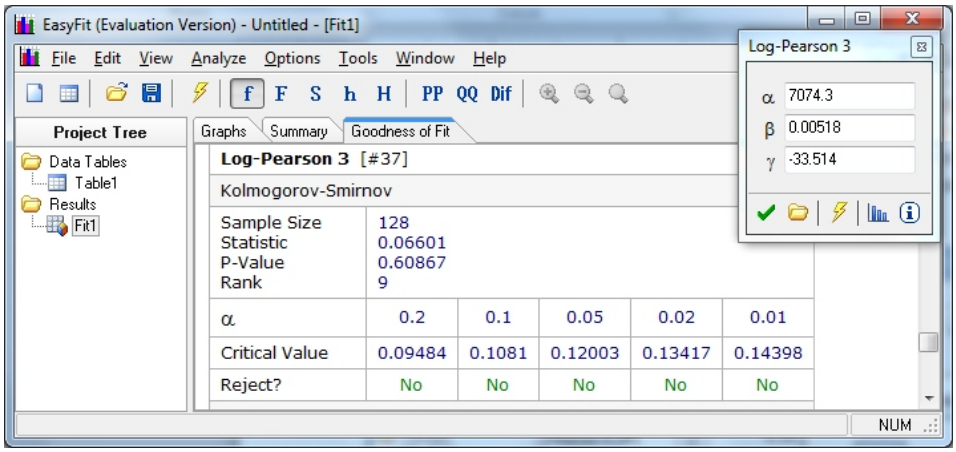
Summary sú uvedené odhady parametrov zvoleného rozdelenia. Súčasne sa ukážu aj v hornej časti obrazovky. V záložke Goodness of Fit sú výsledky (hodnota testovacej štatistiky a poradie vhodnosti) troch testov dobrej zhody – Kolmogorovho – Smirnovovho, Andersonovho-Darlingovho a Chí kvadrát testu. (Obrázok 4). Je to z toho dôvodu, že každý z uvedených testov je viac alebo menej citlivý na odchýlku empirických údajov od teoretického rozdelenia. Po kliknutí na názov testu sa pravdepodobnostné rozdelenia zoradia podľa poradia, ako sa umiestnili v danom teste.



Obr. 3: Ukážka funkcie hustoty Log-Pearsonovho rozdelenia a histogramu empirických dát vrátane odhadu jeho parametrov

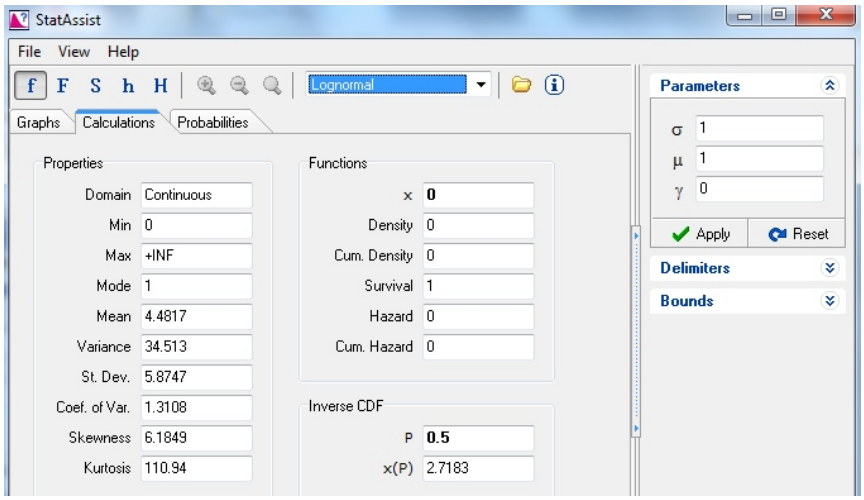


Obr. 4: Súhrn výsledkov testov dobrej zhody troch testov pre jednotlivé teoretické rozdelenia

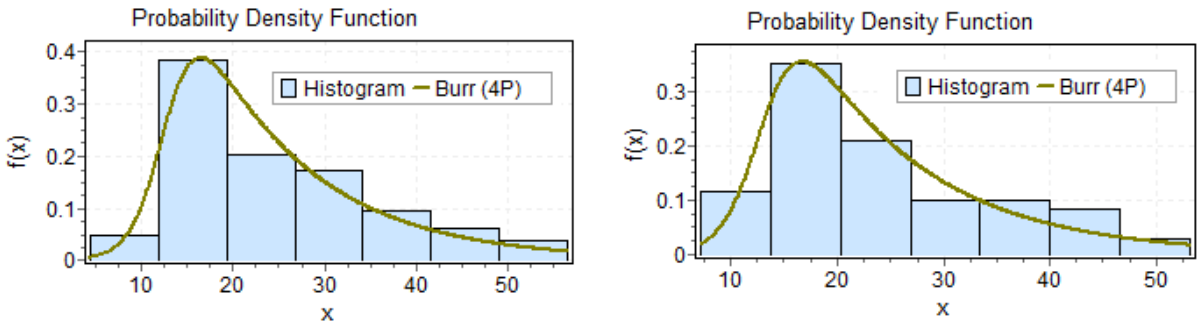


Obr. 5: Výsledok Kolmogorovho-Smironovovho testu dobrej zhody pre Log-Pearsonovo rozdelenie

Po kliknutí na názov pravdepodobnostného rozdelenia sa zobrazí rozsah výberu (Sample Size), hodnota testovacej štatistiky (Statistic), P hodnota (P-Value) a poradie (Rank) pre daný test (v našom prípade Kolmogorovov-Smirnovov test). Ďalšia tabuľka obsahuje hladiny významnosti (α), kritickú hodnotu pre danú hladinu významnosti (Critical Value) a rozhodnutie (Reject ?), či nulovú hypotézu o zhode empirického a teoretického rozdelenia (v našom prípade Log-Pearsonovho 3)na zvolenej hladine významnosti zamietame (Yes), resp. nemôžeme zamietnúť (No). Okrem samotného testu dobrej zhody , program umožňuje grafické znázornenia PP grafov (PP), QQ grafov (QQ), rozdielov pravdepodobností (Dif). Umožňuje tiež vykresliť funkciu hustoty (f), distribučnej funkcie (F), funkciu prežitia (S), rizikovú fukciu (h) a kumulatívnu rizikovú fukciu (H). Užívateľ iste ocení informáciu pod ikonkou i. Obsahuje okrem popisu parametrov, definičného oboru, funkciu hustoty, distribučnú funkciu, príkazy na odhad priemeru, smerodajnej odchýlky, modusu a ďalších. Program je prepojený s programom StatAssist. Tieto charakteristiky sa dajú znázorniť v záložke výpočty (Calculations) (Obrázok). V záložke Pravdepodobnosti (Probabilities) môžeme počítať hodnoty distribučnej funkcie v danom bode, hodnotu pravdepodobnosti, že náhodná premenná nadobúda hodnoty zo zvoleného intervalu a mnohé ďalšie.



Obr. 6: Popisné charakteristiky pre dané teoretické rozdelenie, jeho parametre a možnosť výpočtu hodnoty funkcie hustoty, distribučnej funkcie, funkcie prežitia a rizikovej funkcie vo zvolenom bode x



Obr. 7: Burrovo štvorparametrické rozdelenie pre modelovanie miery rizika chudoby v EÚ v rokoch 2008 a 2010 na úrovni NUTS 2

Spoločné vhodné rozdelenie pre oba hodnotené roky 2008 a 2010 sa ukázalo štvorparametrické Burrovo rozdelenie s kladnými parametrami α, β, k a lokujúcim parametrom γ, definované na intervale (γ, ∞) s funkciou hustoty

$$f(x) = \frac{\alpha k \left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\beta\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{k+1}}.$$

Odhady parametrov za roky 2008 a 2010 sú nasledovné:

2008: $k = 0{,}15364, \alpha = 99701{,}0, \beta = 1{,}8663.10^5, \gamma = -1{,}8662.10^5$,

2010: $k = 0{,}16157, \alpha = 2{,}8073.10^8, \beta = 5{,}2686.10^8, \gamma = -5{,}2686.10^8$.

Vhodnosť štvorparametrického Burrovho rozdelenia potvrdili v oboch rokoch všetky tri testy. Výsledky sú zhrnuté v tabuľke 1.

Tab. 1: Testy dobrej zhody miery rizika chudoby v EÚ

Rok	Test	Testovacia štatistika	P-hodnota	Kritická hodnota $\alpha = 0{,}01$	Kritická hodnota $\alpha = 0{,}05$
2008	Kolmogorov Smirnovov	0,06606	0,60776	0,14398	0,120003
	Andersonov-Darlingov	0,63389	0,63389	3,9074	2,5018
	Chí kvadrát test	5,9115	0,43317	16,812	12,592
2010	Kolmogorov Smirnovov	0,06662	0,06662	0,14398	0,120003
	Andersonov-Darlingov	0,64966	0,64933	3,9074	2,5018
	Chí kvadrát test	5,9185	0,43238	16,812	12,592

Zdroj: Vlastné výpočty z údajov Eurostatu

Interpretáciou zmien treba byť veľmi opatrný. Zmeny môžu byť nielen dôsledkom zlepšenej situácie na trhu práce, ale aj zmenou demografických ukazovateľov, zmien sociálnej politiky a iných ukazovateľov. Netreba tiež zabúdať, že pracujeme s podielmi osôb pod 60 percentnou hranicou mediánu národného ekvivalentného príjmu, ktorý je v jednotlivých štátoch EÚ rôzny. V roku 2010 je rozdelenie viac zošikmené doľava - pribudlo území s nižšou mierou rizika, ale aj území s hodnotou rizika chudoby na 40 percent. Smerodajná odchýlka sa mierne zvýšila z hodnoty 12,607 percent na 13,544 percent.

Záver

Výskyt chudoby v silných a vyspelých ekonomikách môže byť signálom zlyhania riadenia a vládnutia v spoločnosti. Preto sa mnohí snažia tento problém pomenovať a nájsť vysvetlenie, nájsť riešenie tohto problému, alebo aspoň nájsť nástroje na jeho predchádzanie. Jedným z možných príspevkov je nájsť pravdepodobnostné rozdelenie, akým sa riadi. Druhým výsledkom je poukázanie a upozornenie na program EasyFit, ktorý umožní nájsť pravdepodobnostné rozdelenie aj iných ukazovateľov a tak prispeje k lepšiemu spoznaniu ďalších závažných skutočností dnešnej doby.

Poďakovanie

Príspevok bol napísaný s podporou Vedeckej grantovej agentúry MŠ SR a SAV v rámci riešenia vedecko-výskumného projektu *VEGA 1/0127/11 Priestorová distribúcia chudoby v EÚ*.

Literatúra

[1] EASYFIT. Dostupné na internete. 29.10.2012 http://www.mathwave.com/

[2] EUROSTAT.EU. *People at risk of poverty or social exclusion by NUTS 2 regions*. Dostupné na internete. [cit. 29.10.2012]. URL: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_peps11&lang=en

[3] WORLD BANK. *Poverty Manual* (Chapter 1: The Concept of Poverty and Well-Being***).*** Washington DC: The World Bank, 2001.

Adresa autora (-ov):

Jana Bednáriková, Ing.
FEP Paneurópska vysoká škola
Tematínska 10
851 05 Bratislava
janka.bednarikova@gmail.com

Beáta Stehlíková, prof. RNDr. CSc.
FEP Paneurópska vysoká škola
Tematínska 10
851 05 Bratislava
stehlikovab@gmail.com

Minimum variance portfolio: A comparison of robust and classic approach

Portfólio s minimálnou disperziou: porovnanie robustného a klasického prístupu

Martin Boďa

Abstract: In the paper empirical behaviour of minimum variance portfolio selection is explored in the case when input parameters are estimated either by classic statistical methods or by their robust variants.

Key words: portfolio selection; minimum variance portfolio; classical method; robust method; fast minimum covariance determinant estimator.

Abstrakt:V článku sa skúma empirické správanie úlohy výberu portfólia s minimálnou disperziou v prípade, keď vstupné parametre sa odhadujú klasickými štatistickými metódami alebo ich robustnými variantmi.

Kľúčové slová: výber portfólia, portfólio s minimálnou disperziou, klasická metóda, robustná metóda, rýchly MCD estimátor.

JEL classification: G11.

Introduction

This paper is an empirical exercise into the effect of using robust estimates of the mean vector and the covariance matrix in the process of portfolio choice that is oriented on selecting the minimum variance efficient portfolio. The task is to choose, in the set of all mean-variance efficient portfolios (i. e. portfolios that achieve the highest expected return at a given level of risk as expressed or measured by standard deviation / volatility), the portfolio with the lowest attainable standard deviation / volatility (or variance). However, it is customary in practical application of this optimization approach to employ classical estimates of statistical inputs necessary for the optimization task: the mean vector of asset returns of which the portfolio is to be composed and their covariance matrix (Kanderová, 2007, 2011). In the paper as a competitive approach to utilization of classical statistical inputs in the portfolio selection of this task utilization of robust statistical estimates of both the mean vector and the covariance matrix are considered. The focus is only restricted to detecting the influence of a particular decision to employ robust estimates instead of classic ones on portfolio selection, and the possible effect on portfolio performance is left aside and not evaluated as it is more complex for practical considerations. Out of a selection of 10 stocks represented in the S&P 500 Index (each stock chosen at random from one of the ten Global Industry Classification Standards (GICS) sectors) in a sliding manner a minimum variance efficient portfolio is selected for each business day of the period from 2009/09/30 up to 2012/10/01 using both classical statistical inputs and robust statistical inputs updated in the same sliding basis, and the differences in the composition of selected portfolios are marked and explored. As might be expected, it was found that there has been a substantial difference stemming from using different methods of estimating necessary statistical inputs. However, no attention is paid to performance evaluation of selected portfolios in individual business days and their comparisons which is open to further investigations and improvement of the approach.

Regardless of this introduction and the final concluding section, the paper consists of two core sections. In the following, methodological, section details on the portfolio selection

problem and on the estimation of input parameters are clarified and the design of the conducted empirical investigation is presented. The third section gives an overview of the empirical results that were obtained in the study.

1. Methodological issues

Suppose that n risky asset returns are represented by a random vector $\mathbf{R} = (R_1, ..., R_n)'$ that have an expectation $\mathtt{m} = (\mu_1, ..., \mu_n)'$ and an $n \times n$ covariance matrix $S = (\Sigma_{ij})_{n \times n}$ (the diagonal elements Σ_{ii} are variances σ_i^2 of individual returns and non-diagonal elements are respective covariances). Assume for now that the both the expectation $\mathtt{m}$ and the covariance matrix S are known. Any portfolio Π with a set of n weights $w = (\omega_1, ..., \omega_n)'$ that decide allocation of available financial funds across individual risky assets has expected return $w'\mathtt{m}$, variance $w'Sw$ and standard deviation (volatility) $\sqrt{w'Sw}$. All attainable portfolios are represented by coordinates $[w'\mathtt{m}, \sqrt{w'Sw}]$ in the Cartesian plane, the first coordinate is given by the expected return of a given portfolio whilst the second represents its standard deviation (volatility). The two-dimensional (expected return × standard deviation) space of portfolios generated (spanned) by risky assets is called frequently the mean-variance space and it can be shown that it is in the form of a hyperbole which intersects the Cartesian plane and determines the set of all attainable portfolios (c. f. Prigent, 2007, pp. 73-74). The upper arc of this hyperbole is made up of those portfolios that attain the highest expected return possible at the given level of risk expressed by standard deviation / volatility. These portfolios are addressed as efficient in the sense of Markowitz (also known as Markowitz-efficient or mean-variance efficient) portfolios and form the efficient frontier. The portfolio with the minimum standard deviation (volatility) is called the minimum variance portfolio (or the mean variance portfolio). Denote a vector of n ones by $\mathbf{1}$ and introduce the following quantities

$$A = \mathbf{1}'S^{-1}\mathtt{m}, \quad B = \mathtt{m}'S^{-1}\mathtt{m}, \quad C = \mathbf{1}'S^{-1}\mathbf{1} \quad \text{and} \quad D = BC - A^2. \tag{1}$$

When shortselling is allowed and there are no constraints save the weights must sum to one, the coordinates of the minimum variance portfolio in the mean-variance space are then given by

$$[1/\sqrt{C}, \quad A / C], \tag{2}$$

and the vector of weights $w^{\#}$ of its allocation across the n risk assets is given by

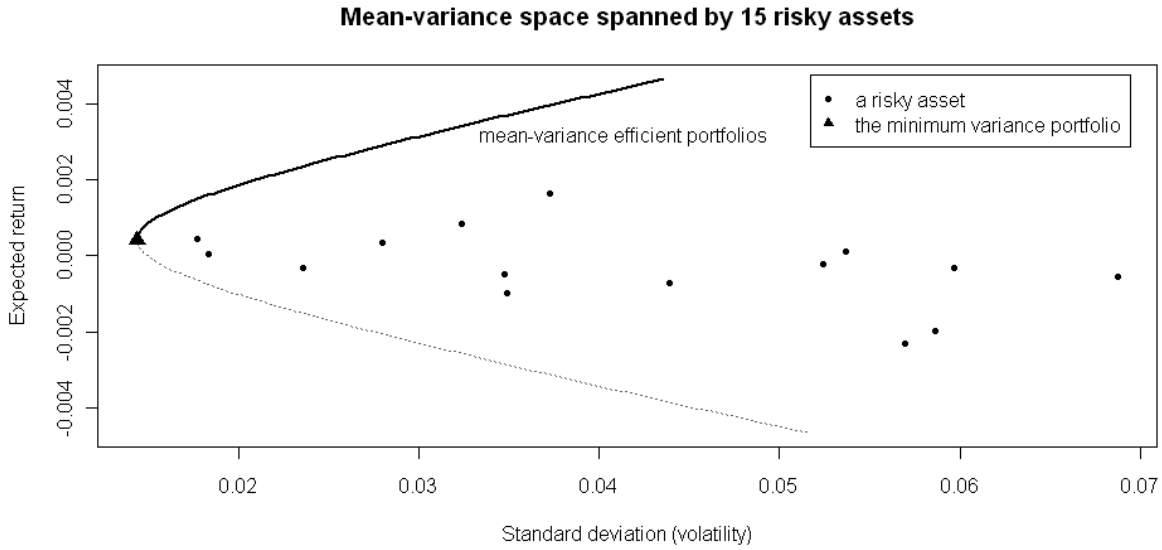
$$w^{\#} = D^{-1}(BS^{-1}\mathbf{1} - AS^{-1}\mathtt{m}) - A(CD)^{-1}(AS^{-1}\mathbf{1} - CS^{-1}\mathtt{m})) \tag{3}$$

(see Prigent, 2007, pp. 72-73).

Graph 1 displays an illustration of mean-variance space generated by 15 risky assets (indicated by the dot symbols). Two arcs of the hyperbole demark the set of all attainable portfolios composed of the 15 risky assets under consideration and the upper bold-line arc distinguishes the efficient frontier. The triangle symbol highlights the minimum variance portfolio.

Naturally, in practical applications it is necessary to estimate the expectation vector $\mathtt{m}$ and the covariance matrix S and these estimates are formed and computed out of time series of historical observations on n asset returns. Ordinarily, the expectation vector $\mathtt{m}$ is estimated by simple or (exponentially) weighted averaging of individual historical asset returns and the covariance matrix S by an unbiased estimator or by an (exponentially) weighted estimator. Since these estimators are well known and used by default, their description is omitted here. As alternative to this, these inputs may be estimated by some robust procedure. In the paper, they are estimated by the fast Minimum Covariance Determinant estimator (MCD) proposed

by Rousseeuw and van Driessen (1999). The MCD method yields a robust multivariate location and scale (covariance matrix) estimate with a high breakdown point. The primal idea of the MCD estimator is to trim available data points on the basis of the squared Malahanobis distance from their center and to use classical estimators of location and scale (covariance matrix) on the set of those observations that are closest to the center of all data points. The MCD estimate of location (the expectation vector `m`) is obtained as the average of the trimmed data points and the MCD estimate of scale (the covariance matrix S) is obtained as their classic covariance matrix multiplied by a consistency factor and a finite sample bias correction factor. This procedure is explained in detail in the original article by Rousseeuw and van Driessen (1999) or in Maronna et al. (2006, pp. 189-190).



Graph 1 An illustration of mean-variance space (the source: the author)

In the empirical exercise, out of the stocks represented in the S&P 500 Index one stock was chosen randomly by each of the 10 GICS sectors and this selection of 10 stocks were available for portfolio selection. The sample of these stocks is displayed in Table 1.

Table 1: The stocks participating in the empirical exercise (the source: the author)

GICS sector	Consumer Discretionary	Consumer Staples	Energy	Financial	Health Care
Stock (& ticker)	Lennar Corp. (LEN)	ConAgra Foods Inc. (CAG)	Denbury Resources Inc (DNR)	Capital One Financial (COF)	Quest Diagnostics (DGX)
GICS sector	Industrials	Information Technology	Materials	Telecommunications Services	Utilities
Stock (& ticker)	Avery Dennison Corp. (AVY)	Salesforce.com (CRM)	FMC Corporation (FMC)	Sprint Nextel Corp. (S)	TECO Energy (TE)

The time span of 252 business days (approx. one calendar year) and the first window from 2008/10/01 to 2009/09/30 were set. The expectation vector `m` and the covariance matrix S were estimated by both the classical estimators and the MCD estimator on the basis of 252 daily historical observations of the sampled 10 stocks. With all these inputs and allowing short positions, the composition of an optimal portfolio utilizing classic estimates of the expectation vector `m` and the covariance matrix S and their robust counterparts was determined with help of (3). The window of 252 business days slid on by one business day and the procedure was repeated with updated estimates until 2012/10/01. The portfolio selection procedure was thus repeated for each method of estimate provision 758 times, in

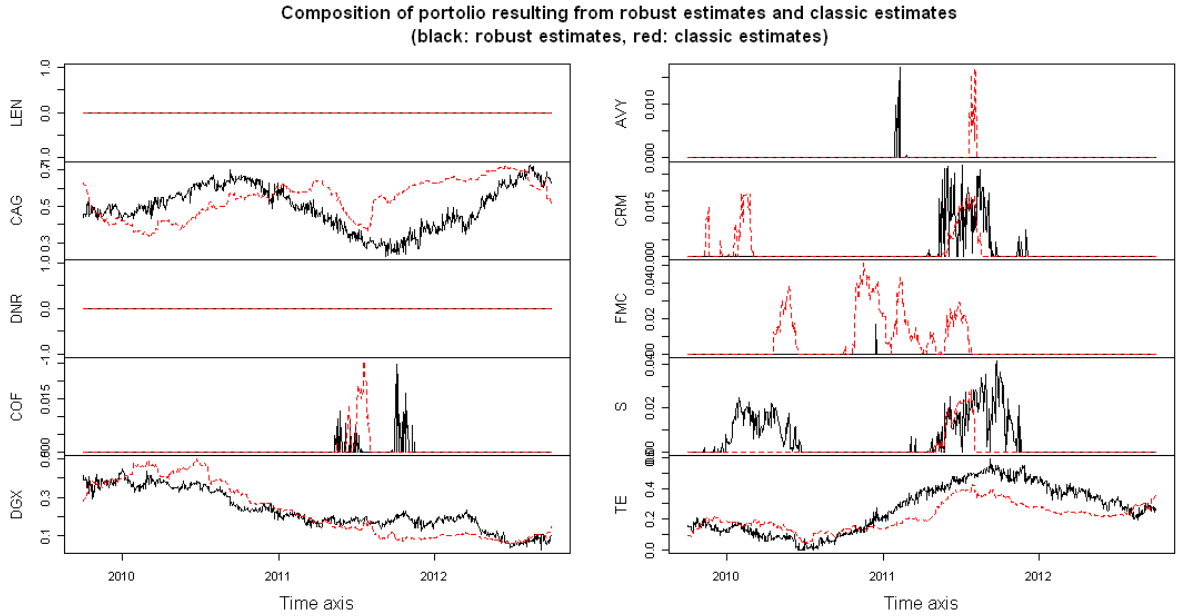
consequence of which a total of $2 \times 758 = 1516$ portfolios were determined. Their compositions are evaluated in the next section.

Please note that using daily data might imply the investment horizon of one business day but this need not be true. It is possible to convert daily estimates of the expectation vector `m` and to scale daily estimates of the covariance matrix S into a longer time horizon (say 1 year), but the results on the composition of the portfolio would remain the same. Another point is that it is not expedient and reasonable to evaluate performance of these portfolio selections on the basis of the type of estimates used. If the investment horizon were, say one year, the composition of the portfolio would be over this one-year horizon be an issue of revaluation. Of course, it is not possible to adjust the portfolio every day for this would incur immense transactional costs. Yet, there are many circumstances on which this would be inevitable as confronting the current composition of the portfolio with new information on financial markets.

2. Empirical results

In computations and preparing graphical presentations, the software R (R Development Core Team, 2012) was employed and several of its libraries, PerformanceAnalytics (Carl et al., 2012), timeSeries (Wuertz and Chalabi, 2012) and fPortfolio (Rmetrics Core Team and Wuertz, 2011).

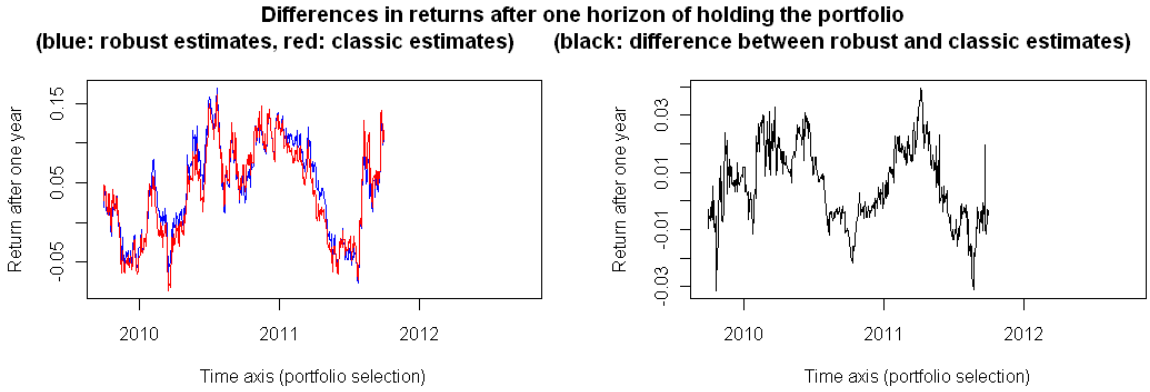
Empirical results are summarized, for convenience, in the form of graphs. Graph 2 presents individual compositions of selected portfolios when using robust input estimates and classic input estimates, and Graph 3 gives an impression on the return of selected portfolios after a one-year holding period (precisely, 252 business days), provided the composition of individual portfolios remained intact.



Graph 2 The composition of portfolios over the out-of-sample period (the source: the author)

It is important to remark that notwithstanding the fact that short sales were allowed, they were chosen in the procedure. On the whole, the composition of portfolios when using robust estimates and the composition of portfolios when using classic estimates do not share their evolution pattern over time. With each stock there are long periods when the compositions with both type of estimates perfectly match and this is only when there are zero

weights on the given stock. This may be due to the fact that in some cases (when there are silent periods on financial markets) the MCD estimator when trimming historical observations in fact does little trimming and yields the same or comparable estimates as do the traditional estimators. This is also indicated by almost perfect matching the portfolio return after a simulated one-year holding horizon without readjustment. Graph 2 additionally manifests another property of portfolio selection based on different estimates: even the weights of individual stocks with robust estimates do not appear to be stable and immune to sudden perturbations or distractions. All though there are no apparent advantages of using robust estimates when inspecting Graph 4 that might overturn possible computational complications associated with using robust estimators since returns after one-year holding of selected portfolios are practically identical, it still may be safer to prefer robust estimators in order to achieve a higher stability if this maximization principle were employed on a daily basis with intention of recomposing the portfolio.



Graph 3 Differences in portfolio returns after one year of holding (the source: the author)

Conclusion

In the background of the paper was an attempt to find out whether robustness may be of avail in portfolio selection. Of course, given the extent of the empirical exercise contained in the paper this cannot be decided and there is no intention to answer this complex question on the ground of modest results that are submitted in the paper. In the paper effect of using robust parameter estimates in minimum variance portfolio selection was evaluated in comparison to using classic estimates. However, the design of study is only restricted to a random selection of ten S&P 500 Index components, each being a representative of a different GICS sector; for each business day over the period from 2009/09/30 to 2012/10/01 the portfolio minimizing standard deviation / volatility was chosen in two variants: by use of the robust MCD estimates of input parameters and by use of their classic estimates. It has been found that robustness need not be a vital requirement for practical portfolio selection as the use of robust estimates did not yield special benefits projected in higher returns of selected portfolios.

The financial support of the grant scheme VEGA No. 1/0765/12 Research into possibilities and perspectives of employing traditional and alternative approaches in financial management and financial decision-making in the changing economic environment is kindly appreciated.

References

[1] PRIGENT, J. L. 2007. *Portfolio Optimization and Performance Analysis*. Boca Raton, FL: Chapman & Hall/CRC, 2007. ISBN 1-58488-578-5.

[2] FABOZZI, F. J., FOCARDI, S. M., KOLM, P. N. *Financial Modeling of the Equity Market: from CAPM to Cointegration*. Hoboken, NJ: Wiley, 2006. ISBN 0-471-6990-4.

[3] KANDEROVÁ, M. 2011. Alternatívne kritéria výberu portfólia investičných nástrojov. *Forum Statisticum Slovacum*, vol. III, iss. 7/2011, pp. 84-89. ISSN 1336-7420.

[4] KANDEROVÁ, M. 2007. Optimalizácia portfólia investičných nástrojov v prostredí MS Excel. *Forum Statisticum Slovacum*, vol. III, iss. 6/2007, pp. 68-73. ISSN 1336-7420.

[5] ROUSSEEUW, P. J. DRIESSEN VAN, K. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, vol. 41, pp. 212-223, August 1999.

[6] MARONNA, R. A. ET AL. 2006. *Robust Statistics: Theory and Methods*. Chichester, UK: Wiley, 2006. ISBN 0-470-01092-4.

[7] R DEVELOPMENT CORE TEAM: *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2012, http:// www.r-project.org/.

[8] CARL, P. et al. 2012. *PerformanceAnalytics: Econometric tools for performance and risk analysis*. R package, version 1.0.4.4 of 2012-03-31, http://cran.r-project.org/web/ packages/PerformanceAnalytics/index.html.

[9] WUERTZ, D., CHALABI, Y. 2012. *timeSeries: Rmetrics - Financial Time Series Objects*. R package, version 2160.95 of 2012-08-07, http://cran.r-project.org/web/packages/ timeSeries/index.html.

[10] RMETRICS CORE TEAM, WUERTZ, D. *fPortfolio: Rmetrics - Portfolio Selection and Optimization*. R package, version 2130.80 of 2011-02-10, http://cran.r-project.org/web/ packages/fPortfolio/index.html

The author's address

Martin Boďa, Ing. et Bc.
Univerzita Mateja Bela v Banskej Bystrici
Ekonomická fakulta
Katedra kvantitatívnych metód a informačných systémov
Tajovského 10, 975 90 Banská Bystrica
martin.boda@umb.sk

Integrovanie a teória pravdepodobnosti na časových škálach
Integration and probability theory on time scales

Eva Brestovanská

Abstract: In this article, I modify some concepts of probability theory on time scales. Delta probability function has been constructed on time scales. I pay a special attention the general formulae for delta integration some elementary functions. The main goal this article is to establish the basics of probability theory on time scales and apply those results. I collect studies mainly from the work of Bohner and Peterson [1, 2] and the dissertation Thomas Matthews [4].

Abstrakt: V úvode tohto článku sú uvedené príklady, na ktorých je vysvetlený pojem miery zrnitosti. Špeciálnu pozornosť venujem delta integrácii elementárnych funkcii. Hlavným cieľom článku je ukázať možnosti využitia časovej škály (miery zrnitosti. delta integrácie) v teórii pravdepodobnosti. V závere práce uvádzam definíciu delta miery, delta pravdepodobnosti, delta distribučnej funkcie s príkladmi. Teoretická časť bola spracovaná na základe práce Bohnera a Petersona [1, 2] (miera zrnitosti, delta integrácia) a práce Thomasa Matthewsa [4] (delta pravdepodobnosť, delta distribučná funkcia).

Key words: delta integration, delta probability function, time scales, graininess

Kľúčové slová: delta integrácia, delta pravdepodobnostná funkcia, časová škála, zrnitosť

JEL classification: C02

1. Úvod

V roku 1988, Stefan Hilger predstavil koncept časovej škály (time scales) vo svojej dizertačnej práci. Základné vety boli rozpracované v knihe Martina Bohnera a Allana Petersona [1], [2]. Vznikla dôsledku potreby zjednotiť dve samostatné teórie - diskrétnu a spojitú analýzu, ktoré boli skúmané doteraz samostatne. Delta miera ako aj delta pravdepodobnosť bola prvý krát definovaná v práci [6] Guseinova v roku 2003 a neskôr rozpracovaná v práci [7] Guseinova a Bohnera v roku 2003.

2. Miera zrnitosti, predný $s(t)$ a spätný $r(t)$ operátor skoku

Definícia 1.: [1] Časová škála (time scales) je ľubovoľná neprázdna uzavretá podmnožina reálnych čísel R, v ďalšom texte označovaná písmenom T.

Dva najvýznamnejšie príklady časovej škály sú R (reálne), Z (celé čísla), N (prirodzené čísla), N_0 (prirodzené čísla vrátane nuly), [1, 2] U [5, 6] (zjednotenie dvoch uzavretých intervalov. Množiny Q (racionálne čísla), R\Q (iracionálne čísla) a (0,3) (otvorený interval) nie sú príklady časovej škály T.

Definícia 2.: [1] Pre $t \in T$ definujeme predný a spätný operátor skoku $\sigma(t)$; $\rho(t)$: $T \to T$, $\sigma(t) := \inf \{s \in T: s > t\}$, $\rho(t) := \sup \{s \in T: s < t\}$.

Definícia 3.: [1] Nech $\sigma(t)$ je predný a $\rho(t)$ je spätný operátor skoku a $t \in T$. Ak $\sigma(t)>t$, potom t je sprava riedky bod; $\rho(t)< t$, potom t je zľava riedky bod; $\sigma(t)=t$, potom t je sprava hustý bod; $\rho(t)=t$, potom t je zľava hustý bod, $\rho(t)<t<\sigma(t)$, potom t je izolovaný bod; $p(t)=t=\rho(t)$, potom t je hustý bod.

Definícia 4.: [1] Zobrazenie $\mu : T \to [0,\infty)$ nazveme funkciou zrnitosti, ktorá je definovaná vzťahom $\mu(t) := \sigma(t) - t$, alebo $\mu(t) := t - \rho(t)$ pre $t \in T$.

Definícia 5.: [1] Nech T je časová škála, funkcia f: T→R a t∈T. Ak existuje $\lim_{s\to t\, s\in T} [f(\sigma(s))-f(t)] / [\sigma(s)-t]$ t.j. $\lim_{s\to t\, s\in T} [f(\sigma(s))-f(t)] / \mu(t)$ nazývame túto limitu delta deriváciou funkcie f v bode t a označujeme ju $f^{\Delta}(t)$.

Príklad 1.: Úlohou je vyjadriť mieru zrnitosti μ(t), predný σ(t) a spätný ρ(t) operátor pre časové škály R (reálne čísla), Z (celé čísla), h*Z .

Riešenie: Pre T=R: μ(t)=0, σ(t)=t, ρ(t)=t; pre T=Z: μ(t)=1, σ(t)=t+1, ρ(t)=t-1; pre T=h*Z: μ(t)=h, σ(t)=t+h, ρ(t)=t-h.

Príklad 2.: V tomto príklade je ukážka riešenia problému čiastočne spojitého a čiastočne diskrétneho charakteru. Nech N je množstvo predaných lyžiarskych permanentiek v sezóne v čase t. Počas mesiacov november až marec prevádzka sa rozbieha. Na začiatku apríla sa predaj zastaví vplyvom teplého počasia, vleky prestanú fungovať, nová prevádzka začne v budúcom roku v novembri. Podobný charakter má prevádzka kúpalísk, sezónnych reštaurácii, alebo podnikov, ktoré prešli inováciou (výroba sa na čas zastaví, továreň sa začne pripravovať na nový typ výrobku).

Riešenie: [3] Najskôr zavedieme časovú škálu $T= P_{a,b}=\bigcup_{k=0}^{\infty}[k(a+b), k(a+b)+a]$ kde a,b>0.

Potom predný operátor skoku je určený vzťahom

$$\sigma(t) \begin{cases} t & \text{pre } t\in \bigcup_{k=0}^{\infty}[k(a+b), k(a+b)+a) \\ t+b & \text{pre } t\in \bigcup_{k=0}^{\infty}\{(a+b)+a\} \end{cases}$$

a funkcia zrnitosti je daná vzťahom

$$\mu(t) \begin{cases} 0 & \text{pre } t\in \bigcup_{k=0}^{\infty}[k(a+b), k(a+b)+a) \\ 1 & \text{pre } t\in \bigcup_{k=0}^{\infty}\{(a+b)+a\}. \end{cases}$$

V tomto príklade sa teda zaujímame o časovú škálu $P_{1,1} = T = \bigcup_{k=0}^{\infty}[2*k;\ 2*k+1]$; kde t=0 je 1.november tohto roku, t=1 je 1.apríl tohto roku, t=2 je 1.november budúceho roku, t = 3 apríl budúceho roku a tak ďalej. Potom mieru zrnitosti zadefinujme nasledovne:

$$\mu(t) \begin{cases} 0 & \text{ak } 2*k \le t < 2*k+1 \\ 1 & \text{ak } t = 2*k+1 \end{cases}$$

a predný operátor skoku je určený vzťahom:

$$\sigma(t) \begin{cases} t & \text{ak } 2*k \le t < 2*k+1 \\ t+1 & \text{ak } t = 2*k+1 \end{cases}$$

3. Pojem integrálu a jeho základné vlastnosti

Definícia 6.: [2, 3] Funkcia f : T → R, sa nazýva regulovaná za predpokladu, že existuje jej limita sprava vo všetkých sprava hustých bodoch a limita zľava vo všetkých zľava hustých bodoch na T.

Definícia 7.: [2, 3] Funkcia f : T → R, sa nazýva rd-spojitá za predpokladu, že je spojitá v sprava hustých bodoch a jej limita zľava existuje v zľava hustých bodoch na T, označujeme ju $C_{rd.}$

Veta 1. : [2, 3] Predpokladajme f : T → R :

(1) Ak f je spojitá, potom f je rd-spojitá;

(2) Ak f je rd-spojitá, potom f je regulovaná;

(3) Operátor "skok (σ)"je rd-spojitý;

(4) Ak f je regulovaná alebo rd-spojitá, potom aj $f(\sigma(t))$ je regulovaná alebo rd-spojitá;

(5) Predpokladajme, že f je spojitá. Ak $g : T \to R$ je regulovaná alebo rd-spojitá, potom $f \circ g$ má tie isté vlastnosti.

Definícia 8.: [2, 3] Každá rd-spojitá funkcia $f : T \to R$ má antideriváciu $F : T \to R$ $F^{\Delta}(t) = f(t)$. Ďalej definujeme neurčitý delta integrál regulovanej funkcie f ako $\int f(t)\,\Delta(t) = F(t) + c$

Definícia 9. : [2,3] Nech $f : T \to R$ je rd-spojitá funkcia, F je antiderivácia funkcie f, potom definujeme určitý delta integrál

$$\int_a^b f(t)\,\Delta(t) = F(b) - F(a) \text{ pre } a,b \in T$$

Veta 2.: [2, 3] Ak funkcia $f : T \to R$, je rd-spojitá pre $t \in T$ $\int_t^{\sigma(t)} f(s)\,\Delta s = \mu(t) * t$

Veta 3.: **[** 2, 3] Ak $a,b,c \in T$ a $f,g \in C_{rd}$, potom

(1) $\int_a^b [f(t) + g(t)]\,\Delta(t) = \int_a^b f(t)\Delta t + \int_a^b g(t)\Delta t$; (2) $\int_a^b (\alpha * f)(t)\,\Delta t = \alpha * \int_a^b f(t)\,\Delta t$; (3) $\int_a^b f(t)\Delta t = -\int_b^a f(t)\,\Delta t$

(4) $\int_a^b f(t)\,\Delta t = \int_a^c f(t)\Delta t + \int_c^b f(t)\Delta t$; (5) $\int_a^b f(\sigma(t)) * g^{\Delta}(t)\Delta t = (f*g)(b) - (f*g)(a) - \int_a^b f^{\Delta}(t) * g(t)\Delta t$;

(6) $\int_a^b f(t) * g^{\Delta}(t)\Delta t = (f*g)(b) - (f*g)(a) - \int_a^b f^{\Delta}(t) * g(\sigma(t))\Delta t$; (7) $\int_a^b f(t)\Delta t = 0$;

Veta 4.: [2, 3] Nech $a,b \in T$ a $f \in C_{rd}$

Ak T=R potom $\int_a^b f(t)\,\Delta(t) = \int_a^b f(t)\,dt$ kde integrál na pravej strane je Riemannov integrál.

Ak T=Z potom ak [a,b] obsahuje iba izolované body, potom

$$\text{potom } \int_a^b f(t)\,\Delta(t) \quad \begin{cases} \sum_{t \in [a,b)} \mu(t)\,f(t) & \text{pre } a < b \\ 0 & \text{pre } a = b \\ -\sum_{t \in [b,a)} \mu(t) f(t) & \text{pre } a > b \end{cases}$$

<u>*Príklad 3.:*</u> Nech T je ľubovoľná časová škála, vypočítajme $\int t\,\Delta(t)$ pre $t \in T$.

<u>*Riešenie.*</u> $\int t\,\Delta(t) = \int t + \sigma(t) - \sigma(t)\,\Delta(t) = t^2 - \int \sigma(t)\,\Delta(t) = t^2 - \int \mu(t) + t\,\Delta(t) = 2 * \int t\,\Delta(t) =$

$t^2 - \int \mu(t)\,\Delta(t) \Rightarrow$ ***výsledok*** $\int t\,\Delta(t) = t^2/2 - 1/2 \int \mu(t)\Delta(t)$ ***je závislý na funkcii zrnitosti*** $m(t)$.

Pre T = R $\mu(t) = 0$; $\int t\,\Delta(t) = t^2/2 - 1/2 * \int 0\,\Delta(t) = t^2/2$

Pre T = Z $\mu(t) = 1$; $\int t\,\Delta(t) = t^2/2 - 1/2 * \int 1\,\Delta(t) = t^2/2 - t/2$

Pre T = h*Z $\mu(t) = h$; $\int t\,\Delta(t) = t^2/2 - 1/2 * \int h\,\Delta(t) = t^2/2 - h*t/2$

<u>*Príklad 4.:*</u> Pre $T \in [0,1] \cup [2, 3]$ vypočítajme integrál $\int_0^t s\Delta s$ nech $t \in T$.

Riešenie.:

$$\int_0^t s\Delta s \quad \begin{cases} \int_0^t sds = t^2/2 \text{ pre } t\in[0,1] \\ \int_0^t sds - \int_1^2 sds = t^2 - 3/2 \text{ pre } t\in[2,3] \end{cases}$$

Príklad 5.: Nech T je (R, Z, h*Z), vypočítajme $\int a^t \Delta(t)$ pre $t\in T$

Riešenie.: Pre T = R; $\int a^t \Delta(t) = \int a^t \Delta(t) = \int a^t dt = a^t/\ln a$

Pre T=Z $a\neq 1$ si najskôr vyjadríme delta deriváciu výrazu pozri [2]
$(a^t)^{\Delta} = \Delta(a^t) = (a-1)\, a^t$
$\int a^t \Delta t = 1/(a-1)* \int (a-1)* a^t \Delta t = 1/(a-1)*a^t$
Pre T=h*Z si opäť najskôr vyjadríme delta deriváciu výrazu
$(a^t)^{\Delta} = \Delta (a^t) = ((a^h-1)/h)* a^t$
$\int a^t \Delta t = h/(a^h - 1)* \int((a^h-1)/h) *a^t\Delta t = h/(a^h-1)\, a^t$, opäť výsledok závisí od T.

4. Pojem delta miery a delta pravdepodobnosti

Definícia 10.: **[**4,5] Pre $t\in T$ delta miera (Δ-miera) je daná vzťahom $\mu_\Delta(t)=\sigma(t)-t$, nech $a,b\in T$; pre $a\leq b$, $\mu_\Delta([a,b))=b-a$; $\mu_\Delta((a,b))=b-\sigma(a)$; $\mu_\Delta((a,b])=\sigma(b)-\sigma(a)$; $\mu_\Delta([a,b])=\sigma(b)-a$

Definícia 11.: [4, 5] Delta pravdepodobnosť (označenie P_Δ) je reálna funkcia definovaná na časovej škále nasledovne.

$$P_\Delta(A) \quad \begin{cases} \mu_\Delta(A)/\ \mu_\Delta(W_T) \text{ pre } A\subset W_T \\ 0 \text{ Inak} \end{cases}$$

Potom platia pravidlá :

$P_\Delta(A)\geq 0$; $P_\Delta(W_T)=1$; $P_\Delta(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P_\Delta(A_i)$

Príklad 6: Nech $W_T=[0,1]\cup[2,3] \cup \ldots \cup [22,23]$ a $A = [2,3] \cup[8,9]$; vypočítajme pravdepodobnosť $P_\Delta(A)$.

Riešenie. $\mu_\Delta(W_T)= \mu_\Delta([0,1]\cup[2,3] \cup \ldots \cup [22,23])=23$
$\mu_\Delta(A)= \mu_\Delta([2,3] \cup[8,9])= \mu_\Delta([2,3]/\ \mu_\Delta(W_T) + \mu_\Delta ([2,3]/\ \mu_\Delta(W_T) = 4/23$

Príklad 7 .: Autobus stojí na stanici každý deň v čase medzi
0,1],[8,9],[12,13],[16,17],[18,19] ,[20,21],[22,23], úlohou je vypočítať $P_\Delta([1,9])$, $P_\Delta([1,9))$, $P_\Delta((1,9])$, $P_\Delta((1,9))$.

Riešenie.
$\mu_\Delta(W_T)= \mu_\Delta([0,1],[8,9],[12,13],[16,17],[18,19] ,[20,21],[22,23])=23$
$P_\Delta([1,9])= 11/23$, $P_\Delta([1,9))=8/23$, $P_\Delta((1,9])=4/23$, $P_\Delta((1,9))=1/23$.

Definícia 11.: [4,5] Nech delta distribučná funkcia (označenie F_Δ) je reálna funkcia definovaná na časovej škále na základe delta pradepodobnosti nasledovne: $F_\Delta =$

$P_\Delta(X\leq x)= \int_{-\infty}^{\infty} f_\Delta(t)\, \Delta t;$

Potom platia nasledujúce pravidlá :
F je rastúca funkcia; $\lim_{x\to-\infty} F_\Delta(x)=0$; $\lim_{x\to\infty} F_\Delta(x)=1$.

$F_\Delta = P_\Delta(X\le a)= \int_{-\infty}^{a} f_\Delta(t)\,\Delta t$, pre A=(x: X<a) a A=(x: X<b); A⊂ B; $P_\Delta(X\le a) \le P_\Delta(X\le b)$

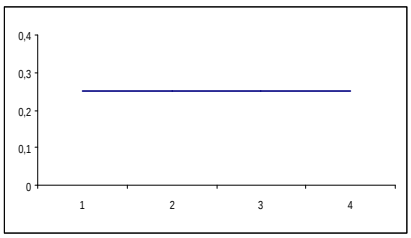
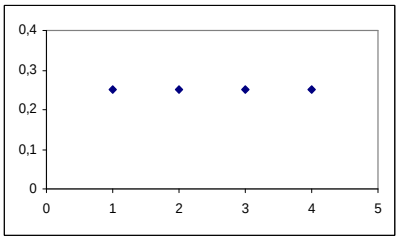
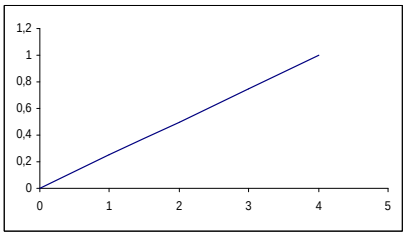
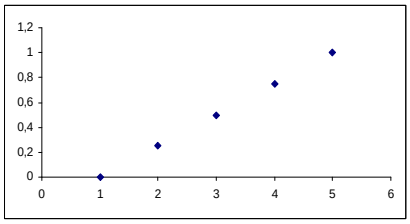
Príklad 8 .: Hustotu rovnomerného rozdelenia môžem napísať na základe delta miery následovne: Pre $T=[t_0, t_1]\cup[t_2, t_3] \cup \ldots \cup [t_{2t}, t_{2t+1}]$; $a= t_0$, $b= t_{2t+1}$;

$$f_\Delta(t) \begin{cases} 1/\sigma(b)\text{-}a & \text{ak } a \le t \le b \quad \text{t.j.} \\ 0 & \text{inak} \end{cases}$$

$$f_\Delta(t) = 1/\mu_\Delta([a,b)\} \text{ ak } t\in T \qquad 0 \text{ inak}$$

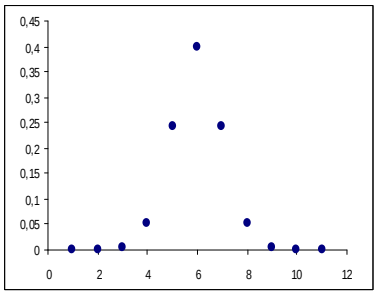
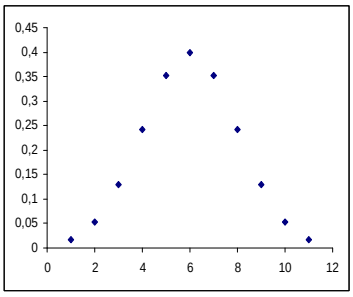
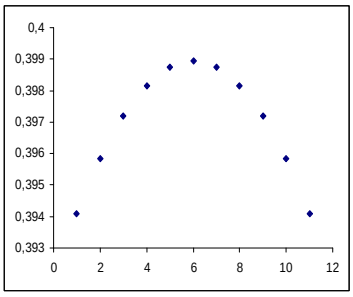
Distribučnú funkciu rovnomerného rozdelenia môžem sformulovať:

$F(x) = \int_0^x f(t)\Delta t = \int_a^x (1/(\sigma(b)\text{-}a)\,\Delta t = (x\text{-}a)/(\sigma(b)\text{-}a)$

Obrázok 1. Grafy hustoty a distribučnej funkcie rovnomerného rozdelenia pre T=R a T=Z.

Príklad 9 .: Hustotu normálneho rozdelenia môžeme napísať na základe delta miery $f_\Delta(t) = c * \prod_{t\in[0,x)} (1+\mu(t))^{-1}$ $x\in T_+$, $x \ge 0$. Ak T=h*Z pre h>0 funkcia musí spĺňať nasledujúce pravidlá: $\sum_{n=1}^{\infty} f_\Delta(t)=1$, $\mu(t)=h$, potom $f_\Delta(hn)=c\prod_{k=0}^{n-1}(1+h)^{-kh} = c*(1+h)^{-hn(1-n)/2}$ pre $n\in N_0$, po substitúcii x=hn: $f_\Delta(x) = c*[(1+h)^{1/h}]^{-x(1-h)/2} = 1$ pre $x\in T$; pre h=1 c=0,60915; pre h=1/2 c=0,35026; pre h=1/4 c=0,0966356; pre h=1/32 c=0,0248695

Obrázok 2. Grafy hustoty normálneho rozdelenia pre T=Z , T=1/2*Z , T=1/32*Z,

5. Záver

V práci sú uvedené definície miery zrnitosti, delta integrácie ako aj ich všeobecné vzťahy s príkladmi. V prvej časti sú uvedené príklady na mieru zrnitosti. V druhej časti je ukážka delta integrácie lineárnej a exponenciálnej funkcie. Na záver sú uvedené príklady uplatnenia tohto nového smeru v teórii pravdepodobnosti, zadefinovaná delta miera, delta pravdepodobnosť a delta distribučná funkcia. Význam práce spočíva v ***zjednotení teórie pravdepodobnosti pre spojitú aj diskrétnu náhodnú premennú.***

Literatúra

[1] BOHNER, M., PETERSON, A.: *Dynamic Equations on Time Scales: An Introduction with Applications*. Birkhauser, Boston, 2001.

[2] BOHNER, M., PETERSON, A.: *Advances in Dynamic Equations on Time Scales*. Birkhauser, Boston, 2003.

[3] PEKÁROVÁ, E.: *Kalkulus na časových škálach a jeho aplikace*. Brno, 2007.

[4] MATTHEWS, T.: Prob*ability theory on time scales and application to finance and ineqalities*. Missouri university of science and technology, 2011.

[5] KAHAMAN, S.: *Probability theory applications on time scales*. Izmír, 2008.

[6] GUSEINOV G.: *Riemann and Lebesgue Integration on Time Scales*. 2003.

[7] BOHNER, M., GUSEINOV G.: *Multiple Lebesgue Integration on Time Scales*. 2006.

Adresa autora :

Eva Brestovanská, Mgr., PhD
FM UK , KEF, Odbojárov 10
820 05 Bratislava 25
Eva.Brestovanska@fm.uniba.sk

Vybrané logistické modely používané pro vyrovnávání a extrapolaci křivky úmrtnosti a jejich aplikace na populace vybraných zemí Evropské unie

Selected logistic models used for leveling and extrapolate mortality curves and their application to the population of the EU countries

Petra Dotlačilová, Jana Langhamrová, Ondřej Šimpach

Abstract: Demographers are constantly trying to find a model that best described the relationship between mortality and age. In the past, for leveling and extrapolate mortality curves most used Gompertz - Makeham model. But at present it is important to develop new models because people are reaching ever higher age. The second reason is better availability of statistical data. Already in the past, was established in several other models used for leveling and extrapolate mortality curves. Currently, come to the fore logistic models. In this paper we will present selected logistic models and we will apply them to data on populations of selected EU countries. The results will be compared with results obtained using the Gompertz - Makeham, Modified Gompertz - Makeham model, mortality tables according to the Czech statistical office methodology and mortality tables without extrapolation.

Abstrakt: Demografové se neustále snaží najít model, který by co nejlépe popisoval vztah mezi úmrtností a věkem. V minulosti se pro vyrovnávání a extrapolaci křivky úmrtnosti nejvíce používal Gompertz – Makehamův model. Ale v současné době je důležité vyvíjet nové modely, protože se lidé dožívají stále vyššího věku. Druhým důvodem je lepší dostupnost statistických dat. Už v minulosti vzniklo v několik dalších modelů používaných pro vyrovnávání a extrapolaci křivky úmrtnosti. V současné době se dostávají do popředí logistické modely. V tomto článku budou představeny vybrané logistické modely a budou aplikovány na data o úmrtnosti populací vybraných zemí Evropské unie. Získané výsledky budou porovnány s výsledky získanými při použití Gompertz – Makehamova, Modifikovaného Gompertz – Makehamova modelu, s metodikou Českého statistického úřadu a s úmrtnostními tabulkami bez extrapolace.

Key words: mortality at the highest ages, logistic models, Gompertz – Makeham function, Modified Gompertz – Makeham function, mortality tables without extrapolation

Klíčová slova: úmrtnost v nejvyšších věcích, logistické modely, Gompertz – Makehamův model, Modifikovaný Gompertz – Makehamův model, úmrtnostní tabulky bez extrapolace

JEL classification: C, C1, C10

1. Úvod

V minulosti se pro vyrovnávání a extrapolaci křivky úmrtnosti nejvíce používal Gompertz – Makehamův model, ale v současné době je důležité vyvíjet další modely. Je to způsobeno především tím, že v minulosti se jen málo lidí dožilo vysokého věku. V dnešní době je situace odlišná. Zvyšuje se úroveň lékařské péče a roste zájem o zdravý životní styl. To způsobuje, že se dnes lidé mohou dožít vyššího věku než jejich předci. Vzhledem ke zlepšování úmrtnostních poměrů se stále více ukazuje, že Gompertz – Makehamův model bude potřeba nahradit nějakým jiným modelem. V současné době se pro vyrovnávání a extrapolaci křivky úmrtnosti nejvíce používají logistické modely. Je však třeba uvědomit si, že logistické modely patří mezi ty optimističtější. Při jejich použití dostaneme vyšší naději dožití než v případě Gompertz – Makehamova modelu.

2. Teoretická část

Pro vyrovnávání specifických úmrtností v nejvyšších věcích je možné použít hned několik již existujících modelů. Po dlouhou dobu se nejvíce požíval Gompertz – Makehamův model. V současné době se naopak upřednostňují logistické modely. V našem příspěvku jsme se zaměřili na dva z nich (tj. na Thatcherův a Kannistův logistický model). Získané výsledky budeme porovnávat s výsledky podle metodiky ČSÚ, s nadějí dožití získanou z úmrtnostních tabulek bez extrapolace a s dosud nejpoužívanějším Gompertz – Makehamovým (resp. Modifikovaným Gompertz – Makehamovým) modelem. Při výpočtu úmrtnostních tabulek podle metodiky ČSÚ nejprve vypočteme empirické hodnoty specifické míry úmrtnosti podle vzorce (ČSÚ, 2012):

$$m_{t,x} = \frac{M_{t,x}}{\overline{S}_{t,x}}, \tag{1}$$

kde $M_{t,x}$ je počet zemřelých x – letých v roce t,

$\overline{S}_{t,x}$ je střední stav počtu žijících x – letých v roce t.

V dalším kroku vyrovnáme empirické hodnoty specifických měr úmrtností nejprve pomocí klouzavých průměrů. V závislosti na věku použijeme různé typy vyrovnání. Pro věk 1 a 2 budou hodnoty vyrovnaných specifických měr úmrtnosti shodné s empirickými hodnotami pro tentýž věk. Pro věky od 3 do 59 – ti let použijeme vyrovnání pomocí klouzavých průměrů (*Fiala.,* 2005):

- vyrovnání ze tří hodnot:

$$\tilde{m}_x^{(3)} = \frac{m_{x-1} + m_x + m_{x+1}}{3}, \qquad x \in <3;5> \tag{2}$$

- vyrovnání z devíti hodnot:

$$\begin{aligned} \tilde{m}_x^{(9)} &= 0{,}2.m_x + 0{,}16.(m_{x-1} + m_{x+1}) + 0{,}12.(m_{x-2} + m_{x+2}) \\ &+ 0{,}08.(m_{x-3} + m_{x+3}) + 0{,}04.(m_{x-4} + m_{x+4}) \end{aligned} \quad x \in <6;29> \tag{3}$$

- vyrovnání z devatenácti hodnot:

$$\begin{aligned} \tilde{m}_x^{(19)} &= 0{,}2.m_x + 0{,}1824.(m_{x-1} + m_{x+1}) + 0{,}1392.(m_{x-2} + m_{x+2}) + \\ &+ 0{,}0848.(m_{x-3} + m_{x+3}) + 0{,}0336.(m_{x-4} + m_{x+4}) - 0{,}0128.(m_{x-6} + m_{x+6}) - \\ &- 0{,}0144.(m_{x-7} + m_{x+7}) - 0{,}0096.(m_{x-8} + m_{x+8}) - 0{,}0032.(m_{x-9} + m_{x+9}) \end{aligned} \quad x \in <30;59> \tag{4}$$

Od 60 – ti do 82 let použijeme Gompertz – Makehamův model (*Thatcher et al.,* 1998):

$$m_x = a + b.c^x, \qquad x \in <60;82>, \tag{5}$$

kde m_x je intenzita úmrtnosti,

a, b a c jsou parametry modelu,

x je věk.

A pro věk od 83 let do 110 – ti let použijeme modifikovaný Gompertz – Makehamův model (*Thatcher et al.,* 1998):

$$m_x = a + b.c^{x_0+\frac{1}{g}.\ln[g.(x-x_0)+1]}, \qquad x \in <83;110> \quad , \tag{6}$$

kde $x>x_0$, x_0 je věk od kterého provádíme vyrovnání pomocí modifikovaného Gompertz – Makehamova modelu,

a, b, c a g jsou parametry modelu.

Použití Modifikovaného Gompertz – Makehamova modelu zohledňuje fakt, že v nejvyšších věcích už nelze přírůstky úmrtnosti s rostoucím věkem považovat za konstantní. Naopak velikost přírůstků se postupně snižuje.

Pro náš příspěvek jsme si z již existujících logistických modelů vybrali Thatcherův a Kannistův model. Při použití zmíněných modelů získáme vyšší naději dožití. Vybrané modely se řadí mezi optimističtější.

Thatcherův model (*Thatcher et al.,* 1998; *Boleslawski & Tabeau*, 2001):

$$m_x = \frac{z}{1+z} + g \ , \tag{7}$$

kde $z = a.e^{b.x}$, a, b a g jsou parametry modelu.

Thatcherův model předpokládá logistický průběh křivky úmrtnosti.

Kannistův model (*Thatcher et al.,* 1998; *Boleslawski & Tabeau*, 2001):

$$m_x = \frac{e^{[\Theta_0+\Theta_1.(x-80)]}}{1+e^{[\Theta_0+\Theta_1.(x-80)]}}, \qquad \text{pro } x \geq 80 \tag{8}$$

kde Θ_0, Θ_1 jsou parametry modelu, které nabývají nezáporných hodnot, m_x je intenzita úmrtnosti ve věku x.

Kannistův model je speciálním případem logistické funkce, kde logitová transformace měr úmrtnosti je vyjádřená jako lineární funkce věku.

Oba zmíněné logistické modely jsme použili pro extrapolaci křivky úmrtnosti pro stejné věkové rozmezí (tj. pro věk od 60 – ti do 85 – ti let).

V další části příspěvku jsme se zabývali výpočtem úmrtnostních tabulek bez extrapolace. K výpočtu byl použit algoritmus pro výpočet úplných úmrtnostních tabulek.

V poslední části jsme pro vyrovnávání specifických měr úmrtnosti v nejvyšších věcích použili již dříve zmíněný Gompertz – Makehamův a Modifikovaný Gompertz – Makehamův model. Oba tyto modely jsme použili pro vyrovnání specifických měr úmrtnosti mezi věky 60 a 85 let. Stejné věkové rozpětí jsme použili i u zmíněných logistickým modelů z důvodu lepší porovnatelnosti výsledků.

3. Praktická část

Pro praktickou aplikaci byla použita data z roku 2009 pro pět vybraných členských zemí Evropské unie. Pro analýzu byly vybrány tyto země: Belgie, Bulharsko, Česká republika, Řecko a Švédsko. Výsledky jsou publikovány zvlášť pro muže a pro ženy.

Výsledky získané při použití zmíněných metod jsme uspořádali do tabulek. Jako ukázku výpočtů jsme vybrali střední délku života v České republice pro muže a pro ženy.

Tab. 5: Naděje dožití v přesném věku – Česká republika - muži

Model	Naděje dožití v přesném věku - Česká republika - muži									
	0	**15**	**20**	**50**	**65**	**80**	**85**	**90**	**95**	**100**
Gompertz	**74,2**	59,6	54,7	26,5	15,2	7,1	5,2	3,7	2,6	1,8
Gompertz-Makeham	**74,2**	59,6	54,7	26,5	15,2	6,7	4,7	3,2	2,1	1,4
Kannistö	**74,4**	59,8	54,9	26,8	15,4	7,2	5,5	4,1	3,1	2,4
Thatcher	**74,2**	59,6	54,8	26,6	15,3	6,8	4,9	3,5	2,6	1,9
Úmrtnostní tabulka ČSÚ	**74,2**	59,6	54,7	26,6	15,3	6,8	4,8	3,4	2,4	1,7
Úmrtnostní tabulka bez extrapolace	**75,0**	60,6	55,7	27,4	15,8	7,2	5,1	3,7	3,3	6,5

Zdroj: vlastní výpočty

Tab. 6: Naděje dožití v přesném věku – Česká republika – ženy

Model	Naděje dožití v přesném věku - Česká republika - ženy									
	0	**15**	**20**	**50**	**65**	**80**	**85**	**90**	**95**	**100**
Gompertz	**80,3**	65,6	60,7	31,6	18,5	8,0	5,5	3,7	2,3	1,4
Gompertz-Makeham	**80,2**	65,5	60,6	31,5	18,4	7,6	5,1	3,2	1,9	1,1
Kannistö	**80,4**	65,7	60,8	31,7	18,6	8,2	5,9	4,1	2,9	2,1
Thatcher	**80,3**	65,6	60,7	31,6	18,5	7,8	5,3	3,5	2,4	1,8
Úmrtnostní tabulka ČSÚ	**80,3**	65,7	60,7	31,7	18,6	7,8	5,4	3,7	2,5	1,7
Úmrtnostní tabulka bez extrapolace	**81,2**	66,7	61,8	32,6	19,4	8,5	5,9	4,2	3,5	6,1

Zdroj: vlastní výpočty

4. Závěr

Ze získaných výsledků můžeme usuzovat, že Kannistův model patří mezi optimistické modely. Pokud budeme porovnávat získané výsledky s metodikou ČSÚ, tak zjistíme, že nejblíže je Thatcherův model.

Při porovnání získaných výsledků s úmrtnostními tabulkami bez extrapolace zjistíme, že nejblíže je Kannistův model.

Názory na vhodnost použití jednotlivých modelů se pravděpodobně budou v budoucnu měnit. Vše bude záviset na budoucím vývoji populace. A také na kvalitě poskytovaných statistických dat.

V současné době dávají demografové přednost logistickým modelů, které patří mezi ty optimističtější (*Gavrilov – Gavrilova*, 2011). Mezi nejfrekventovanější patří model Kannista.

5. Literatura

[1] BOLESLAWSKI, LECH, TABEAU, EWA 2001. Comparing Theoretical Age Patterns of Mortality Beyond the Age of 80. In: Tabeau, Eva, van den Berg Jeths, A. a Heathcote, Ch. (eds.) 2001. *Forecasting Mortality in Developed Countries: Insights from a Statistical, Demographic and Epidemiological Perspective*. s. 127 – 155. ISBN 978-0-7923-6833-5.

[2] BURCIN, BORIS, TESÁRKOVÁ, KLÁRA A ŠÍDLO, LUDĚK: "Nejpoužívanější metody vyrovnávání a extrapolace křivky úmrtnosti a jejich aplikace na českou populaci." *Demografie* 52, 2010: 77 – 89.

[3] ČSÚ 2012. 19. 1. 2012.
<http://www.czso.cz/csu/redakce.nsf/i/umrtnostni_tabulky_metodika>

[4] EUROSTAT. 4. 11. 2012. <http://ec.europa.eu/eurostat>

[5] FIALA, TOMÁŠ: *Výpočty aktuárské demografie v tabulkovém procesoru.* Praha: Vysoká škola ekonomická v Praze, 2005. ISBN 80-2450821-4.

[6] GAVRILOV, LEONID A., GAVRILOVA, NATALIA S.: "Mortality measurement at advanced ages: a study of social security administration death master file." *North American actuarial journal* 15 (3): 432 – 447.

[7] GAVRILOV, LEONID A., GAVRILOVA, NATALIA S.: "Stárnutí a dlouhověkost: Zákony a prognózy úmrtnosti pro stárnoucí populace." *Demografie* 53, 2011: 109 – 128.

[8] HUMAN MORTALITY DATABASE. 23. 8. 2012. <www.mortality.org>

[9] KOSCHIN, FELIX: "Jak vysoká je intenzita úmrtnosti na konci lidského života?" *Demografie* 41 (2), 1999: 105 – 109.

[10] PAVLÍK, ZDENĚK, KALIBOVÁ, KVĚTA: *Monohojazyčný demografický slovník.* Praha: Česká demografická společnost, 2005

[11] THATCHER, ROGER A., KANISTÖ, VÄINÖ A VAUPEL, JAMES W. 1998.: *The Force of Mortality at Ages 80 to 120.* 1998. ISBN 87-7838-381-1.

Příspěvek byl zpracován v rámci projektu VŠE IGA 29/2011 „Analýza stárnutí obyvatelstva a dopad na trh práce a ekonomickou aktivitu".

Adresa autorů

Petra Dotlačilová, Ing.
VŠE v Praze, katedra demografie
Nám. W. Churchilla 4, 130 67 Praha 3
xdotp00@vse.cz

Jana Langhamrová, Bc.
VŠE v Praze, katedra demografie
Nám. W. Churchilla 4, 130 67 Praha 3
jana.langhamrova@vse.cz

Ondřej Šimpach, Ing.
VŠE v Praze, katedra demografie
Nám. W. Churchilla 4, 130 67 Praha 3
ondrej.simpach@vse.cz

Jaká migrace by zajistila optimální vývoj populace České republiky?

Extent of migration ensuring optimal development of the population of the Czech Republic

Tomáš Fiala, Jitka Langhamrová, Martina Miskolczi

Abstract: Many demographers are asking if foreign migration is able to offset low fertility and to prevent the population decrease and population ageing. It has been demonstrated that the annual number of immigrants needed to restrict the decline in the population is relatively small, while the number of immigrants preventing population ageing had to be unrealistically high. (See, eg. Burcin, Drbohlav, Kučera, 2007.) The article shows, what extent of migration would be necessary to stabilise the size and sex-and-age structure of the population of the Czech Republic in the sense of the convergence to the modified stationary population model. The analysis is based on the calculation of the population projection with one variant for the development of mortality and various variants of the development of fertility and migration.

Abstrakt: Řada demografů si klade otázku, zda je zahraniční migrace schopna kompenzovat nízkou plodnost a zabránit úbytku obyvatelstva a stárnutí populace. Ukazuje se, že roční počet imigrantů potřebný k zamezení úbytku obyvatelstva je relativně malý, zatímco počty imigrantů potřebné k zabránění stárnutí populace by musely být nerealisticky vysoké. (Viz např. Burcin, Drbohlav, Kučera, 2007.) Článek ukazuje, jaký objem migrace by byl potřebný ke stabilizaci velikosti a věkové a pohlavní struktury obyvatelstva České republiky ve smyslu konvergence k modifikovanému modelu stacionárního obyvatelstva. Analýza je založena na výpočtu populační projekce s jednou variantou vývoje úmrtnosti a různými variantami vývoje plodnosti a migrace.

Key words: Population ageing, population projection, stationary population, replacement migration, the Czech Republic.

Klíčová slova: Stárnutí populace, projekce obyvatelstva, stacionární populace, náhradová migrace, Česká republika.

JEL classification: J11, F22.

Úvod

Za v jistém smyslu optimální (či „přirozený“) populační vývoj lze pro ekonomicky vyspělé země považovat takový vývoj, kdy nedochází k velkým změnám počtu obyvatel ani se příliš nemění demografická struktura. Pokud by byla úmrtnost stabilní, tj. nedocházelo by ke změnám specifických měr úmrtnosti a k dalšímu růstu délky života, bylo by z hlediska dalšího populačního vývoje optimální, aby byly roční počty živě narozených konstantní. Demografická struktura populace by pak byla totožná se strukturou stacionární populace vypočtené na základě odpovídajících úmrtnostních tabulek. Tato populace by měla neměnnou velikost i strukturu, nedocházelo by ke stárnutí populace.

V nejbližších desetiletích je však realističtější předpokládat, že úmrtnost bude dále klesat, střední délka života poroste. Je logické, že při takovém vývoji nelze zabránit stárnutí populace jinak než trvalým zvyšováním ročních počtů narozených resp. ročních počtů imigrantů, tj. trvalým zvyšováním počtu obyvatel. Tento vývoj by z dlouhodobého hlediska nebyl trvale udržitelný.

1. Model stacionární populace a jeho modifikace

Pokud bude nadále klesat úmrtnost, bude se rovněž měnit model stacionární populace, který je určen právě úrovní úmrtnosti. Demografické složení stacionární populace o celkové velikosti *S* odpovídající úmrtnosti v roce *t* lze určit pomocí vzorců (viz např. Roubíček, 1997)

$$S_{t,x}^{(M)} = k_t \cdot (1-d) \cdot L_{t,x}^{(M)}, \qquad S_{t,x}^{(Ž)} = k_t \cdot d \cdot L_{t,x}^{(Ž)}, \quad ; \tag{1}$$

kde

δ je podíl děvčat při narození,

$L_{t,x}^{(M)}$, resp. $L_{t,x}^{(Ž)}$ je tabulkový počet žijících mužů (resp. žen) pro rok *t* ve věku *x* a

$$k_t = \frac{S}{(1-d) \cdot T_{t,0}^{(M)} + d \cdot T_{t,0}^{(Ž)}}, \tag{2}$$

přičemž $T_{t,0}^{(M)}$, resp. $T_{t,0}^{(Ž)}$ je tabulkový počet let života mužů (resp. žen) přesného věku 0 pro rok *t*, tj. $T_{t,0}^{(M)} = \sum_{x=0}^{v-1} L_{t,x}^{(M)}$, resp. $T_{t,0}^{(Ž)} = \sum_{x=0}^{v-1} L_{t,x}^{(Ž)}$.

Hodnotu koeficientu k_t lze proto vyjádřit pomocí váženého průměru střední délky života mužů a žen

$$k_t = \frac{S}{l_{0\cdot} \cdot \left[(1-d) \cdot e_{t,0}^{(M)} + d \cdot e_{t,0}^{(Ž)}\right]} \tag{3}$$

Počet živě narozených v této populaci je pak roven $N_t = k_t \cdot l_0$, tj.

$$N_t = \frac{S}{(1-d) \cdot e_{t,0}^{(M)} + d \cdot e_{t,0}^{(Ž)}}, \tag{4}$$

je tedy nepřímo úměrný střední délce života.

2. Modelová projekce stacionární populace ČR

Výpočty všech projekcí byly prováděny komponentní metodou (viz např. Bogue, Arriaga, Anderton, 1993). Velikost stacionární populace se předpokládala 10 500 000, podíl děvčat při narození δ=0,485 po celé období projekce. Budeme předpokládat, že střední délka života mužů i žen v ČR po celé období projekce poroste, roční nárůst střední délky života mužů i žen se však bude plynule snižovat, současně se bude zmenšovat rozdíl mezi střední délkou života mužů a žen. V roce 2011 byla střední délka života mužů 74,69 roku, střední délka života žen pak 80,74 roku. Do roku 2100 se předpokládá nárůst střední délky života u mužů na 88,4 roku, žen na 92,4 roku. Viz Tab. 1.

Tab. 1: Předpokládaný vývoj střední délky života

Pohlaví	2011	2020	2030	2040	2050	2060	2070	2080	2090	2100
Muži	74,7	76,8	78,9	80,8	82,6	84,1	85,5	86,7	87,6	88,4
Ženy	80,7	82,7	84,7	86,5	88,1	89,6	90,8	91,9	92,7	93,4

Zdroj: Vlastní předpoklady.

Počty živě narozených odpovídající stacionární populaci o předpokládané velikosti 10,5 miliónu při tomto vývoji úmrtnosti – viz (4) – uvádí Tab. 2. Zatímco při úmrtnosti na úrovni roku 2011 by bylo pro zachování velikosti populace ČR potřeba zhruba 135 tisíc živě narozených ročně, při úmrtnosti na předpokládané úrovni roku 2100 by byl dostačující počet živě narozených přibližně o 20 tisíc menší.

Tab. 2: Počty živě narozených pro stacionární populaci ČR velikosti 10,5 miliónu osob

Rok	2011	2020	2030	2040	2050	2060	2070	2080	2090	2100
Živě narození	135 263	131 824	128 480	125 599	123 122	121 007	119 223	117 743	116 545	115 607

Zdroj: Vlastní výpočet

Vypočteme nejprve projekci populace ČR bez migrace za hypotetického předpokladu, že by současná demografická struktura obyvatelstva ČR (k 1. 1. 2012) byla pravidelná a odpovídala stacionární populaci pro rok 2011. Při výše uvedeném scénáři vývoje úmrtnosti a počtu živě narozených jsou základní výsledky této projekce uvedeny v Tab. 3.

Tab. 3: Hlavní výsledky modelové projekce stacionární populace

Charakteristika	2011	2020	2030	2040	2050	2060	2070	2080	2090	2100
Počet obyvatel (k 31. 12.)	10 500 000	10 581 467	10 737 448	10 890 966	11 016 883	11 106 278	11 156 532	11 168 205	11 144 045	11 089 148
Úhrnná plodnost	2,075	2,025	1,972	1,934	1,933	1,947	1,962	1,976	1,989	2,002
Počet živě narozených	135 263	131 824	128 480	125 599	123 122	121 007	119 223	117 743	116 545	115 607
Průměrný věk (k 31.12.)	40,1	40,5	41,4	42,5	43,5	44,6	45,5	46,2	46,8	47,2
Počet 20–64letých	5 840 727	5 853 919	5 881 069	5 892 150	5 864 868	5 803 097	5 711 339	5 598 250	5 494 127	5 403 894
Podíl 20–64letých (%)	55,6	55,3	54,8	54,1	53,2	52,3	51,2	50,1	49,3	48,7
Podíl 65+/20–64 (%)	33,6	35,0	37,9	41,4	45,1	49,0	53,0	57,0	60,0	62,1
Podíl důch.věk/ produkt.věk (%)	46,5	41,7	37,9	37,2	36,7	36,6	36,0	35,6	34,5	33,1

Zdroj: Vlastní výpočet

Vzhledem k poklesu úmrtnosti počet obyvatel roste, i když počet narozených klesá. Počet 20–64letých osob však klesá. Populace stárne. Průměrný věk by se zvýšil o více než 7 let, dochází ke snížení podílu 20–64letých osob i ke zvyšování poměru 65+letých ku 20–64letým. Na druhou stranu poměr počtu osob v důchodovém věku ku počtu osob v produktivním věku (se zohledněním zvyšování důchodového věku v ČR podle současné právní úpravy) klesá.

Je tedy zřejmé, že i v případě pravidelné výchozí věkové struktury je možné při předpokládaném dalším poklesu úmrtnosti zastavit stárnutí populace pouze trvalým zvyšováním počtu narozených či trvalou imigrací. To by mělo za následek trvalé a poměrně výrazné zvyšování počtu obyvatel, jak ukázaly další dvě varianty projekce stacionární populace. Pro udržení průměrného věku populace ČR na výchozí úrovni (40,1 roku) by musela úhrnná plodnost do roku 2020 vzrůst téměř na 2,4 a na konci století být vyšší než 2,7, což je naprosto nerealistické. Počet živě narozených by se na konci století blížil 300 tisícům ročně, počet obyvatel by překročil 18 milionů. Přesto by došlo k poklesu podílu 20–64letých osob v populaci i k poklesu poměru 65+letých ku 20-64letým. Aby byla hodnota poměru 65+letých ku 20–64letým koncem století na současné úrovni, musel by být růst plodnosti ještě vyšší, v roce 2080 by musela být úhrnná plodnost vyšší než 3,5, roční počet živě narozených by již dosahoval téměř 370 000. (Podrobné výsledky nejsou vzhledem k omezenému rozsahu článku uvedeny.)

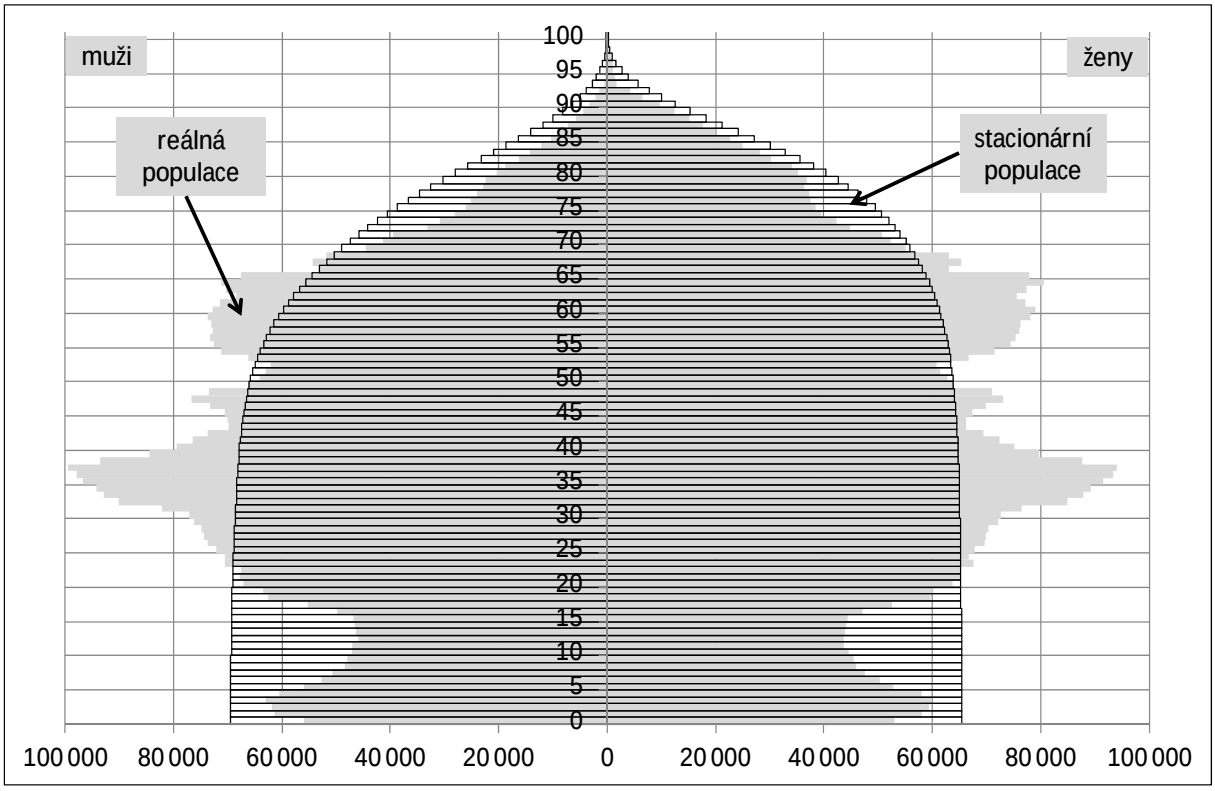
Vidíme, že ani v ideálním hypotetickém případě, kdy by současná demografická struktura ČR byla strukturou stacionární populace a úhrnná plodnost by se pohybovala na úrovni prosté reprodukce, není za předpokladu dalšího růstu délky života realistické předpokládat zastavení dalšího stárnutí populace.

3. Projekce reálné populace ČR

Současná demografická struktura se od struktury stacionární populace poměrně výrazně liší (viz Obr. 1). V reálné populaci jsou v porovnání se stacionární populaci vyšší počty osob zhruba od 20 do 65 let, počty osob do 20 let a od 65 let jsou naopak nižší. Rovněž hodnoty úhrnné plodnosti se v posledních letech pohybují hluboko pod úrovní potřebnou pro zajištění prosté reprodukce obyvatelstva.

Zabývejme se otázkou, zda a do jaké míry by mohlo v tomto století dojít k postupnému přiblížení složení populace ČR ke složení modelové populace vzniklé projekcí stacionární populace při variantě poklesu počtu narozených (popsané v předchozí části). K tomuto

přiblížení může dojít buď zvýšením plodnosti, nebo migrace, případně obou těchto charakteristik. Ve variantě bez migrace budeme předpokládat pouze zvyšování plodnosti, v dalších třech variantách budeme uvažovat zvyšování migrace při daných úrovních plodnosti. Vývoj úmrtnosti předpokládáme ve všech variantách stejný – viz Tab. 1.



Obr. 1: Struktura stacionární a reálné populace (k 1.1.2012) ČR

Zdroj: Vlastní výpočet

Uvažujme tedy 4 varianty projekce obyvatelstva ČR.

a) Varianta bez migrace (BM), předpokládající pouze výrazné zvyšování plodnosti
b) Varianta vysoké úrovně plodnosti (VP), kde se předpokládá postupný nárůst úhrnné plodnosti na 2,0 do roku 2050, v dalších letech zachování její hodnoty na této úrovni a zvyšování migrace
c) Varianta střední úrovně plodnosti (SP) uvažuje postupný nárůst úhrnné plodnosti na 1,7 do roku 2040, v dalších letech zachování její hodnoty na této úrovni a zvyšování migrace
d) Varianta nízké úrovně plodnosti (NP) předpokládá zachování úhrnné plodnosti na současné úrovni 1,4 po celé období projekce a zvyšování migrace.

Tab. 4: Varianty vývoje úhrnné plodnosti

Plodnost	2012	2020	2030	2040	2050	2060	2070	2080	2090	2100
Vysoká (VP)	1,45	1,70	1,80	1,90	2,00	2,00	2,00	2,00	2,00	2,00
Střední (SP)	1,44	1,60	1,65	1,70	1,70	1,70	1,70	1,70	1,70	1,70
Nízká (NP)	1,42	1,40	1,40	1,40	1,40	1,40	1,40	1,40	1,40	1,40

Zdroj: Vlastní předpoklady

Konvergenci složení reálné populace ČR k modelové populaci budeme posuzovat podle 4 následujících charakteristik:

a) Konvergence celkového počtu obyvatel.
b) Konvergence počtu 20–64letých osob.
c) Konvergence průměrného věku.

d) Konvergence poměru počtu 65letých a starších ku počtu 20–64letých.

První dvě charakteristiky se tedy týkají velikosti populace, další dvě jsou charakteristiky její struktury a míry stárnutí populace.

Cílem je zjistit, jaký by musel být vývoj plodnosti (ve variantě bez migrace), resp. jak velké by muselo být migrační saldo (při jednotlivých variantách předpokládaného vývoje plodnosti), aby hodnota příslušné charakteristiky konvergovala k hodnotě ve stacionárním modelu, tj. byla v roce 2060 a v roce 2100 rovna odpovídající hodnotě pro modelovou populaci. Výsledky jsou uvedeny v Tab. 5.

Tab. 5: Hlavní výsledky projekce

Varianta	2011	2020	2030	2040	2050	2060	2070	2080	2090	2100
Projekce konvergence počtu obyvatel ke stacionárnímu modelu										
Úhr. plod. (BM)	1,42	2,03	2,27	2,51	2,64	2,78	2,08	1,38	1,38	1,38
Migrace při VP	16 889	21 542	23 339	25 137	26 194	27 252	16 539	5 826	2 913	0
Migrace při SP	16 889	27 876	32 121	36 365	38 862	41 359	39 509	37 659	26 626	15 594
Migrace při NP	16 889	37 102	44 911	52 721	57 314	61 908	63 144	64 379	55 172	45 965
Projekce konvergence počtu 20–64letých osob ke stacionárnímu modelu										
Úhr. plod. (BM)	1,42	2,55	2,85	3,15	2,17	1,19	1,19	1,19	1,19	1,19
Migrace při VP	16 889	27 053	30 980	34 906	37 216	39 526	0	0	0	0
Migrace při SP	16 889	30 313	35 499	40 686	43 737	46 787	32 402	18 016	18 016	18 016
Migrace při NP	16 889	36 066	43 475	50 884	55 242	59 601	55 161	50 722	50 722	50 722
Projekce konvergence průměrného věku ke stacionárnímu modelu										
Úhr. plod. (BM)	1,42	1,94	2,15	2,35	2,47	2,59	2,03	1,48	1,48	1,48
Migrace při VP	16 889	53 061	67 036	81 011	89 232	97 453	48 726	0	0	0
Migrace při SP	16 889	89 344	117 338	145 332	161 800	178 267	169 320	160 373	151 427	142 480
Migrace při NP	16 889	160 102	215 434	270 767	303 315	335 864	372 318	408 773	485 614	562 455
Projekce konvergence průměrného věku ke stacionárnímu modelu										
Úhr. plod. (BM)	1,42	2,86	3,24	3,62	2,35	1,08	1,08	1,08	1,08	1,08
Migrace při VP	16 889	51 626	65 047	78 468	86 362	94 257	40 313	0	0	0
Migrace při SP	16 889	56 929	72 400	87 870	96 970	106 070	77 524	48 978	48 978	48 978
Migrace při NP	16 889	66 395	85 522	104 650	115 901	127 153	116 548	105 944	105 944	105 944

Zdroj: Vlastní výpočet

Vysvětlivky:

Úhr. plod. (BM) – potřebná úhrnná plodnost při projekci bez migrace

Migrace při VP, SP, resp. NP – potřebné migrační saldo při projekci s vysokou, střední, resp. nízkou úrovní plodnosti

Vzhledem k omezenému rozsahu článku nemůžeme uvést podrobnější výsledky analýzy ve formě tabulek vývoje všech základních demografických charakteristik pro každou z uvedených projekcí.

4. Závěr

Analýza potvrzuje, že dosáhnout stabilizace počtu a demografické struktury obyvatelstva ČR bez migrace není prakticky možné. Příčinou je především nerovnoměrná věková struktura v důsledku vývoje v minulém století, kdy se střídaly silné a slabší ročníky narození. Současné nízké počty osob ve věku do 20 let budou mít za následek další pokles počtu narozených a mohly by být v budoucnu bez migrace vykompenzovány pouze nárůstem úhrnné plodnosti vysoko nad hodnoty nutné pro zajištění prosté reprodukce. To jednak není reálné a navíc by to mělo za následek buď výrazný nárůst počtu obyvatel (pokud by se takto vysoká plodnost udržela i nadále) nebo pokračování nepravidelností věkové struktury (pokud by plodnost ve druhé polovině tohoto století opět poklesla pod hranici prosté reprodukce).

Migrace proto může výraznou měrou přispět ke stabilizaci demografického vývoje ČR. Počty migrantů potřebné k tomu, aby se udržoval nebo dokonce mírně rostl počet obyvatel ČR, se pohybují v řádu desítek tisíc ročně, a to i za předpokladu, že se nezvýší současná nízká úroveň plodnosti.

Na druhou stranu možnosti migrace zpomalit stárnutí populace na optimální úroveň (odpovídající vývoji při plodnosti na úrovni prosté reprodukce) jsou omezené. Při současné

nízké plodnosti by budoucí roční počty migrantů musely být vyšší než 100 tisíc. To by samozřejmě vedlo nejen k nárůstu počtu obyvatel, ale i k výraznému zvyšování podílu cizinců v naší populaci.

K optimálnímu populačnímu vývoji by proto nejspíše přispělo postupné zvyšování plodnosti na úroveň úhrnné plodnosti blízké 2,0 a současná trvalá imigrace ze zahraničí, jejíž objem by se pak mohl postupně snižovat.

5. Literatura

[1]BURCIN, B. – DRBOHLAV, D, – KUČERA, T. 2007. Koncept náhradové migrace a jeho aplikace v podmínkách České republiky. Demografie 49, No. 3., pp. 170–181.

[2]BOGUE, D. J. – ARRIAGA, E. E. – ANDERTON, D., L. (eds.). 1993. Readings in Population Research Methodology Vol. 5. Population Models, Projections and Estimates. United Nations Population Fund, Social Development Center, Chicago, Illinois.

[3]ČSÚ (CZECH STATISTICAL OFFICE). 2009. Projekce obyvatelstva České republiky do roku 2065. http://www.czso.cz/csu/2009edicniplan.nsf/p/4020-09.

[4]FIALA, T., LANGHAMROVÁ, J., PRŮŠA, L. 2011. Projection of the Human Capital of the Czech Republic and its Regions to 2050. *Demografie* 53, No. 4, pp. 304–320.

[5]FIALA, T. AND LANGHAMROVÁ, J. 2009. Human resources in the Czech Republic 50 years ago and 50 years after. In: IDIMT-2009 System and Humans – A Complex Relationship. Trauner Verlag universität, Linz.

[6]ROUBÍČEK, V. 1997. Úvod do demografie. 1. vyd. Praha. Codex Bohemia.

Adresa autorů:

Tomáš Fiala, RNDr., CSc.
Katedra demografie FIS VŠE
Nám. W. Churchilla 4, 130 67 Praha 3
Česká republika
fiala@vse.cz

Jitka „Langhamrová, doc., Ing., CSc.
Katedra demografie FIS VŠE
Nám. W. Churchilla 4, 130 67 Praha 3
Česká republika
langhamj@vse.cz

Martina Miskolczi, Mgr., Ing., MBA
Katedra demografie FIS VŠE
Nám. W. Churchilla 4, 130 67 Praha 3
Česká republika
martina.miskolczi@seznam.cz

Článek vznikl za podpory Interní grantové agentury Vysoké školy ekonomické v Praze F4/29/2011 *Analýza stárnutí obyvatelstva a dopad na trh práce a ekonomickou aktivitu*.

Porovnání dvou dvojnásobně zleva cenzorovaných výběrů typu I z Weibullova rozdělení

Comparison of two Type I Doubly Left-Censored Samples from Weibull Distribution

Michal Fusek, Jaroslav Michálek

Abstract: Left-censored data with one or more detection limits often arise in environmental contexts. The computational procedure for calculation of maximum likelihood estimators of the parameters for Type I left-censored data from underlying Weibull distribution is suggested and used considering two detection limits. The LR (likelihood ratio) test statistic for comparison of two Type I doubly left-censored Weibull populations is constructed, programmed in Matlab environment (version 7.12, R2011a), and applied in the statistical analysis of real chemical data.

Abstrakt: V environmentálních vědách se často vyskytují zleva cenzorovaná data s jedním nebo více detekčními limity. V příspěvku je navržena procedura pro výpočet maximálně věrohodných odhadů parametrů zleva cenzorovaných výběrů z Weibullova rozdělení s dvěma detekčními limity, přičemž cenzorování je typu I. Dále je odvozena testovací statistika LR, která je založena na věrohodnostním poměru (z anglického Likelihood Ratio), pro porovnání dvou dvojnásobně zleva cenzorovaných výběrů z Weibullova rozdělení a sestaven program v Matlabu (verze 7.12, R2011a) pro její výpočet. Použití testu je demonstrováno při analýze reálných chemických dat.

Key words: Fisher information matrix, likelihood ratio test, maximum likelihood estimator, double left censoring, Type I censoring, Weibull distribution.

Klíčová slova: Fisherova informační matice, test poměrem věrohodností, maximálně věrohodný odhad, dvojnásobné cenzorovaní zleva, cenzorování typu I, Weibullovo rozdělení.

JEL classification: C02, C12, C13, C34

1. Úvod

V technické praxi často narazíme na případy, kdy náhodný výběr není úplný. Například při sledování experimentálních jednotek může nastat situace, kdy rizikový jev (např. porouchání součástky) není pozorován u všech jednotek. V takovém případě mluvíme o neúplných nebo také cenzorovaných náhodných výběrech.

Obecně rozlišujeme dva druhy cenzorování, a to zleva a zprava. Cenzorování zprava se často užívá v analýze přežití, kdy nemáme možnost pozorovat experimentální jednotky po celou dobu jejich provozu až do poruchy. Cenzorování zleva se používá při analýze environmentálních nebo chemických dat, např. když se analyzovaná látka vyskytuje v tak nízké koncentraci, že nepřesáhne detekční limit měřicího zařízení. Oba druhy cenzorování se mohou lišit v závislosti na době pozorování experimentálních jednotek nebo detekčním limitu měřicího zařízení. Jestliže je detekční limit fixní, mluvíme o cenzorování typu I nebo též cenzorování časem. V takovém případě je počet cenzorovaných experimentálních jednotek náhodná veličina. Jestliže je fixní počet cenzorovaných jednotek, mluvíme o cenzorování typu II nebo též cenzorování poruchou.

Různé techniky cenzorování a metody statistické inference cenzorovaných dat jsou detailně popsány v mnoha monografiích, např. v [3], [4]. Většina autorů se zabývá cenzorováním zprava, které je v literatuře dobře rozpracováno pro všechna běžná rozdělení pravděpodobností. V případě cenzorování zleva již nejsou literární prameny tak obsáhlé.

V mnoha pracích je cenzorování zleva založeno na normálním rozdělení, což je také dobře rozpracováno v literatuře (viz např. [5], [6]). Toto rozdělení však není příliš vhodné v situacích, kdy měřená veličina (např. koncentrace chemické sloučeniny) nabývá pouze kladných hodnot a rozdělení této veličiny má kladnou šikmost. Proto se budeme zabývat zleva cenzorovanými výběry z Weibullova rozdělení s parametrem měřítka λ, parametrem tvaru τ, distribuční funkcí

$$F(x) = \begin{cases} 1 - e^{-\left(\frac{x}{\lambda}\right)^{\tau}}, & x \geq 0, \\ 0, & x < 0 \end{cases} \tag{1}$$

hustotou

$$f(x) = \begin{cases} \frac{\tau}{\lambda^{\tau}} x^{\tau-1} e^{-\left(\frac{x}{\lambda}\right)^{\tau}}, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2}$$

a dvěma detekčními limity d_1, d_2. V takovém případě máme k dispozici pouze pozorování nad detekčním limitem d_2 a počty pozorování pod zbývajícími detekčními limity. V dalším odstavci popíšeme model dvojnásobně zleva cenzorovaného výběru z Weibullova rozdělení včetně metody odhadu neznámých parametrů. Budeme-li dále mluvit o cenzorovaném výběru, vždy budeme mít na mysli zleva cenzorovaný výběr a cenzorování typu I.

2. Model pro cenzorovaný výběr

Uvažujme cenzorovaný náhodný výběr $X_1, \ldots, X_n$, přičemž $X_{(1)}, \ldots, X_{(n)}$ bude značit uspořádaný výběr $X_1, \ldots, X_n$. Abychom zjednodušili zápis některých výrazů, položíme $d_0 = 0$ a $\ln(d_0) = 0$. Veličina N_i označuje počet pozorování v intervalu $(d_{i-1}; d_i\rangle$ a N_0 je počet necenzorovaných pozorování. Tedy pro $N_0 > 0$ jsou pozorování $X_{(n-N_0+1)}, \ldots, X_{(n)}$ necenzorovaná.

K odhadu neznámých parametrů použijeme metodu maximální věrohodnosti (viz např. [2], [9]). Užitím rovnic (1) a (2), a s využitím výsledků uvedených v [4], můžeme věrohodnostní funkci cenzorovaného výběru zapsat ve tvaru

$$L(\lambda, \tau) = \frac{n!}{N_1!\, N_2!} [F(d_1)]^{N_1} [F(d_2) - F(d_1)]^{N_2} \prod_{i=n-N_0+1}^{n} f\left(X_{(i)}\right), \tag{3}$$

přičemž součin $f\left(X_{(i)}\right)$ je roven jedné pro $N_0 = 0$. Logaritmická věrohodnostní funkce je pak tvaru

$$l(\lambda, \tau) = \ln \frac{n!}{N_1!\, N_2!} + N_1 \ln F(d_1) + N_2 \ln[F(d_2) - F(d_1)] + \sum_{i=n-N_0+1}^{n} \ln f\left(X_{(i)}\right). \tag{4}$$

Věrohodnostní rovnice pro odhad parametrů λ a τ získáme derivováním logaritmické věrohodnostní funkce (4) podle jednotlivých parametrů a jsou tvaru

$$\begin{aligned} \frac{\partial l}{\partial \lambda} &= \sum_{i=1}^{2} N_i H_i^{\lambda} - N_0 \frac{\tau}{\lambda} + \frac{\tau}{\lambda^{\tau+1}} \sum_{i=n-N_0+1}^{n} X_{(i)}^{\tau} = 0, \\ \frac{\partial l}{\partial \tau} &= \sum_{i=1}^{2} N_i H_i^{\tau} + N_0 \frac{1 - \tau \ln \lambda}{\tau} + \sum_{i=n-N_0+1}^{n} \ln X_{(i)} + \frac{\ln \lambda}{\lambda^{\tau}} \sum_{i=n-N_0+1}^{n} X_{(i)}^{\tau} \\ &\quad - \frac{1}{\lambda^{\tau}} \sum_{i=n-N_0+1}^{n} X_{(i)}^{\tau} \ln X_{(i)} = 0, \end{aligned} \tag{5}$$

kde

$$H_i^{\lambda} = \frac{\tau\left[d_{i-1}^{\tau} e^{-\left(\frac{d_{i-1}}{\lambda}\right)^{\tau}} - d_i^{\tau} e^{-\left(\frac{d_i}{\lambda}\right)^{\tau}}\right]}{\lambda^{\tau+1}\left[e^{-\left(\frac{d_{i-1}}{\lambda}\right)^{\tau}} - e^{-\left(\frac{d_i}{\lambda}\right)^{\tau}}\right]},$$
$$H_i^{\tau} = \frac{d_i^{\tau} \ln\left(\frac{d_i}{\lambda}\right) e^{-\left(\frac{d_i}{\lambda}\right)^{\tau}} - d_{i-1}^{\tau} \ln\left(\frac{d_{i-1}}{\lambda}\right) e^{-\left(\frac{d_{i-1}}{\lambda}\right)^{\tau}}}{\lambda^{\tau}\left[e^{-\left(\frac{d_{i-1}}{\lambda}\right)^{\tau}} - e^{-\left(\frac{d_i}{\lambda}\right)^{\tau}}\right]}, i = 1,2. \quad (6)$$

Numerickým řešením soustavy rovnic (5) získáme maximálně věrohodné odhady $\hat{\lambda}$ a $\hat{\tau}$ parametrů λ a τ. Odhady parametrů je rovněž možné získat maximalizací věrohodnostní funkce (3) nebo logaritmické věrohodnostní funkce (4).

V dalším odstavci rozšíříme výše uvedený model a zkonstruujeme test poměrem věrohodností pro porovnání dvou cenzorovaných výběrů z Weibullova rozdělení.

3. Test rovnosti dvou cenzorovaných výběrů

Mějme nyní dva cenzorované náhodné výběry $X_{11}, \dots, X_{1n}$ a $X_{21}, \dots, X_{2n}$ z Weibullova rozdělení s parametry λ_1 a τ_1, respektive λ_2 a τ_2, s distribuční funkcí (1) a s hustotou (2). Dále $X_{(j1)}, \dots, X_{(jn)}, j = 1,2$, bude opět značit uspořádaný výběr $X_{j1}, \dots, X_{jn}$ a veličiny N_{ji} jsou četnosti odpovídající četnostem N_i, $i = 0,1,2$, z odstavce 2, přičemž j značí číslo výběru ($j = 1,2$). Logaritmická věrohodnostní funkce pro dva výběry je pak tvaru

$$l_1(\lambda_1, \tau_1, \lambda_2, \tau_2) = \sum_{j=1}^{2} \ln \frac{n_j!}{N_{j1}!\, N_{j2}!} + N_{j1} \ln F\left(d_{j1}, \lambda_j, \tau_j\right)$$
$$+N_{j2} \ln\left[F\left(d_{j2}, \lambda_j, \tau_j\right) - F\left(d_{j1}, \lambda_j, \tau_j\right)\right] + \sum_{i=n_j-N_{j0}+1}^{n_j} \ln f\left(X_{(ji)}\right). \quad (7)$$

Proveďme nyní takovou reparametrizaci, že $\lambda_2 = \lambda_1 + \alpha$ a $\tau_2 = \tau_1 + \beta$. Pak logaritmická věrohodnostní funkce nového modelu bude tvaru

$$l_2(\lambda_1, \tau_1, \alpha, \beta) = l_1(\lambda_1, \tau_1, \lambda_1 + \alpha, \tau_1 + \beta). \quad (8)$$

Maximalizací logaritmické věrohodnostní funkce (8) získáme maximálně věrohodné odhady $\hat{\lambda}_1, \hat{\tau}_1, \hat{\alpha}, \hat{\beta}$ parametrů $\lambda_1, \tau_1, \alpha, \beta$. Můžeme použít například numerickou proceduru založenou na Nelderově – Meadově simplexovém algoritmu (viz [8]), který je implementován v Matlabu (verze 7.12, R2011a)

Při testování rovnosti dvojic parametrů dvou Weibullových rozdělení lze využít asymptotický test poměrem věrohodností (viz [1]). V uvedené reparametrizaci jsou λ_1 a τ_1 rušivé parametry. Proto pro nulovou hypotézu H_0 rovnosti parametrů měřítka a tvaru dostaneme formální zápis H_0: $\boldsymbol{\theta} = \mathbf{0}^{\mathrm{T}}$, kde $\boldsymbol{\theta} = (\alpha, \beta)^{\mathrm{T}}$. Alternativní hypotéza je H_1: H_0 neplatí. Testovací statistika, která má χ^2 rozdělení s dvěma stupni volnosti, je tvaru

$$LR = 2\left[l_2\left(\hat{\lambda}_1, \hat{\tau}_1, \hat{\alpha}, \hat{\beta}\right) - l_2\left(\tilde{\lambda}_1, \tilde{\tau}_1, 0,0\right)\right], \quad (9)$$

kde vlnka nad parametrem značí maximálně věrohodný odhad rušivých parametrů za platnosti nulové hypotézy. V dalším odstavci aplikujeme výše uvedené výsledky na reálných datech uvedených v [7].

4. Aplikační příklad

Měřením koncentrace musk sloučenin (phantolid, traseolid, galaxolid, tonalid, musk ambrette, musk xylen, musk tibeten, musk keton) v tkáni ryb vylovených před a za čističkou odpadních vod byly získány dva cenzorované výběry, které lze popsat pomocí Weibullova rozdělení (viz [7]). Maximalizací logaritmické věrohodnostní funkce (8) nejprve získáme odhady parametrů Weibullova modelu, které použijeme k výpočtu testovací statistiky (9) pro jednotlivé sloučeniny. Z výsledků testů poměrem věrohodností, které jsou uvedeny v tabulce 1, plyne, že vliv čističky odpadních vod na koncentraci uvedených musk sloučenin v rybí tkáni je na hladině významnosti 0,05 statisticky neprůkazný.

Tab. 1: Test poměrem věrohodností pro porovnání dvou modelů koncentrací musk sloučenin. H_0 označuje nulovou hypotézu, přičemž 0 znamená, že hypotézu nezamítáme na hladině významnosti 0,05. Dále p označuje p-hodnotu daného testu, LR je hodnota testovací statistiky s kritickou hodnotou $\chi^2_{0,95}(2) = 5.99$.

Sloučenina	$\hat{\lambda}_1$	$\hat{\tau}_1$	$\hat{\alpha}$	$\hat{\beta}$	$\tilde{\lambda}_1$	$\tilde{\tau}_1$	**H_0**	**p**	**LR**
Phantolid	0,562	2,925	-0,215	-2,122	0,482	1,098	0	0,18	3,39
Traseolid	0,918	2,562	0,074	0,005	0,953	2,527	0	0,83	0,37
Galaxolid	23,977	1,612	5,381	0,436	26,688	1,799	0	0,36	2,06
Tonalid	5,175	1,523	1,613	-0,046	5,987	1,465	0	0,28	2,52
Musk ambrette	0,890	2,000	0,000	0,000	0,873	1,929	0	1,00	0,00
Musk xylen	0,460	0,496	0,002	0,264	0,429	0,556	0	0,37	2,00
Musk tibeten	0,003	0,203	0,090	1,765	0,004	0,240	0	0,24	2,88
Musk keton	2,454	1,340	-0,228	0,313	2,335	1,439	0	0,32	2,28

5. Závěr

V příspěvku byla navržena procedura pro výpočet maximálně věrohodných odhadů parametrů dvojnásobně zleva cenzorovaných výběrů typu I z Weibullova rozdělení. Dále byl navržen test poměrem věrohodností pro porovnání dvou cenzorovaných výběrů z Weibullova rozdělení, který byl následně počítačově implementován a aplikován na reálných chemických datech.

6. Literatura

[1] ANDĚL, J. 2005. *Základy matematické statistiky*. Praha: Matfyzpress, 2005. 358 s. ISBN 80-86732-40-1.

[2] BARNDORFF-NIELSEN, O. E. – COX, D. R. 1994. *Inference and Asymptotics*. London: Chapman and Hall/CRC, 1994. 360 s. ISBN 978-0412494406.

[3] COHEN, A. C. 1991. *Truncated and Censored Samples*. New York: Marcel Dekker, 1991. 312 s. ISBN 978-0824784478.

[4] COX, D.R. – OAKES, D. 1984. *Analysis of Survival Data*. New York: Chapman and Hall/CRC, 1984. 201 s. ISBN 978-0412244902.

[5] EL-SHAARAWI, A. H. – DOLAN, D. M. 1989. Maximum likelihood estimation of water concentrations from censored data. *Canadian Journal of Fisheries and Aquatic Sciences* 46, 1989, s. 1033 – 1039.

[6] EL-SHAARAWI, A. H. – ESTERBY, S.R. 1992. Replacement of censored observations by a constant: An evaluation. *Water Research* 26 (6), 1992, s. 835 – 844.

[7]FUSEK, M. – ZOUHAR, L. – MICHÁLEK, J. – VÁVROVÁ, M. 2012. *Comparison of Evaluations of Biotic Matrices Contamination Based on Incomplete Data Sets*. 22nd Annual Conference of The International Environmetrics Society, Hyderabad, India, s. 48.

[8]LAGARIAS, J. C. – REEDS, J. A. – WRIGHT, M. H. – WRIGHT,P. E. 1998. Convergence properties of the Nelder-Mead simplex method in low dimensions. SIAM *J. Optim.* 9 (1), 1998, s.112 – 147.

[9]LEHMANN, E. L. – CASELLA, G. 1998. *Theory of Point Estimation*. 2nd Edition. New York: Springer-Verlag, 1998. 589 s. ISBN 978-0387985022.

Adresy autorů:

Michal Fusek, Ing.
Fakulta strojního inženýrství
Vysoké učení technické v Brně
Technická 2896/2, 616 69 Brno
fusek.m@gmail.com

Jaroslav Michálek, Doc. RNDr., CSc.
Fakulta strojního inženýrství
Vysoké učení technické v Brně
Technická 2896/2, 616 69 Brno
michalek@fme.vutbr.cz

Kvartilová analýza produktivity spotreby produktívnych faktorov meranej tržbami v divíziách 58-63 SK NACE v MS Excel

Quartile Analysis of the Consumption of Productive Factors of Productivity Measured in Sales Divisions 58-63 SK NACE using MS Excel

Jozef Chajdiak

Abstract: The paper contains the analysis of productivity consumption of productive factors measured by revenues with the percentile and box plot for the data of enterprises Section J Information and communication in 2010 and a description creation box plot in Excel.

Abstrakt: Príspevok obsahuje analýzu produktivity spotreby viazaných produktívnych faktorov meranej tržbami pomocou percentilov a box plotu na údajoch podnikov sekcie J Informácia a komunikácia za rok 2010 a popis tvorby box plotu v Exceli

Key words: Productivity, consumption of productive factors, sales, percentiles, box plot, box plot creation

Kľúčové slová: produktivita, spotreba viazaných produktívnych faktorov, tržby, percentily, box plot, tvorba box plotu..

JEL classification: C20, C67, D24, I2, Z10

1. Úvod

Sekciu J – INFORMÁCIE a KOMUNIKÁCIA tvoria divízie 58 až 63 (Tab.1):

Tab. 1: Divízie Sekcie J SK NACE

		SEKCIA J - INFORMÁCIE A KOMUNIKÁCIA
Divízia		
58		Nakladateľské činnosti
59		Výroba filmov, videozáznamov a televíznych programov, príprava a zverejňovanie zvukových nahrávok
60		Činnosti pre rozhlasové a televízne vysielanie
61		Telekomunikácie
62		Počítačové programovanie, poradenstvo a súvisiace služby
63		Informačné služby

SCB, s.r.o. poskytol k dispozícii podnikové údaje za Sekciu J o Tržbách za predaný tovar a Tržbách za vlastné výrobky a služby, ktorých súčet sme označili Tržby. Ďalej boli k dispozícii údaje o osobných nákladoch (riadok 12 Výkazu ziskov a strát Úč POD 2-01) a odpisoch (riadok 18 + riadok 20) súčet ktorých sme označili SVPZ (spotreba viazaných produktívnych zdrojov (faktorov)). Z nich sme vypočítali produktivitu spotreby viazaných produktívnych zdrojov ako podiel PP = TRZBY / SVPZ

K analýze sa použili len údaje za podniky s kladnou hodnotou tržieb a kladnou hodnotou spotreby viazaných produktívnych zdrojov a za jednotlivé divízie sme vypočítali minimálnu hodnotu, 25., 50. a 75. percentil a maximálnu hodnotu.

2. Kvartily produktivity v jednotlivých divíziách

V tab.2 sú uvedené hodnoty kvartilov produktivity spotreby viazaných zdrojov meranej tržbami (v eurách tržieb na 1 euro spotreby produktívnych viazaných zdrojov).

Tab.2: Kvartily produktivity v divíziách 58 až 63 (SR, rok 2010)

	A	B	C	D	E	F	G
1	pp	58	59	60	61	62	63
2	min	0,02	0,00	0,25	0,13	0,03	0,00
3	0,25	2,26	2,41	1,13	2,78	2,24	2,32
4	0,5	4,86	4,77	2,54	4,62	4,25	4,62
5	0,75	12,17	12,95	4,58	9,99	9,51	11,48
6	max	9699,40	758,78	62,00	192,81	8398,20	17631,67

Prameň: Vlastný výpočet

Vypočítané hodnoty kvartilov môžeme prezentovať v grafickej podobe box plotom.

3. Box plot (Krabicový graf)

Box plot už patrí ku klasickej výbave štatistika – analytika.. K jeho konštrukcii sa využívajú kvartily rozdelenia analyzovaného znaku (x_{min}, $x_{0,25}$, $x_{0,50}$, $x_{0,75}$, x_{max}). Variačné rozpätie hodnôt analyzovaného znaku je podľa početností výskytu rozdelené na štvrtiny a tak dáva príslušnú predstavu o rozdelení hodnôt skúmaného znaku. Obdĺžnik v centre ležatého box plotu je v prostriedku rozdelený hodnotou mediánu (druhého kvartilu) $x_{0,50}$ na dve časti. Ľavý okraj obdĺžnika je určený hodnotou prvého kvartlilu $x_{0,25}$ a pravý okraj obdĺžnika je určený hodnotou tretieho kvartilu $x_{0,75}$. Zo stredu ľavej a stredu pravej strany obdĺžnika vychádzajú úsečky, ktoré končia hodnotou minima x_{min} resp. maxima x_{max}, t.j. prezentujú rozpätie medzi minimom a prvým kvartilom a rozpätie medzi tretím kvartilom a maximom.

Napriek všeobecnej vypovedacej schopnosti Excel priamou voľbou box plotu nedisponuje. Dá sa však poskladať z iných grafov v ponuke. Určitým ohraničením je , že k vytvoreniu box plotu by sme mali vytvoriť aspoň štyri box ploty (aspoň štyri premenné k analýze alebo jednu premennú štyri krát sa opakujúcu sa).

Prvým krokom je výpočet jednotlivých kvartilov jednotlivých premenných v políčkach B2 až G6. V Tab.2 v B2 až G2 sú minimálne hodnoty, v políčkach B3 až G3 hodnoty prvých kvartilov, v políčkach B4 až G4 mediány, v políčkach B5 až G5 hodnoty tretích kvartilov a v políčkach B6 až B6 maximálne hodnoty.

Z hodnôt kvartilov zostavíme východiskovú tabuľku hodnôt pre zostavenie box plotov v políčkach B2 až G13.

Tab.3a

	A	B	C	D	E	F	G
8		58	59	60	61	62	63
9	rad1	0,02	0,00	0,25	0,13	0,03	0,00
10	rad2	2,25	2,41	0,88	2,65	2,21	2,31
11	rad3	2,60	2,35	1,41	1,84	2,01	2,30
12	rad4	7,31	8,18	2,04	5,37	5,26	6,87
13	rad5	9687,23	745,84	57,42	182,82	8388,70	17620,18

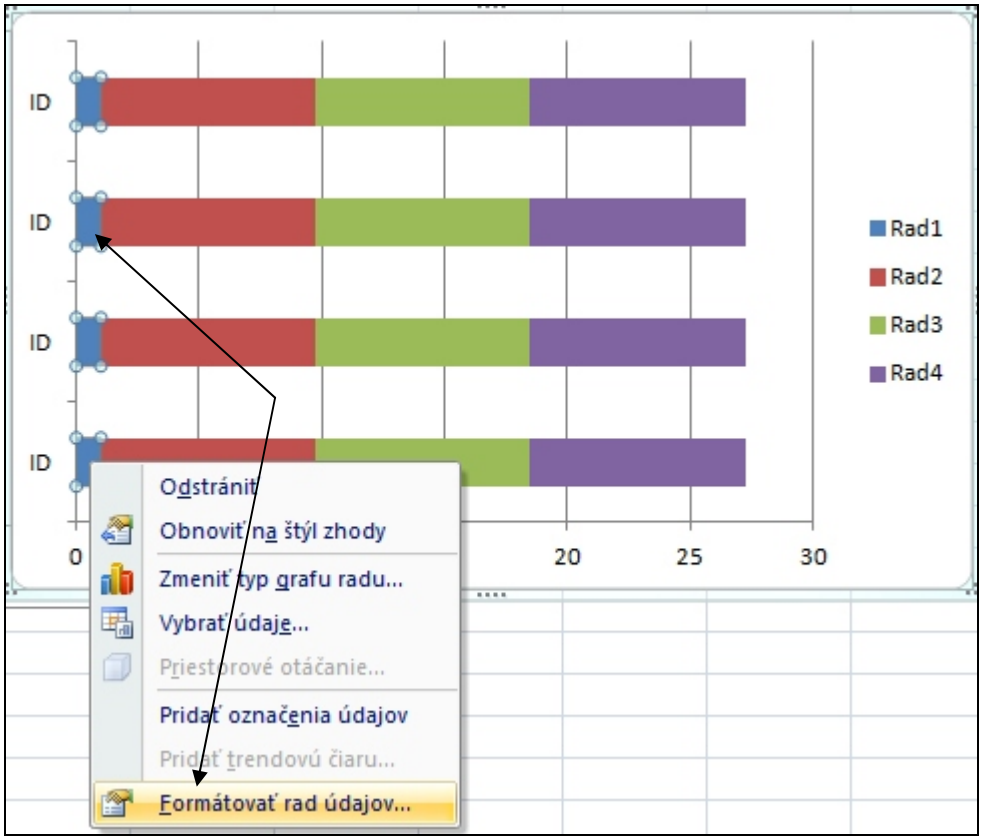
Tab.3b Zápis v tvare vzorcov

	A	B	C	D	E	F	G
8		58	59	60	61	62	63
9	rad1	=B2	=C2	=D2	=E2	=F2	=G2
10	rad2	=B3-B2	=C3-C2	=D3-D2	=E3-E2	=F3-F2	=G3-G2
11	rad3	=B4-B3	=C4-C3	=D4-D3	=E4-E3	=F4-F3	=G4-G3
12	rad4	=B5-B4	=C5-C4	=D5-D4	=E5-E4	=F5-F4	=G5-G4
13	rad5	=B6-B5	=C6-C5	=D6-D5	=E6-E5	=F6-F5	=G6-G5

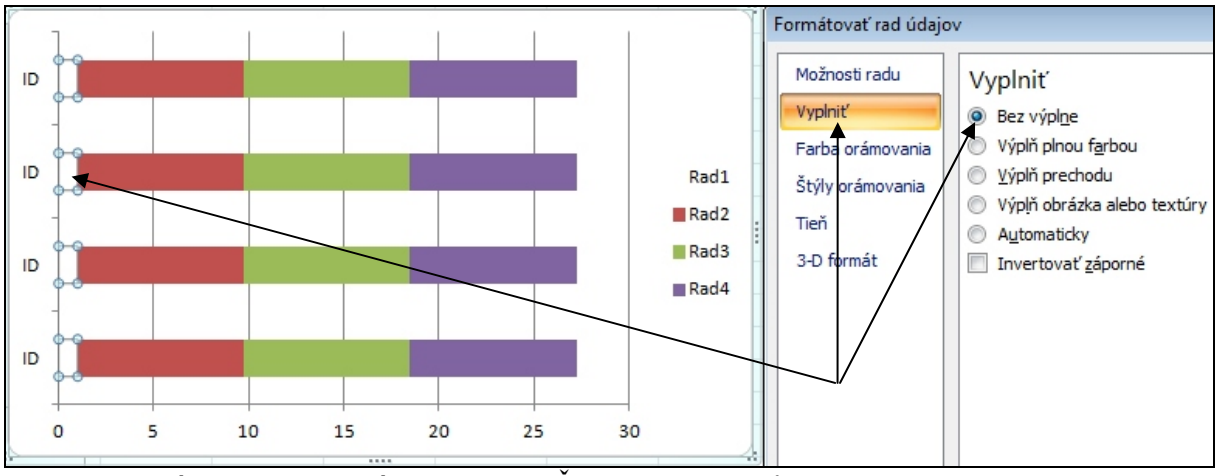
V políčkach A9 až G13 špecifikujeme Menovky Rad1 až Rad5. V 9. riadku sú hodnoty Rad1 pre jednotlivé premenné zodpovedajúce minimám týchto premenných. V 10. riadku sú rozpätia medzi prvým kvartilom a minimom u jednotlivých premenných, v 11. riadku sú rozpätia medzi mediánom a prvým kvartilom u jednotlivých premenných, v 12. riadku sú rozpätia medzi tretím kvartilom a mediánom u jednotlivých premenných a v 13. riadku sú rozpätia medzi maximom a tretím kvartilom.

Postup tvorby box plotu je nasledujúci:

Vysvietime blok A8.G12, t.j. druhú tabuľku bez posledného riadku Rad5. Ťukneme Vložiť, Pruhový, Dvojrozmerný pruhový – Skladaný pruhový.



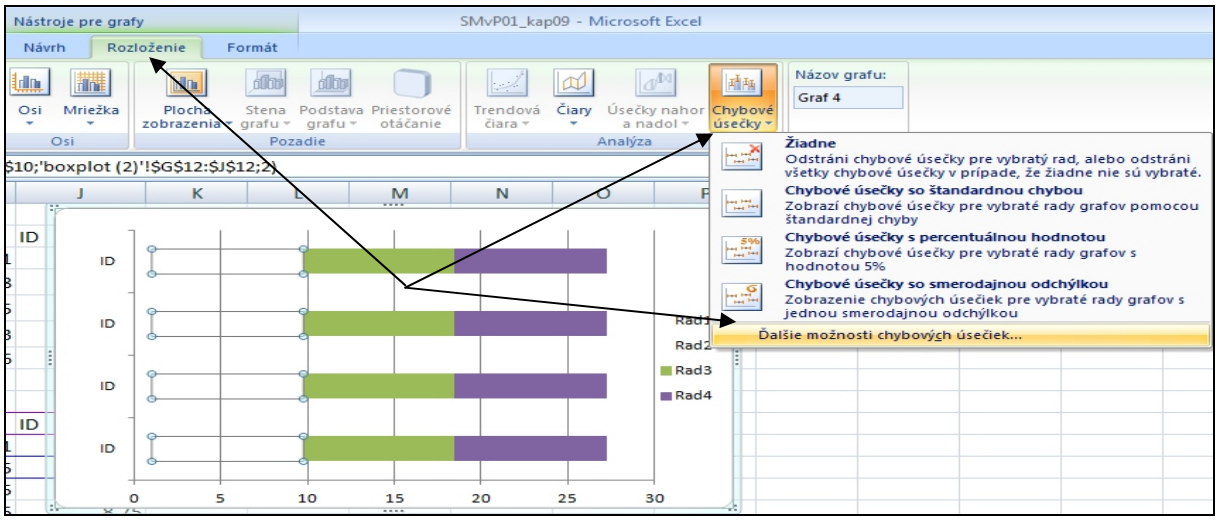
Aktivujeme Rad1. Ťukneme na prvú časť pruhového grafu (zobrazí sa v obdĺžnikoch s krúžkami v rohoch). Ťukne druhým tlačítkom myši do jedného z týchto obdĺžnikov. Objaví sa okno s ponukou, v ktorom ťukneme na Formátovať rad údajov. V časti Vyplniť zvolíme Bez výplne.



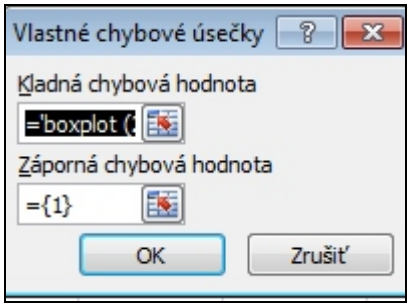
V časti Farba orámovania vyplníme Bez čiar. Ťukneme Zavrieť.

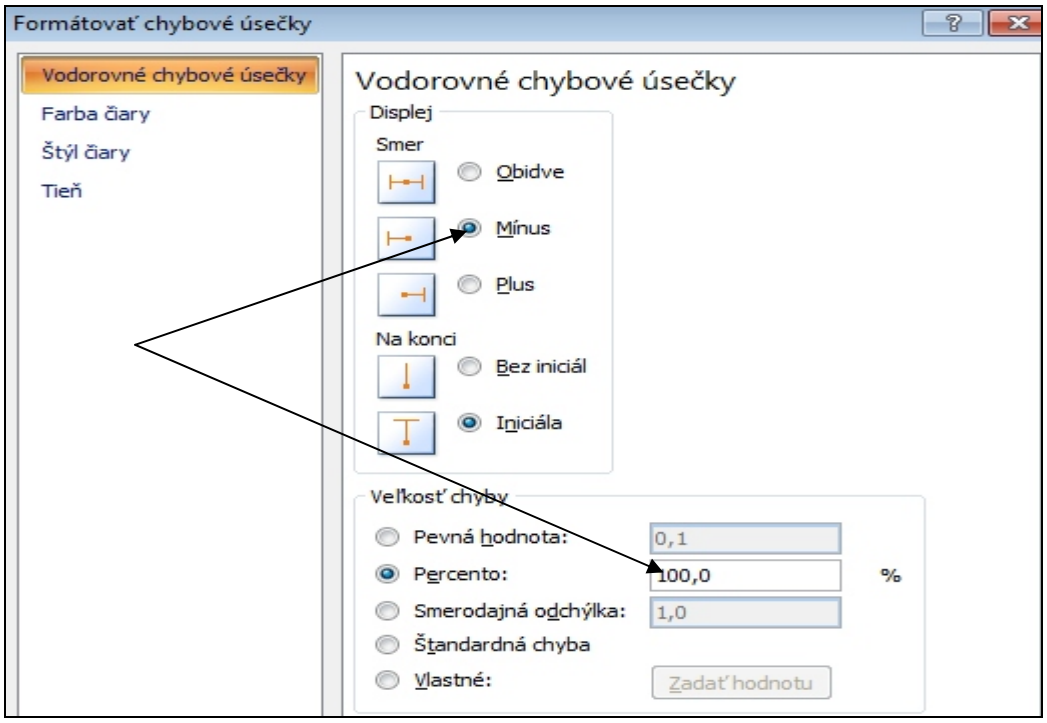
Aktivujeme Rad2. Ťukneme na druhú časť pruhového grafu (zobrazí sa v obdĺžnikoch s krúžkami v rohoch). Ťukne druhým tlačítkom myši do jedného z týchto obdĺžnikov. Objaví sa okno s ponukou, v ktorom ťukneme na Formátovať rad údajov. V časti Vyplniť zvolíme Bez výplne a v časti Farba orámovania zvolíme Bez čiar.

Rad2 je aktivovaný. Ťukneme Nástroje pre grafy / Rozloženie / Analýza/ Chybové úsečky / Ďalšie možnosti chybových úsečiek.

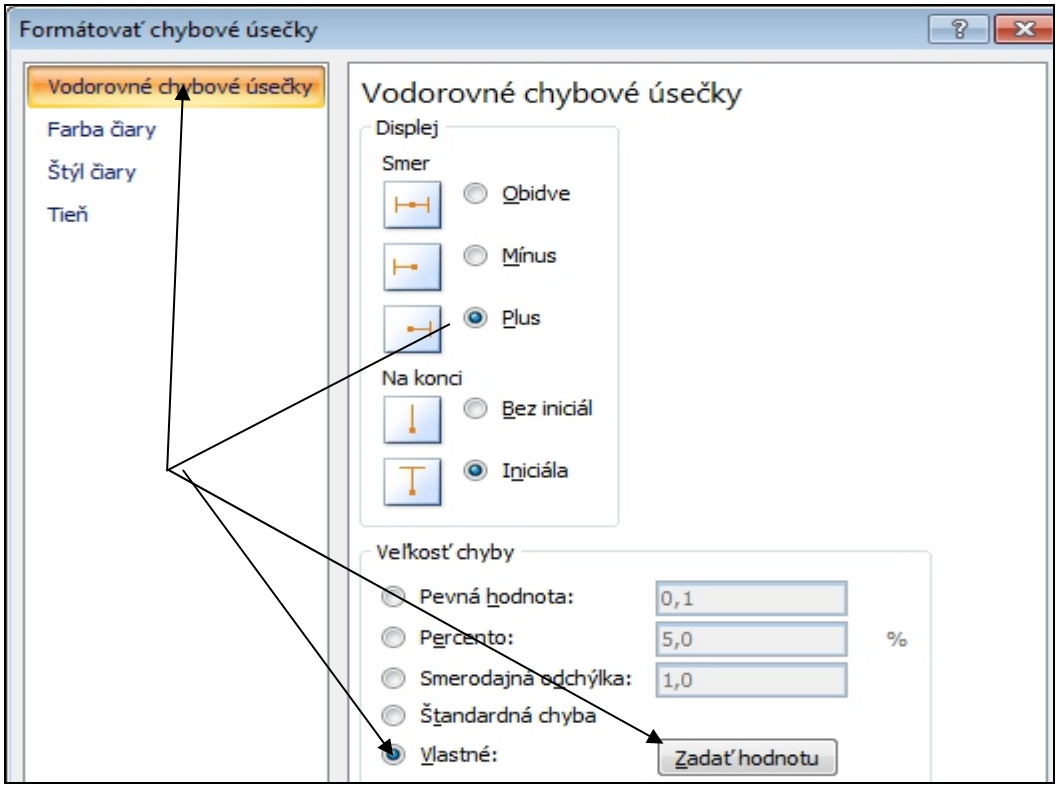


Objaví sa okno Formátovať chybové úsečky. V časti Vodorovné chybové úsečky Smer zvolíme Mínus, Na konci Iniciála a Percento zadáme 100 %. Ťukneme Zavrieť.

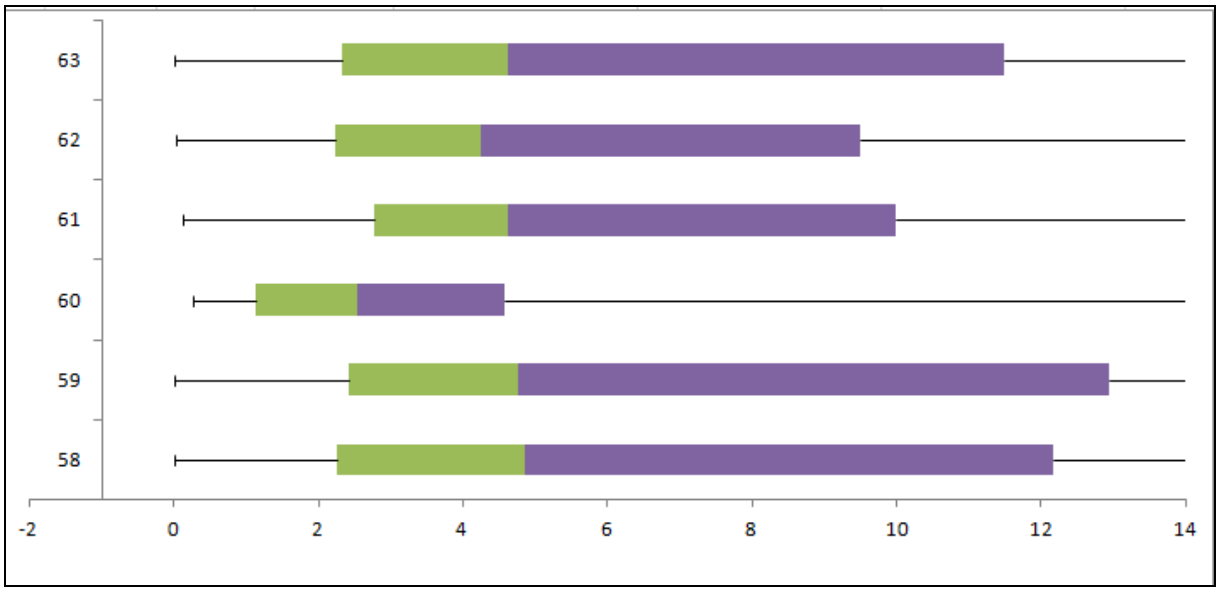




Označíme Rad4. Ťukneme postupnosť príkazov Nástroje pre grafy / Rozloženie / Analýza / Chybové úsečky / Ďalšie možnosti chybových úsečiek. Objaví sa okno Formátovať chybové úsečky. Smer zvolíme Plus a Veľkosť chyby Vlastné:. V okne Vlastné chybové úsečky, v časti Kladná chybová hodnota, zadáme hodnoty Radu5 v bloku B13.G13.



Dostávame päť box plotov pre divízie 58 až 63.



Obr. 2 Box ploty pre produktivitu

Rozdelenie hodnôt produktivity má hyperbolický charakter (maximá sú viacciferné čísla) a preto na obr.2 sú hodnoty maxím v boxplote urezané. Vidíme ale, že hodnoty na úrovní prvého kvartilu sú okolo dvoch a medián je do päť, tretí kvartil od 9 do 12.

4. Záver

Najnižšia produktivita je v divízii 60 (Činnosti pre rozhlasové a televízne vysielanie) a najvyššia v divízii 59 (Výroba filmov, videozáznamov a televíznych programov, príprava a zverejňovanie zvukových nahrávok).

5. Literatúra:

[1] CHAJDIAK, J.: *Štatistika v Exceli 2007*. Bratislava: Statis, 2009. ISBN 978-80-85659-49-8.

[2] CHAJDIAK, J.: *Štatistika jednoducho*. 3.vydanie. Bratislava: Statis, 2010. ISBN 978-80-85659-60-3.

[3] WALKENBACH, J.: *Microsoft Office Excel 2007 Grafy*. Brno: Computer Press, a.s., 2009. ISBN 978-80-251-2305-8.

Adresa autora:

Jozef Chajdiak, doc. Ing. CSc.
ÚM STU
Vazovova 5, Bratislava
chajdiak@statis.biz

Vypracované v rámci riešenia úlohy VEGA č. 1/1164/*12 „Možnosti uplatnenia informačných a komunikačných technológií na zvyšovanie efektívnosti medzinárodnej spolupráce malých a stredných podnikov SR v oblasti inovácií“*.

Úroveň pochopenia pojmu aritmetický priemer
The level of understanding the concept of arithmetic mean

Martina Ivanecká

Abstract: This paper deals with cognitive construction concept of the arithmetic mean in mind of student learning this concept in terms of APOS theory. We conducted a survey among graduates of secondary schools through a task whose solution requires a different level of understanding of the arithmetic mean.

Abstrakt: V predkladanom článku sa zaoberáme kognitívnou konštrukciou pojmu aritmetický priemer v mysli študentov učiacich sa tento pojem z pohľadu APOS teórie. Realizovali sme prieskum medzi absolventmi stredných škôl prostredníctvom úloh, ktorých riešenie si vyžaduje rôznu úroveň pochopenia aritmetického priemeru.

Key words: arithmetic mean, APOS theory, cognitive construction of concept.

Kľúčové slová: aritmetický priemer, APOS teória, kognitívna konštrukcia pojmu.

JEL classification: C 18

Úvod

Dnešný človek je každý deň v kontakte s najrôznejšími štatistickými dátami. V médiách sa často objavujú zavádzajúce interpretácie týchto dát. Mnohokrát je nevyhnutné, aby na základe nich robil vážne životné rozhodnutia. Preto by každý mal disponovať aspoň minimálnymi zručnosťami zahŕňajúcimi schopnosť pracovať so štatistickými informáciami, rozumieť im, vedieť ich interpretovať a ďalej používať. Dnešná spoločnosť si čoraz viac uvedomuje narastajúci význam štatistickej gramotnosti pre každodenný život človeka, ktorý je preplnený informáciami. Preto sa štatistika stáva súčasťou učebných osnov už základných ale aj stredných škôl. Produkovať do istej miery štatisticky gramotných ľudí by malo byť úlohou už stredných škôl. Aj náš štátny vzdelávací program [7] sa o tom zmieňuje. Hovorí sa v ňom: „Dôležitá je aj výučba elementov štatistiky, najmä schopnosť správnej interpretácie štatistických dát, porozumenie štatistickým vyjadreniam, realizácia a posudzovanie jednoduchých štatistických prieskumov“. O úrovni štatistickej gramotnosti žiakov krajín sveta priebežne informuje medzinárodné meranie OECD PISA. Slovenskí žiaci pravidelne dosahujú v oblasti náhodnosť najslabší výkon spomedzi štyroch testovaných oblastí. V tejto oblasti sa umiestňujeme pod priemerom krajín OECD [8].

Štatistika je nielen pre žiakov ale aj pre mnohých učiteľov oblasťou, v ktorej majú málo skúseností, pretože sa len nedávno stala hlavnou oblasťou niektorých učebných osnov. Tradičné spôsoby vyučovania matematiky nenachádzajú veľký úspech v tejto oblasti. Môže to byť čiastočne spôsobené tým, že matematika sa často vyučuje ako predmet zameraný na pracovné postupy. Pri vyučovaní štatistiky je však naopak potrebné klásť väčší dôraz na to, aby sa študenti naučili formulovať otázky, zhromažďovať dáta a vhodne ich využívať pri riešení reálnych problémov [2].

Väčšina výskumov v oblasti štatistického vzdelávania bola realizovaná so žiakmi základných škôl alebo študentmi vysokých škôl. Dôsledkom toho sú nedostatočné vedomosti o predstavách a porozumení základným pojmom v štatistike u študentov stredných škôl. Centrom nášho záujmu je preto úroveň pochopenia vybraných štatistických pojmov práve u absolventov stredných škôl. V príspevku sa venujeme hĺbke pochopenia pojmu aritmetický priemer. O lepšie porozumenie kognitívnej konštrukcii tohto pojmu v mysli študentov sme sa pokúsili prostredníctvom APOS teórie.

1. APOS teória

APOS teória poskytuje základnú kostru štruktúry procesu učenia sa matematických pojmov. Bola vytvorená v USA skupinou Research in Undergraduate Mathematics Education Community (RUMEC) pod vedením profesora Eda Dubinského. Vznikla na základe didaktickej teórie Jeana Piageta. Prepojením Piagetovej teórie a APOS teórie ja myšlienka konštruktivizmu. APOS teória popisuje štyri špecifické kognitívne konštrukcie, ktoré prebiehajú v mysli jedinca pri učení sa matematických pojmov [3], [5], [6].

Akcia

Akcia je transformácia objektu ako reakcia na vonkajšie podnety, ktorou sa získajú ďalšie objekty. Teda študent vie operáciu uskutočniť spamäti, alebo na základe presne daných inštrukcií (predpisu). Hovoríme, že porozumenie daného pojmu má študent na úrovni akcie, ak hĺbka jeho pochopenia je obmedzená len na realizáciu operácií vzťahujúcich sa k danému pojmu. Napríklad v prípade pojmu aritmetický priemer je jeho poňatie na úrovni akcie, ak študent dokáže vypočítať aritmetický priemer z danej množiny dát použitím vzťahu pre jeho výpočet, ale nevie čo vyjadruje a kedy je vhodné či potrebné aritmetický priemer počítať.

Proces

Žiak opakovaním akcie začína získavať nad ňou vedomú kontrolu, dochádza k jej zvnútorňovaniu a akcia sa tak stáva procesom. Študent môže realizovať tú istú akciu bez vonkajších podnetov zo strany učiteľa, dokáže popísať jednotlivé kroky transformácie bez toho, aby ich musel vykonať, môže viesť akciu dopredu aj dozadu a koordinovať ju s inými akciami.

Objekt

Žiak získa objekt, ak uvažuje o procesoch a akciách ktoré realizoval, v jeho mysli dôjde mentálnemu zdvihu a následnej konštrukcii kognitívneho objektu. Objekt je vytvorený prostredníctvom „zapuzdrenia“ procesu. Termín zapuzdrenie sa používa na opis myšlienkovej konštrukcie procesu na kognitívny objekt prostredníctvom transformácie istou akciou. To znamená, že študent si je vedomý úplnosti procesu, dokáže realizovať transformáciu objektov a konštruovať objekty transformáciou. Objekt môže byť aj „odpuzdrený“ získaním procesov, z ktorých objekt vzišiel. To znamená, že študent sa tak môže pohybovať dozadu a dopredu medzi objektom a procesom chápania matematickej myšlienky. Napríklad študent, ktorý chápe smerodajnú odchýlku súboru ako charakteristiku rozptylu, ktorá hrubo odhaduje vzdialenosti od priemeru, hovoríme, že má pojem smerodajnej odchýlky poňatý na úrovni objektu.

Schéma

Schéma je koncepcia žiaka, ktorá zahŕňa akcie, procesy a objekty jednej matematickej tematickej oblasti vrátane vzťahov medzi nimi. Žiak sa dokáže jednoznačne rozhodnúť, či nejaký prvok do schémy patrí alebo nie. Môže uskutočniť jej transformáciu a schéma sa môže stať novým objektom. Matematické objekty môžu tak vyplynúť jednak z procesov, ale aj schém. Konečným cieľom vzdelávania pre každého žiaka je dosiahnutie práve tejto úrovne v procese učenia.

Z pohľadu APOS teórie možno kognitívny vývin žiaka vzhľadom k danej koncepcii popísať nasledovne. Najprv je potrebné aby žiak vykonal akcie zavádzaného pojmu. Následne sú tieto akcie zvnútornené do procesov. Výsledné procesy sú zase zapuzdrené do objektov. Nakoniec žiak koordinuje tieto svoje mentálne konštrukcie do schémy daného pojmu [5].

2. Materiál a metódy

S cieľom zistiť, ako dokážu študenti po absolvovaní strednej školy pracovať so základnými pojmami popisnej štatistiky, aká je úroveň pochopenia týchto pojmov, ako dokážu

interpretovať údaje v kontexte danej situácie a posúdiť relevantnosť tvrdení, sme realizovali prieskum formou pracovného listu obsahujúceho úlohy vyžadujúce rôznu úroveň porozumenia pojmu aritmetický priemer. Tieto úlohy sme volili na základe jednotlivých úrovní pochopenia z hľadiska APOS teórie.

Prieskum bol realizovaný na vzorke 82 študentov prvého a druhého ročníka dennej formy bakalárskeho štúdia prírodovedeckej fakulty. Priemerný vek účastníkov bol 20 rokov. Prieskumná vzorka pozostávala z 69 žien (84%) a 13 mužov (15%). Keďže študentmi prírodovedeckej fakulty sú prevažne absolventi gymnázií, odzrkadlilo sa to aj na našej vzorke. U 79 respondentov je ukončenou školou gymnázium a len dvaja ukončili obchodnú akadémiu a jeden strednú odbornú školu. Pracovný list riešili študenti v priebehu prvej polovice letného semestra školského roka 2011/12. V tomto období títo študenti zatiaľ nemali na vysokej škole žiaden predmet, v ktorom by sa stretli so štatistikou. Teda všetky poznatky zo štatistiky nadobudnuté na strednej škole nemohli byť ovplyvnené vysokoškolskou štatistikou. Riešenia úloh sme podrobili kvantitatívnej ale aj kvalitatívnej analýze, pri ktorej nám pomohli zdôvodnenia riešení študentov, ktoré sme od nich pri jednotlivých úlohách požadovali.

3. Formulácia úloh a sledované javy

Na základe rozhovorov s učiteľmi stredných škôl, štúdia vzdelávacích štandardov, výskumov v tejto oblasti [5] a inšpirovaním sa úlohami štúdie PISA realizovanej na Slovensku v roku 2003 sme do pracovného listu zameraného na prieskum štatistických vedomostí absolventov stredných škôl vybrali úlohy vyžadujúce rôznu úroveň pochopenia pojmu aritmetický priemer.

V tejto časti uvedieme znenia jednotlivých úloh, konkretizáciu javov, ktoré sme v každej úlohe sledovali spolu s úrovňou matematických vedomostí a zručností potrebných pre riešenie jednotlivých úloh.

Úloha 1: *Včera skontrolovali revízori cestovne lístky cestujúcich dvadsiatich autobusov v Košiciach, pričom v jednotlivých autobusoch natrafili postupne na 1, 3, 2, 4, 2, 5, 3, 3, 4, 0, 3, 2, 7, 1, 4, 5, 2, 6, 3 a 4 čiernych pasažierov. Zistite, koľko čiernych pasažierov pripadá v priemere na jeden kontrolovaný autobus.*

Skúmané javy a úroveň matematických vedomostí a zručností:

- práca s aritmetickým priemerom v štandardnej situácii,
- reprodukčná úroveň.

Úloha patrí medzi klasické úlohy bežne sa vyskytujúce v každej stredoškolskej učebnici. Pri riešení prvej časti úlohy je potrebné použiť všeobecne známy základný vzťah pre výpočet aritmetického priemeru, s ktorým sa stretli už aj žiaci základných škôl a bežne sa využíva aj v iných oblastiach matematiky nielen v rámci štatistiky. Prostredníctvom tejto úlohy sme chceli zistiť, či študenti ovládajú algoritmus výpočtu priemeru. Teda, či majú procedurálne pochopenie pojmu aritmetický priemer a z hľadiska APOS teórie, či majú tento pojem poňatý na úrovni akcie.

Úloha 2: *Triedu navštevuje 17 dievčat a 12 chlapcov. Dievčatá narástli za posledný rok priemerne o 10 cm, chlapci narástli priemerne o 7 cm. Priemerne o koľko centimetrov sa zväčšila výška žiakov v triede?*

Skúmané javy a úroveň matematických vedomostí a zručností:

- práca s aritmetickým priemerom v neštandardnej situácii,
- úroveň prepojenia.

Úloha nepatrí medzi úlohy s vysokou frekvenciou výskytu v súčasných učebniciach a zbierkach matematiky pre stredné školy. Aj napriek tomu patrí medzi úlohy, ktorej riešenie nevyžaduje vedomosti nad rámec základnej školy. Jedinou podmienkou je poznanie vzťahu

pre výpočet aritmetického priemeru a schopnosť manipulácie a interpretácie tohto vzťahu. Úloha bola prevzatá zo zbierky úloh pre základné školy [1]. Touto úlohou sme sledovali, či pochopenie pojmu aritmetický priemer je u študentov na úrovni procesu.

Úloha 3: *V košickej pôrodnici sa 1. januára 2012 narodilo päť zdravých novorodencov. Novorodenec s najväčšou hmotnosťou vážil 4200 gramov. Priemerná hmotnosť týchto novorodencov bola 3 400 gramov, pričom váha ani jedného z nich nebola menšia ako 2700 gramov.*

Mohli dvaja novorodenci vážiť 4200 gramov? Svoju odpoveď zdôvodnite.

Mohli traja novorodenci vážiť 2700 gramov? Svoju odpoveď zdôvodnite.

Aká mohla byť váha prvých piatich obyvateľov Košíc v roku 2012? Uveďte aspoň dva rôzne možné prípady.

Skúmané javy a úroveň matematických vedomostí a zručností:

- práca s aritmetickým priemerom a jeho interpretácia v neštandardnej situácii,
- tvorba štatistického súboru dát pri daných určitých obmedzeniach,
- úroveň prepojenia.

Študenti sa v rámci vyučovania štatistiky stretávajú predovšetkým s úlohami, v ktorých majú vypočítať aritmetický priemer istého súboru dát. Úlohy zamerané na zostavenie súboru dát s daným aritmetickým priemerom a inými požiadavkami už nie sú až také časté. Ich riešenie si vyžaduje opačný postup ako pri zisťovaní aritmetického priemeru, na ktorý sú študenti zvyknutí. Vo väčšine prípadov ide o úlohy, ktoré majú viacero riešení. Zaujímali nás stratégie, ktoré budú študenti pri zostavovaní štatistického súboru voliť. Aké štatistické súbory zostavia, či budú obsahovať hraničné hodnoty alebo priemernú hodnotu a čím sa budú líšiť štatistické súbory zostavené jedným študentom. Z pohľadu APOS teórie sme touto úlohou sledovali chápanie pojmu aritmetický priemer na úrovni procesu.

Úloha 3: *Istý podnik sa hrdo chválil priemerným platom svojich zamestnancov, ktorý sa výrazne vymykal nízkemu regionálnemu štandardu. Samotní zamestnanci však tvrdili, že podnik falšuje výkazy, pretože ich vlastné príjmy sa vraj pohybovali pod uvádzaným priemerom. Nasledujúca tabuľka sumarizuje skutočné príjmy všetkých 200 zamestnancov danej firmy.*

Tab. 1: Príjmy zamestnancov

Platová kategória	Počet zamestnancov	Platová kategória	Počet zamestnancov
230-300	41	1330-1400	-
330-400	48	1430-1500	6
430-500	40	1530-1600	6
530-600	19	1630-1700	7
630-700	15	1730-1800	-
730-800	7	1830-1900	-
830-900	5	1930-2000	1
930-1000	-	Spolu	200
1030-1100	-		
1130-1200	5		
1230-1300	-		

Priemer z týchto údajov predstavuje sumu 575 eur. To naozaj zodpovedá oficiálne zverejnenému priemernému príjmu. Klamali teda zamestnanci? Svoju odpoveď zdôvodnite [4].

Skúmané javy a úroveň matematických vedomostí a zručností:

- vnímanie vplyvu extrémnych hodnôt na hodnotu aritmetického priemeru,
- posúdenie vhodnosti použitia aritmetického priemeru pri charakterizácii daného súboru dát,
- interpretácia aritmetického priemeru pri formulácií záverov v neštandardnej situácii,
- úroveň reflexie.

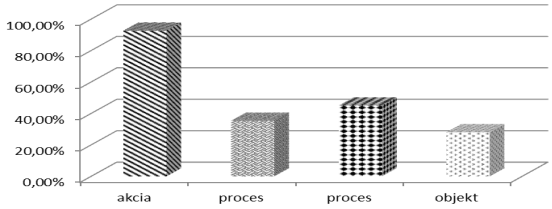
Študenti sa v médiách bežne stretávajú s podobnými vyhláseniami ohľadne priemerného platu, no v stredoškolských učebniciach matematiky sa analogická úloha hľadá len veľmi ťažko. Pre jej zaradenie sme sa rozhodli z dôvodu aktuálnosti a možnosti prostredníctvom nej hlbšie preniknúť do pochopenia pojmu aritmetický priemer u študentov. Predovšetkým nás zaujímalo ako zohľadnia veľký rozptyl súboru pri formulácií záverov. Sledovali sme ňou pochopenie pojmu aritmetický priemer na úrovni objektu.

4. Výsledky a diskusia

Cieľom popísaného experimentu bolo zistenie vedomostí absolventov stredných škôl zo štatistiky. Zamerali sme sa predovšetkým na vyšetrenie hĺbky porozumenia pojmu aritmetický priemer.

Kvantitatívne vyhodnotenie jednotlivých úloh uvádzame v nasledujúcej tabuľke a grafe.

	Úloha 1 (akcia)	Úloha 2 (proces)	Úloha 3 (proces)	Úloha 4 (objekt)
Správne	92,7%	35.4%	45.1%	28.1%
Nesprávne	6.1%	51.2%	20.7%	25.6%
Neriešené	1.2%	13.4%	34.2%	46.3%



Obr. 1: Kvantitatívne vyhodnotenie úloh sledujúcich úroveň pochopenia pojmu aritmetický priemer z pohľadu APOS teórie

Pri bližšom kvalitatívnom pohľade na jednotlivé riešenia sme pozorovali, že:

- študentom nerobí problém vypočítať aritmetický priemer jednoduchého súboru dát,
- predstava väčšiny študentov o aritmetickom priemere je založená len na algoritme „súčet všetkých hodnôt vydelený ich počtom“,
- aj napriek poznaniu vzťahu pre výpočet aritmetického priemeru majú študenti problémy s jeho interpretáciou a manipuláciou v jazyku algebry,
- zostavovanie štatistického súboru spĺňajúceho určité podmienky, čo je proces opačný k procesu zisťovania aritmetického priemeru, je pre študentov do istej miery náročný, mnohí pri tomto procese neuplatňujú poznatok o stálosti súčtu aritmetického priemeru,
- väčšina študentov nevníma vplyv extrémnych hodnôt na aritmetický priemer.

Na základe týchto pozorovaní a porovnaní úspešnosti jednotlivých úloh vyžadujúcich si rôznu úroveň pochopenia sledované pojmu sa domnievame, že študenti chápu aritmetický priemer iba ako abstraktný vzorec, chýba im však jeho konceptuálny význam. Z pohľadu jednotlivých úrovni poznania popisovanými APOS teóriou, úspešnosť jednotlivých predložených úloh nasvedčuje tomu, že tento pojem má väčšina študentov poňatý na úrovni procesu, len málo študentov preukázalo poznanie na úrovni objektu.

5. Záver

Ukázalo sa, že u takmer 72% respondentov našej prieskumnej vzorky nedošlo k mentálnej konštrukcii pojmu aritmetický priemer ani na úrovni objektu. Svedčí to o povrchnom chápaní tohto pojmu. Cieľom vyučovania štatistiky by malo byť priviesť študentov k skutočnému porozumeniu základných štatistických pojmov a myšlienok. Preto tradičná výučba spojená s memorovaním a numerickými cvičeniami by mala byť v rámci vyučovania štatistiky skôr v úzadí. Z konštruktivizmu a z APOS teórie vyplýva, že úlohou učiteľa nie je odovzdať žiakom svoje chápanie pojmu, ale namiesto toho vytvoriť situácie, v ktorých si žiaci sami budú konštruovať tieto akcie, procesy, objekty a schémy. Pedagóg praktizujúci pedagogický prístup založený na tejto teórii by mal štrukturovať činnosti, ktoré študentovi poskytnú základné skúsenosti s akciami, procesmi a objektmi štatistiky v snahe pomôcť mu budovať základné mentálne konštrukcie a organizovať ich do koherentných schém a tak ho priviesť k skutočnému poznaniu štatistických pojmov a myšlienok nezaťažených formalizmom.

Literatúra

[1]BÁLINT, Ľ. 2010. *Kombinatorika, pravdepodobnosť, štatistika, logika, grafy – zbierka riešených úloh pre žiakov 2. stupňa základných škôl a nižších ročníkov gymnázia s osemročným štúdiom.* 1. vydanie. Bratislava: Príroda, 2010. s. 102. ISBN 978-80-07-01841-9.

[2]BEGG, A. AT AL. 2004. *The school statistics curriculum: Statistics and probability education literature review*. Auckland: Auckland Uniservices Ltd, University of Auckland, 2004.

[3]DUBINSKY, E., MCDONALD, M. 2001. APOS. A Constructivist Theory of Learning in Undergraduate Mathematics Education Research. In: Holton, D. et. (Eds.) *The teaching and Learning of Mathematics at University Level: An ICMI Study*. Kluwer Academic Publishers, 2001. s. 273-280.

[4]FERJENČÍK, J. 2006. *Základy štatistických metód v sociálnych vedách*. Košice: FVS UPJŠ, 2006. s. 51.

[5]MATHEWS, D., CLARK, J. M. *Successful Students' Conceptions of Mean, Standard Deviation, and The Central Limit Theorem*.

[6]STEHLÍKOVÁ, N. 2004. *Structural Understanding in Advanced Mathematical Thinking*. Praha: PF UK, 2004. s. 11-20. ISBN 80-7290-9.

[7]ŠTÁTNY PEDAGOGICKY ÚSTAV. 2009. *Štátny vzdelávací program: MATEMATIKA (Vzdelávacia oblasť: Matematika a práca s informáciami), Príloha ISCED 3A*. Bratislava: Štátny pedagogický ústav, 2009.

[8]UHRINOVÁ, E. 2011. Porovnanie štatistickej gramotnosti krajín sveta. In: *Forum Statisticum Slovacum*, 2/2011, s. 193-198. ISSN 1336-7420.

Adresa autora (-ov):

Martina Ivanecká, RNDr.
Ústav matematických vied, PF UPJŠ
Jesenná 5, 040 01 Košice
martina.ivanecka@student.upjs.sk

Tento článok vznikol s podporou grantu VEGA 1/1331/12.

Analýza dát z oblasti neživotného poistenia metódou blokového maxima
Non-life insurance data analysis by block maxima method

Matej Juhás, Valéria Skřivánková

Abstract: This paper deals with analysis of non-life insurance data by block maxima method. Our aim is to give estimation of parameters of generalized extreme value distribution and to find the optimal choice of block size using some graphical and analytical methods.

Abstrakt: V príspevku sa zaoberáme analýzou poistných dát metódou blokového maxima. Naším cieľom je odhadnúť neznáme parametre zovšeobecneného rozdelenia extrémnych hodnôt a nájsť optimálnu veľkosť jednotlivých blokov využitím grafických a analytických metód.

Key words: Generalized extreme value distribution, parameter estimations, LR test, K-S test.

Kľúčové slová: Rozdelenie extrémnych hodnôt, odhady parametrov, LR test, K-S test.

JEL classification: C13, C16, G22

Úvod

Cieľom štatistického modelovania extrémnych hodnôt je analýza pozorovaných extrémov, odhad ich rozdelenia a predpoveď ďalších extrémnych udalostí. Prvé aplikácie teórie extrémnych hodnôt môžeme nájsť v oblasti hydrológie a s ňou spojenej oblasti klimatológie. V článku [4] z roku 1941 Gumbel, ktorý sa považuje za priekopníka v tejto oblasti, rozpracoval problematiku doby návratu záplav. V súčasnosti sa teória extrémnych hodnôt aplikuje na modelovanie zriedkavých udalostí ako sú napríklad extrémne záplavy či snehové búrky, extrémne nárazy vetra, extrémne teploty, vysoká fluktuácia výmenných kurzov vo finančnej sfére. V oblasti poisťovníctva sa teória extrémnych hodnôt využíva pri modelovaní výskytu extrémnych poistných plnení, stanovení hranice zaistenia a rizikového kapitálu. Pre poisťovňu je nevyhnutné vedieť predpovedať výskyt extrémnych poistných plnení a na základe toho stanoviť výšku poistného pre jednotlivé zmluvy.

Existujú viaceré prístupy pri modelovaní extrémnych hodnôt. Historicky prvý prístup je založený na metóde blokového maxima a vychádza z prvej vety teórie extrémnych hodnôt, ktorá popisuje limitné rozdelenia extrémov (Fisher a Tippett, 1928). Pri tejto metóde sa za extrémne hodnoty považujú maximá v blokoch rovnakej dĺžky. Neskôr sa do popredia dostáva prístup založený na odhade chvosta rozdelenia, základom ktorého je druhá veta teórie extrémnych hodnôt (Balkema a de Haan, 1974; Pickands, 1975). Táto metóda sa nazýva metóda excedentov ponad prah (POT-peaks over treshold) a študuje správanie sa veľkých pozorovaní, ktoré prevyšujú vopred zvolenú hranicu (prah). Tento príspevok bude venovaný metóde blokového maxima. Metóde POT sme sa venovali v práci [5].

1. Blokové maximum

Jedným z možných prístupov pri práci s extrémnymi hodnotami je zoskupenie dát do blokov rovnakej dĺžky. Pri tomto type registrácie sa za extrém považuje maximálna hodnota z každého bloku, pričom jednotlivé bloky môžu reprezentovať denné, mesačné alebo ročné pozorovania. Tak hovoríme o dennom, mesačnom či ročnom maxime, stručne o blokovom maxime. Takáto metóda registrácie extrémnych hodnôt sa nazýva metóda blokového maxima.

Nech $X_1, X_2, \ldots$ je postupnosť nezávislých rovnako rozdelených (iid) náhodných veličín s distribučnou funkciou F. Označme maximum z nich ako $M_n = max\{X_1, X_2,\ldots, X_n\}$, pre $n \geq 1$. Pre distribučnú funkciu takto definovaného blokového maxima platí :

$P(M_n \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) = F^n(x)$, pre x reálne a n prirodzené. Limitné rozdelenie blokového maxima je dané nasledujúcou Fisher–Tippettovou vetou [3].

Veta 1 (Fisher-Tippett, 1928).
Nech $\{X_n, n \geq 1\}$ je postupnosť iid náhodných veličín. Ak existujú normovacie konštanty $c_n > 0$, $d_n \in R$ a nedegenerovaná distribučná funkcia H taká, že

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} H, \quad n \to \infty \tag{1}$$

potom H je jednou z nasledujúcich troch typov štandardných distribučných funkcií extrémnych hodnôt:

Gumbel:	$\Lambda(x) = exp\,(-e^{-x}),\ x \in R$	(2)
Fréchet:	$\Phi_\alpha(x) = exp\,(-x^{-\alpha}),\ x > 0, \alpha > 0$	(3)
Weibull:	$\Psi_\alpha(x) = exp\,(-(-x)^{-\alpha}),\ x \leq 0, \alpha < 0.$	(4)

Dôkaz. Náčrt dôkazu možno nájsť v [2], strana 122.

Fisher-Tippettova veta tvrdí, že limitné rozdelenie maxima je jedným z troch typov rozdelení, popísaných vo Vete 1, nezávisle na rozdelení pôvodných dát. Teda rozdelenie chvosta pozorovaných dát môžeme modelovať (odhadovať) jedným z uvedených troch typov rozdelení. Konštanty c_n a d_n pre niektoré vybrané rozdelenia je možné nájsť v [2], strana 153.

Rozdelenia dané vzťahmi (2)-(4) sú špeciálnym prípadom zovšeobecnenej distribučnej funkcie extrémnych hodnôt (GEV) definovanej vzťahom

$$H_x(x) = \begin{cases} \exp\left[-\left(1+x\left(\frac{x-m}{s}\right)\right)^{-1/x}\right], & x \neq 0,\ 1+x\left(\frac{x-m}{s}\right) > 0 \\ \exp(-e^{-\frac{x-m}{s}}), & x = 0, x \in R, \end{cases} \tag{5}$$

kde $\mu \in R$, $\sigma > 0$, ξ je parameter rozdelenia GEV a nazýva sa index extrémnych hodnôt (EVI). Distribučná funkcia H_ξ zodpovedá

- Gumbelovmu rozdeleniu pre $\xi = 0$, $x \in$ R.
- Fréchetovmu rozdeleniu pre $\xi = \alpha^{-1} > 0$, $x > \xi^{-1}$
- Weibullovmu rozdeleniu pre $\xi = \alpha^{-1} < 0$, $x < \xi^{-1}$

Na otestovanie toho, či pozorované maximá pochádzajú z Gumbelovho rozdelenia, môžeme použiť test pomeru vierohodností (Likelihood Ratio-LR) test pre Gumbelov model (pozri [6] strana 118). Testovaná hypotéza je tvaru H_0: $\xi = 0$, alternatívna hypotéza H_1: $\xi \neq 0$. Testovacia štatistika je

$$T_{LR} = 2ln\frac{\prod_{i=1}^{k} h_{\hat{\mu},\hat{\sigma},\hat{\xi}}(x_i)}{\prod_{i=1}^{k} h_{\tilde{\mu},\tilde{\sigma},0}(x_i)}, \tag{6}$$

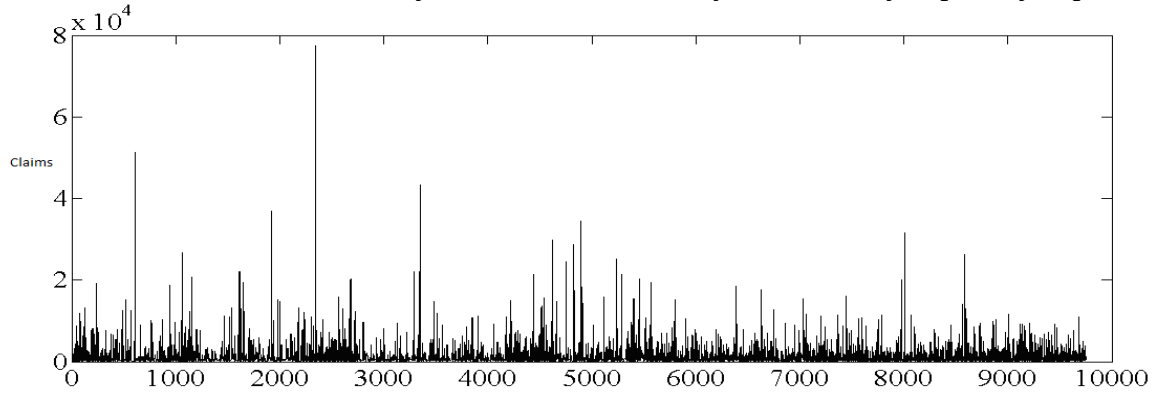
kde $\hat{\mu}, \hat{\sigma}, \hat{\xi}$ a $\tilde{\mu}, \tilde{\sigma}$ sú maximálne vierohodné odhady pre GEV a Gumbelovo rozdelenie a h je hustota GEV. Hodnoty x_i , $i = 1, 2, ...,k$ sú pozorované maximá. Asymptotické rozdelenie LR štatistiky pre $k \to \infty$ je χ^2 s jedným stupňom voľnosti. H_0 zamietame ak T_{LR} je viac ako príslušný kvantil.

Nevýhodou metódy blokového maxima je to, že využíva iba jednu hodnotu z každého bloku, pričom v bloku môžu byť aj iné vysoké hodnoty, ktoré táto metóda zanedbáva.

2. Analýza poistných dát metódou blokového maxima

K dispozícii máme súbor dát zo slovenského poistného trhu z rokov 1998 až 2008 a tieto dáta pozostávajú z 9748 poistných plnení v eurách. Uvažujeme iba tie poistné plnenia, ktoré prevyšujú 150 eur. Analýzu dát sme uskutočnili využitím softwaru R (balík extremes), Matlab (balík evim), Excel (balík EasyFitXL a Kerneldensity) a Maple.

Najprv si znázorníme časový rad (Obr. 1), na ktorom môžeme vidieť štruktúru dát v čase. Tento obrázok nám umožňuje lokalizovať a určiť výšku extrémnych poistných plnení.



Obr. 1: Vývoj poistného plnenia v čase

Základné charakteristiky dát sú uvedené v Tabuľke 1. Vidíme, že rozpätie dát je 77234.3 eur, priemer je relatívne malý a variancia dosť veľká. Podľa koeficientu šikmosti (šikmosť = 8.5) vieme, že rozdelenie výberového súboru je pozitívne zošikmené. Keďže špicatosť dát je 134.997, čo je podstatne viac ako 3(špicatosť normovaného normálneho rozdelenia), tak môžeme usúdiť, že dáta pochádzajú z rozdelenia s tučným chvostom, čo indikuje prítomnosť extrémnych hodnôt.

Tab.1: Základné charakteristiky dát

Počet	Min	Max	Priemer	Medián	Rozptyl	Šikmosť	Špicatosť
9748	150.2	77384.5	1348.458	571.015	6831693	8.5	134.997

Extrémy chceme modelovať metódou blokového maxima. Podľa Vety 1 jediné možné limitné rozdelenia blokových maxím sú Gumbelovo, Fréchetovo a Weibullovo. Preto sa pokúsime modelovať naše dáta pomocou zovšeobecnenej distribučnej funkcie extrémnych hodnôt, ktorá je daná vzťahom (5) a podľa odhadnutého parametra ξ určiť typ rozdelenia.

Keďže naše dáta pôvodne neboli zaradené do blokov, pri triedení sme zvolili viacero variant veľkosti blokov. Po dôkladnej analýze dát (na základe p hodnôt, P-P grafov, Q-Q grafov a K-S štatistiky) sa nám ako najvhodnejšie modely zdali tie, pre ktoré boli dáta zatriedené do blokov dĺžky 15, 30, 50 a 70. V nasledujúcom uvedieme maximálne vierohodné (maximum likelihood-ML) odhady neznámych parametrov pre nami uvažované veľkosti blokov, vykonáme Kolmogorov-Smirnovov (K-S) test, Anderson-Darlingov (A-D) test a LR test pre vybrané rozdelenia. Na určenie kvality modelu využijeme P-P a Q-Q grafy.

Základné charakteristiky maximálnych hodnôt v jednotlivých blokoch sú uvedené v Tabuľke 2.

Tab.2: Základné charakteristiky blokových maxím

Dĺžka bloku	Počet blokov	min	max	priemer	odchýlka	Medián
15	650	366.89	77384.5	5906.033	6243.095	4279.25
30	325	1329.55	77384.5	8312.351	7704.048	6256.62
50	195	2560.71	77384.5	10566.011	8735.034	7970.66
70	140	2608.11	77384.5	12109.03	9703.33	9377.53

Z Tabuľky 2 vidíme, že zvyšovaním dĺžky bloku *n* nám narastajú minimálne hodnoty, priemer, odchýlka a medián. Nárast minimálnej hodnoty je spôsobený zväčšovaním dĺžky bloku a tým aj zvyšovaním počtu dát v rámci bloku. Nárast odchýlky je spôsobený znižovaním počtu blokov *k*, teda znižovaním počtu dát z ktorých robíme odhady. Jednotlivé hodnoty odhadovaných parametrov spolu so štandardnými chybami pre jednotlivé bloky sú uvedené v nasledujúcej tabuľke.

Tab.3: ML odhady neznámych parametrov GEV rozdelenia a ich štandardné chyby

Dĺžka bloku	μ	σ	ξ	$Se(\mu)$	$Se(\sigma)$	$Se(\xi)$
15	3222.159	2352.609	0.390	106.984	95.400	0.039
30	4929.184	2973.683	0.389	190.372	169.573	0.055
50	6668.108	3457.517	0.386	286.24	255.461	0.070
70	7819.190	4110.603	0.355	399.24	351.980	0.079

Z odhadov neznámych parametrov vidíme, že hodnota odhadu parametra ξ sa pohybuje nad úrovňou 0.3 pre všetky bloky, čo je viac ako nula, teda jednotlivé GEV rozdelenia zodpovedajú Fréchetovmu rozdeleniu. To že sa skutočne nejedná o Gumbelovo rozdelenie overíme LR testom pre Gumbelovo rozdelenie. Jeho výsledky sú uvedené v Tabuľke 4. Ak porovnáme hodnoty testovacích štatistík s príslušnými kvantilmi χ^2rozdelenia, tak vidíme, že na hladine významnosti $\alpha = 0.05$ zamietame Gumbelovu hypotézu pre všetky zvolené modely. Z p-hodnôt pre jednotlivé testy vidíme, že zamietame zhodu dát s Gumbelovým rozdelením pre všetky štyri uvažované modely aj na hladine významnosti $\alpha = 0.01$.

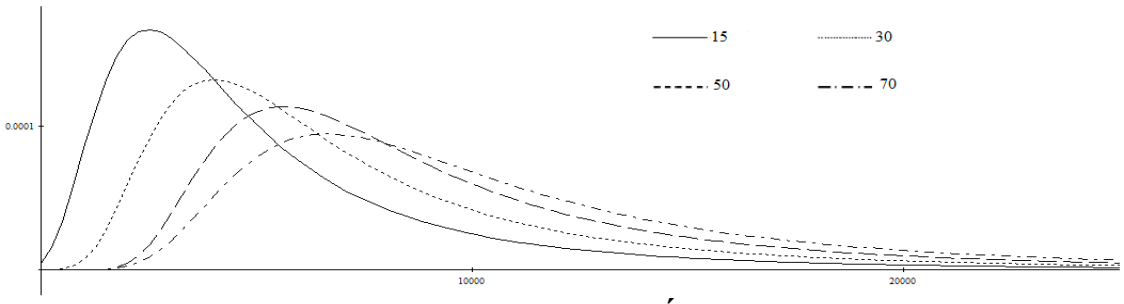
Tab.4: LR test pre jednotlivé modely

Dĺžka bloku	Štatistika	$\chi^2_{0,05}(1)$	p-hodnota	Zamietnutie
15	195.879	3.841	$1.656e^{-44}$	Áno
30	96.587	3.841	$8.539e^{-23}$	Áno
50	55.731	3.841	$8.31e^{-44}$	Áno
70	36.225	3.841	$1.75e^{-9}$	Áno

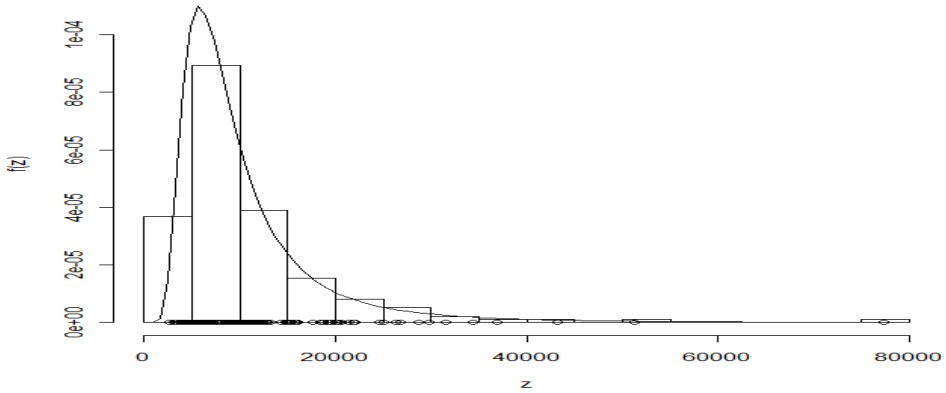
K overeniu zhody teoretického rozdelenia s empirickým sme použili K-S a A-D test dobrej zhody. Ani jeden z nich nezamieta hypotézu o zhode dát s GEV rozdelením s odhadnutými parametrami pre zvolené dĺžky blokov (Tabuľka 5). Odpovedajúce hustoty sú znázornené na Obrázku 2. Ak by sme ako kritérium pre optimálny model zvolili čo najvyššiu p-hodnotu K-S alebo A-D testu, tak ako najlepší sa javí model s dĺžkou bloku 50. Pre tento model sú p-hodnoty oboch testov najvyššie.

Tab.5: K-S a A-D testy

Dĺžka bloku	15	30	50	70
K-S test				
Statistics	0.0260	0.0297	0.0286	0.0451
p-value	0.7731	0.9374	0.9972	0.9386
A-D test				
Statistics	0.2955	0.2421	0.1160	0.1614
p-value	0.9415	0.9744	0.9999	0.997

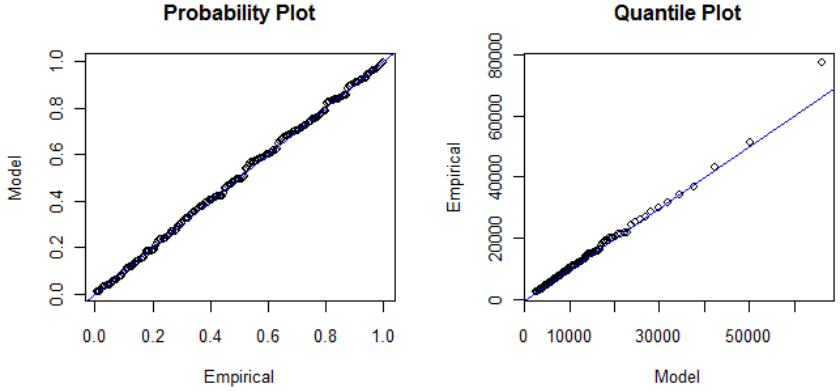


Obr 2. Hustoty pravdepodobnosti pre modely s dĺžkou bloku 15, 30, 50 a 70.



Obr.3: Histogram a hustota odhadnutého GEV rozdelenia pre n = 50

Na záver analýzy uvedieme P-P graf a Q-Q graf blokových maxím, pre dĺžku bloku 50, ktoré nám umožnia posúdiť kvalitu zvoleného modelu vizuálne (Obr. 4). Z uvedených grafov vyplýva, že vybraný model je vhodný na modelovanie blokových maxím v našich dátach.



Obr. 4: P-P a Q-Q grafy pre model s dĺžkou bloku 50

3. Záver

V príspevku sme sa venovali analýze reálnych dát z oblasti neživotného poistenia metódou blokového maxima. Dáta pôvodne neboli vhodne triedené, mali sme k dispozícii ročné údaje, teda iba 10 blokov, čo nestačilo na použitie metódy. Preto naším cieľom bolo nájsť optimálnu dĺžku bloku respektíve počet blokov. Pre zvolený model sme potom odhadli rozdelenie chvosta. Pomocou neho sme určili maximálnu výšku poistného plnenia, ktoré nebude prekročené s pravdepodobnosťou napríklad 0.995. Pre naše dáta a dĺžku bloku 50 je táto hodnota $x_{0,995}$ = 66 886.67976 eur.

Literatúra

[1]BEIRLANT, J. at al. 2004. *Statistics of Extremes: Theory and applications*. Wiley, New York. ISBN 0-471-97647-4

[2]EMRECHTS, P. at al. 1997. *Modelling extremal events for insurance and finance*. Springer-Verlag. ISSN 0172-4568

[3]FISHER, R. A., TIPPETT L. H. C 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Cambridge Philosophical Society*, vol. 24:
180-190 .

[4]GUMBEL, E. 1941. The return period of flood flows. *Ann Math Stat*, vol. 12:163–90.

[5]JUHÁS, M., SKŘIVÁNKOVÁ, V. 2010. Analýza extrémnych hodnôt v poistení motorových vozidiel metódou POT. *Forum Statisticum Slovacum* 5/2010. 97-102. ISSN 1336-7420.

[6]REISS D.-R., THOMAS M. 2007. *Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields*. Birkhäuser Verlag, Basel.
ISBN 978-3-7643-7230-9.

Adresa autorov:

Matej Juhás, RNDr.
Ústav matematických vied, PF UPJŠ
Jesenná 5, 040 01 Košice
matej.juhas@student.upjs.sk

Valéria Skřivánková, doc. RNDr. CSc
Ústav matematických vied, PF UPJŠ
Jesenná 5, 040 01 Košice
valeria.skrivankova@upjs.sk

Tento príspevok vznikol s podporou grantu VEGA No.1/0410/11 a VVGS-PF-2012-45.

FDH DEA analýza efektívnosti verejných vysokých škôl na Slovensku
FDH DEA efficiency analysis of public universities in Slovakia

Samuel Korónyi, Štefan Hronec

Abstract: Príspevok sa zaoberá aplikáciou FDH DEA modelu analýzy efektívnosti na dostupné ukazovatele slovenských verejných vysokých škôl za roky 2009-2010 v oblasti vzdelávania a vedy a techniky. Podľa FDH DEA vstupovo orientovaného modelu sú tri z nich efektívne vo všetkých analýzach: UJS Komárno, TUAD Trenčín a UK Bratislava.

Abstract: The paper deals with FDH DEA efficiency analyses applications to available indicators of Slovak public universities from the years 2009 - 2010 in the fields of education and of science and technology. According to FDH input DEA model three of them are efficient in all analyses: UJS in Komarno, TUAD in Trencin and Comenius University in Bratislava.

Keywords: Public Universities, FDH DEA Analysis

Kľúčové slová: verejné vysoké školy, FDH DEA analýza

JEL classification: I23, C61

Úvod

Príspevok prezentuje predbežné výsledky výskumu v súlade s tretím cieľom podporeného projektu VEGA č. 1/0969/11 „Matematicko-ekonomické metódy hodnotenia efektívnosti verejných vysokých škôl na Slovensku.“ Ako hlavná vedecká metóda bol použitý vstupovo orientovaný FDH DEA model. Cieľom práce bolo vyjadrenie optimálnych hodnôt ukazovateľov vzdelávania a vedy a techniky na neefektívnych verejných vysokých školách.

V prvej časti príspevku je uvedený zdroj dát, ich vymedzenie a dostupné analyzované ukazovatele.

1. Vymedzenie materiálu skúmania

Podkladom pre FDH DEA analýzy efektívnosti verejných vysokých škôl boli údaje “Výročnej správy o stave vysokého školstva“ za roky 2009 a 2010 a Centrálneho registra evidencie publikačnej činnosti.

Objektom skúmania sú verejné vysoké školy (ďalej „školy“) v pôsobnosti rezortu Ministerstva školstva SR. Analyzované boli dostupné ukazovatele charakterizujúce pedagogický proces a úroveň vedy a techniky na školách za roky 2009 a 2010. Univerzita J. Selyeho v Komárne bola založená v roku 2004. Napriek pomerne krátkemu obdobiu činnosti bola zaradená do súboru skúmaných vysokých škôl pre získanie celkového prehľadu.

V personálnej oblasti boli ako vstupy použité údaje o absolútnom počte všetkých pedagogických (profesorov, docentov, odborných asistentov, asistentov, lektorov) a vedecko-výskumných zamestnancov. Pre analýzu doktorandského štúdia bol vstupom počet profesorov a docentov.

Ako výstupné ukazovatele pre analýzu efektívnosti pedagogického procesu sme použili počty študentov a absolventov na prvom (bakalárskom), druhom (magisterskom) a treťom (doktorandskom) stupni štúdia. Osobitne pre dennú aj externú formu štúdia.

Z dostupných výkonových ukazovateľov v oblasti vedy a techniky boli pre potreby analýzy využité údaje o celkových objemoch finančnej podpory na projekty (VEGA, KEGA, APVV a zahraničné) spolu a tiež údaje o absolútnom celkovom počte vedeckých a odborných publikácií.

Riešenie problému efektívnosti vzdelávania je náročnejšie ako pri vede a technike (Gavurová 2011). Jedným z dôvodov je nemožnosť maximalizácie výstupov (napr. počty absolventov) pre rôznu personálnu a technickú náročnosť študijných odborov a udržanie štandardu kvality vzdelávacieho procesu. Preto sme z analýzy vylúčili umelecky orientované školy (AU B. Bystrica, VŠMU a VŠVU Bratislava) a UVL Košice, ktoré majú podstatne vyššie koeficienty personálnej náročnosti študijných odborov.

V ďalšej časti príspevku je stručne uvedený postup pri FDH DEA analýze.

2. Metódy skúmania

Pri analýze efektívnosti škôl bola aplikovaná optimalizačná FDH DEA metóda (Jablonský 2004). Prvé práce o DEA analýze bola uverejnené pred tridsiatimi rokmi s cieľom hodnotenia efektívnosti produkčných jednotiek aj s viacerými výstupmi. DEA metódy boli použité pri analýze efektívnosti jednotiek rôznych odvetví (vrátane školstva) a sektorov.

Ak chceme analyzovať efektívnosť skupiny jednotiek a máme k dispozícii jeden vstup a jeden výstup, potom stačí vyjadriť efektívnosť jednotiek pomerovým ukazovateľom typu výstup/vstup a zoradiť ich.

Často je však potrebné zohľadniť súčasne niekoľko vstupov a výstupov a tak dostávame netriviálnu matematickú úlohu - usporiadať súbor jednotiek vo viacrozmernom priestore, ktorý je navyše rozdelený osobitne na oblasť vstupov a výstupov. Základom vzniku DEA metód je zovšeobecnenie pomerového ukazovateľa na pomer váženého súčtu výstupov a váženého súčtu vstupov. Na výpočet optimálnych hodnôt vstupov alebo výstupov sa väčšinou používa lineárna kombinácia niekoľkých jednotiek (u nás škôl). To v našom prípade vyvoláva otázku o reálnosti fungovania školy, ktorá je „hybridom“ iných škôl.

Principiálne odlišné riešenie uvedenej základnej úlohy bolo uverejnené v roku 1984 (Deprins et al. 1984). Pri ňom sa upúšťa od klasickej podmienky pri modelovaní produkčného procesu – konvexnosti množiny produkčných možností. Ďalším dôležitým rozdielom voči ostatným DEA modelom je to, že hodnotené jednotky sa porovnávajú iba so skutočnými efektívnymi jednotkami. Optimálne (dominantné) jednotky sú na tzv. FDH (angl. Free Disposal Hull) obale produkčnej množiny, ktorý je definovaný jednoducho ako

$$FDH(x_0, y_0) = \{(x, y) : x \geq x_0; y \leq y_0\}. \quad (1)$$

Jednotky produkčnej množiny sú porovnávané s optimálnou jednotkou na FDH tak, že žiadna nemá menší vstup a väčší výstup. Pre každé $(x, y) \in FDH(x_0, y_0)$ potom platí aspoň jedna z troch podmienok $x > x_0, y = y_0; x = x_0, y < y_0; x > x_0, y < y_0$. Optimálna jednotka teda produkuje rovnaký výstup pri menšom vstupe alebo väčší výstup pri rovnakom vstupe alebo väčší výstup pri menšom vstupe. Formálny zápis vstupovo orientovaného FDH modelu je

$$\min q, X\,l \leq q\,x_0, Y\,l \geq y_0, e\,l = 1, l \in \{0,1\}, \quad (2)$$

kde θ je vstupná efektívnosť, X a Y sú vstupné a výstupné matice a λ je alternatívny znak s hodnotami nula alebo jedna. Výhoda FDH DEA metódy je v jednoduchosti výpočtu. Nie je potrebné používať optimalizačné programovacie metódy, stačí urobiť párové porovnania jednotiek pri zohľadnení daných podmienok. Nevýhodou je možná existencia formálne efektívnych jednotiek spôsobená prítomnosťou tzv. sklzov, ale tým „trpia“ aj klasické radiálne CCR a BCC DEA modely.

Na získanie výsledkov FDH DEA modelu sme použili software DEA Solver. V tretej časti uvádzame výsledky aplikovania DEA softwaru na uvedené ukazovatele.

3. Výsledky FDH DEA analýzy

Ako prvú sme z hľadiska FDH analýzy podrobili úroveň efektívnosti na prvom (bakalárskom) a druhom (magisterskom) stupni štúdia a obidvoch jeho formách (dennej a externej). Za vstup sme zvolili celkový počet pedagogických a vedecko-výskumných zamestnancov (ďalej „zamestnancov“), ktorí sa podieľajú na vzdelávacom procese na prvom a druhom stupni štúdia. Z celkového počtu 32 (16 škôl za 2 obdobia) škôl zaradených do FDH analýzy efektívnosti prvého a druhého stupňa štúdia je dvanásť škôl neefektívnych.

Tabuľka 1 obsahuje výsledky FDH DEA analýzy efektívnosti vzdelávacieho procesu na prvom a druhom stupni štúdia. V stĺpci I/O sú názvy použitých analyzovaných vstupných a výstupných ukazovateľov. Potom nasleduje prvý blok údajov v poradí: neefektívna škola s hodnotami v ukazovateľoch, jej odpovedajúca efektívna (vzorová) škola s hodnotami v ukazovateľoch. Ďalšie dva stĺpce obsahujú absolútny (Prír. - prírastok) a relatívny (%) rozdiel medzi hodnotami odpovedajúcej efektívnej a neefektívnej školy. Ako vzor je najčastejšia EU Bratislava za rok 2009 (trikrát), dvakrát je to v prípade PU Prešov 2009, SPU Nitra 2009 a TUAD Trenčín 2010. Jedenkrát ide o TUKE Košice 2009, TVU Trnava 2009 a UCM Trnava 2009. Z dvanástich vzorov je jedenásť škôl z roku 2009 a iba jeden z roku 2010 – TUAD Trenčín. To svedčí o celkovom medziročnom poklese efektívnosti vzdelávania na prvom a druhom stupni štúdia.

Tab. 1: Výsledky FDH DEA analýzy vzdelávania prvého a druhého stupňa štúdia

I/O	EU10	EU9	Prír.	%	TVU10	TVU9	Prír.	%
PEDVED	677	633	-44	-6.5%	346	309	-37	-10.7%
STAB12D	13 406	14 605	1 199	8.9%	5 995	6 430	435	7.3%
STAB12E	3 312	3 706	394	11.9%	3 373	3 487	114	3.4%
I/O	PU10	PU9	Prír.	%	UCM10	UCM9	Prír.	%
PEDVED	562	558	-4	-0.7%	286	257	-29	-10.1%
STAB12D	9 428	9 696	268	2.8%	5 926	6 736	810	13.7%
STAB12E	4 863	5 625	762	15.7%	1 805	2 557	752	41.7%
I/O	SPU10	PU9	Prír.	%	UPJS10	SPU9	Prír.	%
PEDVED	561	558	-3	-0.5%	698	555	-143	-20.5%
STAB12D	9 203	9 696	493	5.4%	8 849	8 977	128	1.5%
STAB12E	3 678	5 625	1 947	52.9%	1 206	3 970	2 764	229.2%
I/O	TUKE10	TUKE9	Prír.	%	UPJS9	SPU9	Prír.	%
PEDVED	978	973	-5	-0.5%	706	555	-151	-21.4%
STAB12D	16 208	16 685	477	2.9%	8 824	8 977	153	1.7%
STAB12E	4 190	4 625	435	10.4%	1 033	3 970	2 937	284.3%
I/O	TUZV10	TUAD10	Prír.	%	ZU10	EU9	Prír.	%
PEDVED	339	235	-104	-30.7%	805	633	-172	-21.4%
STAB12D	4 267	4 506	239	5.6%	11 580	14 605	3 025	26.1%
STAB12E	1 966	3 817	1 851	94.2%	2 965	3 706	741	25.0%
I/O	TUZV9	TUAD10	Prír.	%	ZU9	EU9	Prír.	%
PEDVED	342	235	-107	-31.3%	772	633	-139	-18.0%
STAB12D	4 153	4 506	353	8.5%	11 713	14 605	2 892	24.7%
STAB12E	1 838	3 817	1 979	107.7%	3 435	3 706	271	7.9%

Prvá neefektívna škola v tabuľke je EU Bratislava v roku 2010. Je vzorom sama sebe v roku 2009, t.j. medziročne mierne zhoršila svoju efektívnosť vstupov (o 6,5 %) pedagogického procesu. Je to spôsobené jednak prírastkom počtu zamestnancov zo 633 v roku 2009 na 677 v roku 2010 (absolútne o 44; relatívne o 6,5 %). A navyše klesol medziročne aj počet

študentov a absolventov: na dennom štúdiu zo 14 605 na 13 406 (absolútne -1 199; relatívne 8,9 %), na externom z 3 706 na 3 312 (absolútne -394; relatívne 11,9 %).

Podobné prípady medziročného zhoršenia efektívnosti vzdelávania v zmysle zmeny z efektívnej školy na neefektívnu sú na PU Prešov, TUKE Košice, TVU a UCM Trnava.

Je to o. i. spôsobené klesajúcim počtom potenciálnych uchádzačov o štúdium, ktorý je dôsledkom nepriaznivého demografického vývoja na Slovensku. To sa udialo na uvedených školách pri súčasnom náraste počtu zamestnancov. Najmarkantnejší prípad je na UCM Trnava, kde medziročne narástol počet zamestnancov o 10,1 %, pri súčasnom poklese študentov a absolventov dennej formy o 13,7 %. Pri externej forme dokonca o 41,7 %.

Z tabuľky sú zrejmé aj ďalšie zaujímavé skutočnosti:

Neefektívna škola SPU Nitra za rok 2010 má približne rovnaký počet zamestnancov a tiež študentov a absolventov na dennom štúdiu, ako mala PU Prešov v roku 2009. Ale prešovská škola mala takmer o 2 000 študentov viac na externom štúdiu.

Trenčianska TUAD v roku 2010 mala v porovnaní s TUZV Zvolen v rovnakom roku o 30,7 % menej zamestnancov, ale o 94,2 % viac externých študentov a absolventov. Podobne je to v roku 2009, kedy mala externých študentov viac o 107,7 %.

Ďalší rozdiel je medzi UPJS Košice v roku 2009 a 2010 a SPU Nitra za rok 2009, kde pri počte zamestnancov menšom o 143 až 151 (20,5 % resp. 21,4 %), mala SPU počet externých študentov a absolventov väčší takmer o 3 000. Tento fakt vyplýva pravdepodobne z toho, že na UPJS sú zastúpené aj exaktné prírodovedné študijné odbory (napr. matematika, fyzika, chémia), ktoré patria medzi najnáročnejšie. O to viac to platí pri externom štúdiu.

Ďalšiu FDH analýzu sme použili na tretí (doktorandský) stupeň štúdia. Na rozdiel od predchádzajúceho prípadu sme za vstupný ukazovateľ zvolili celkový počet profesorov a docentov, ktorí sa najväčšou mierou podieľajú na vzdelávacom procese doktorandského štúdia. Z celkového počtu 32 škôl zaradených do FDH analýzy efektívnosti tretieho stupňa štúdia je desať škôl neefektívnych. Tabuľka 2 obsahuje údaje ukazovateľov neefektívnych škôl a k nim príslušných efektívnych vzorov.

Tab. 2: Výsledky FDH DEA analýzy vzdelávania tretieho stupňa štúdia

I/O	KU10	TVU9	Prír.	%	UKF9	EU9	Prír.	%
PROFDOC	143	96	-47	-32.9%	175	175	0	0.0%
STAB3D	183	189	6	3.3%	266	311	45	16.9%
STAB3E	279	320	41	14.7%	277	407	130	46.9%
I/O	**KU9**	**TVU9**	**Prír.**	**%**	**UMB10**	**SPU10**	**Prír.**	**%**
PROFDOC	130	96	-34	-26.2%	204	163	-41	-20.1%
STAB3D	136	189	53	39.0%	290	323	33	11.4%
STAB3E	255	320	65	25.5%	252	264	12	4.8%
I/O	**TUZV9**	**TVU9**	**Prír.**	**%**	**UMB9**	**SPU10**	**Prír.**	**%**
PROFDOC	104	96	-8	-7.7%	192	163	-29	-15.1%
STAB3D	188	189	1	0.5%	242	323	81	33.5%
STAB3E	155	320	165	106.5%	252	264	12	4.8%
I/O	**UCM9**	**TUAD9**	**Prír.**	**%**	**UPJS9**	**SPU10**	**Prír.**	**%**
PROFDOC	82	61	-21	-25.6%	204	163	-41	-20.1%
STAB3D	61	95	34	55.7%	308	323	15	4.9%
STAB3E	45	68	23	51.1%	256	264	8	3.1%
I/O	**UKF10**	**EU9**	**Prír.**	**%**	**ZU9**	**ZU10**	**Prír.**	**%**
PROFDOC	181	175	-6	-3.3%	226	225	-1	-0.4%
STAB3D	283	311	28	9.9%	452	478	26	5.8%
STAB3E	347	407	60	17.3%	351	356	5	1.4%

Ako vzor je najčastejšia TVU Trnava za rok 2009 a SPU Nitra v roku 2010 (trikrát), dvakrát je to v prípade EU Bratislava v roku 2009. Jedenkrát ide o TUAD Trenčín 2009 a ZU Žilina 2010. V tabuľke je aj formálne efektívna UKF Nitra za rok 2009 s koeficientom 1. Ide o prípad, keď vstupný ukazovateľ - počet profesorov a docentov bol rovnaký na dvoch rôznych školách, ale odpovedajúce výstupy - počty študentov a absolventov boli rôzne.

V pravom dolnom rohu tabuľky je jediná škola ZU Žilina 2009, ktorá je vzorom sama sebe v roku 2010. Tu je potrebné uvažovať aj kauzalitu, takže môžeme povedať, že ZU Žilina sa polepšila v efektívnosti doktorandského stupňa štúdia. Z mierne neefektívnej školy roku 2009 sa stala efektívna v roku 2010.

Na UKF Nitra bol v roku 2009 rovnaký počet profesorov a docentov ako na EU Bratislava. Počet študentov a absolventov denného doktorandského štúdia bol však o 16,9 % menší a pri externom štúdiu to bolo dokonca o 46,9 %.

Druhá škola KU Ružomberok v roku 2009 má o 26,2 % väčší počet profesorov a docentov v porovnaní s TVU Trnava, ale počet študentov a absolventov doktorandského štúdia menší o 39 % (denná forma) a o 25,5 % (externá forma). Ešte väčší kontrast je v prípade UCM Trnava voči TUAD Trenčín (rozdiel vo vstupe 25,6 %; rozdiel s opačným znamienkom vo výstupoch 55,7 % - denné štúdium, resp. 51,1 % - externé štúdium).

Vyššia náročnosť už spomínaných exaktných prírodovedných odborov sa znova prejavuje na UPJS Košice, kde v porovnaní s SPU Nitra je o 20,1 % väčší personálny vstup do doktorandského štúdia, ale výstup je mierne nižší.

Poslednú FDH analýzu sme použili na úroveň vedy a techniky na školách. Za vstup sme zvolili celkový počet všetkých pedagogických a vedecko-výskumných zamestnancov, ktorí participujú na výstupoch vedy a techniky. Počet podporených projektov nie je dobrý ukazovateľ, preto sme vytvorili agregovaný celkový finančný objem (v tis. Eur) všetkých typov podporených projektov (VEGA, KEGA, APVV a zahraničné). Na zistenie úrovne vedy a techniky z pohľadu publikačných aktivít sme zvolili dva agregované ukazovatele: celkový počet vedeckých publikácií a celkový počet odborných publikácií. Urobili sme tak preto, aby sme zvýšili vierohodnosť výsledkov vzájomného porovnania škôl.

Tab. 3: Výsledky FDH DEA analýzy úrovne vedy a techniky

I/O	EU9	PU10	Prír.	%	UKF9	SPU10	Prír.	%
PEDVED	633	562	-71	-11.2%	581	561	-20	-3.4%
GRANTY	503.9	532.4	28.5	5.7%	658.4	1 238.6	580.2	88.1%
VEDPUB	2 200	2 617	417	19.0%	2 085	2 312	227	10.9%
ODBPUB	542	562	20	3.7%	354	499	145	41.0%
I/O	KU9	TUZV10	Prír.	%	UMB9	SPU10	Prír.	%
PEDVED	396	339	-57	-14.4%	661	561	-100	-15.1%
GRANTY	281.5	1 292.3	1 010.7	359.0%	644.9	1 238.6	593.7	92.1%
VEDPUB	1 166	1 418	252	21.6%	2 107	2 312	205	9.7%
ODBPUB	221	232	11	5.0%	463	499	36	7.8%
I/O	PU9	SPU9	Prír.	%	UPJS9	UPJS10	Prír.	%
PEDVED	558	555	-3	-0.5%	706	698	-8	-1.1%
GRANTY	450.7	1 435.8	985.1	218.6%	2 535.9	2 772.6	236.7	9.3%
VEDPUB	1 665	1 771	106	6.4%	1 933	2 912	979	50.7%
ODBPUB	396	490	94	23.7%	249	418	169	67.9%
I/O	STU9	STU10	Prír.	%				
PEDVED	1 432	1 420	-12	-0.8%				
GRANTY	7 534.7	7 606.8	72.1	1.0%				
VEDPUB	4 050	5 848	1 798	44.4%				
ODBPUB	947	1 139	192	20.3%				

Z celkového počtu 34 škôl zaradených do FDH analýzy efektívnosti úrovne vedy a techniky bolo sedem neefektívnych (tabuľka 3). Ako vzor je najčastejšia SPU Nitra v roku 2010 (dvakrát). Jedenkrát ide o SPU Nitra z roku 2009. Za rok 2010 sú to: PU Prešov, STU Bratislava, TUZV Zvolen a UPJS Košice.

V rámci analýzy efektívnosti úrovne vedy a techniky sa dvakrát vyskytuje prípad, keď škola je sama sebe vzorom. Ide o STU Bratislava a UPJS Košice. Vzhľadom na to, že boli v roku 2009 obe (aj keď mierne) neefektívne a v roku 2010 už efektívne, tak ide o zlepšenie efektívnosti produkcie výstupov vedy a techniky. Pri STU Bratislava sa personálny vstup a projektový výstup prakticky nezmenil. Počet vedeckých publikácií však relatívne narástol o 44,4 %, odborných o 22,3 %. Podobne je to aj na UPJS Košice, kde sa tiež vstup pracovnej sily takmer nezmenil, finančný objem podporených projektov pritom narástol o 9,3 % a počet vedeckých a odborných publikácií dokonca o 50,7 %, resp. o 67,9 %.

Vysoký rozdiel je v oblasti projektov aj na TUZV Zvolen roku 2010 voči KU Ružomberok v roku 2009 (359 %). Tiež SPU Nitra roku 2009 voči PU Prešov je v projektoch podstatne viac produktívna (218,6 %).

Pre celkový prehľad nám ostáva zhrnutie výsledkov našich FDH DEA analýz.

4. Syntéza výsledkov FDH DEA analýzy

Urobili sme tri FDH analýzy efektívnosti: prvú z efektívnosti pedagogického procesu prvého a druhého stupňa štúdia, druhú z efektívnosti tretieho stupňa štúdia a tretiu z efektívnosti produkcie výstupov vedy a techniky. V tabuľke 4 sú zhrnuté výsledky.

Tab. 4: Prehľad výskytu neefektívnych škôl v analýzach

VS	Počet	SAB12	SAB3	VAT
UPJS9	3	1	1	1
TUZV9	2	1	1	-
ZU9	2	1	1	-
KU9	2	-	1	1
UKF9	2	-	1	1
UMB9	2	-	1	1
EU10	1	1	-	-
PU10	1	1	-	-
SPU10	1	1	-	-
TUKE10	1	1	-	-
TUZV10	1	1	-	-
TVU10	1	1	-	-
UCM10	1	1	-	-
UPJS10	1	1	-	-
ZU10	1	1	-	-
UCM9	1	-	1	-
KU10	1	-	1	-
UKF10	1	-	1	-
UMB10	1	-	1	-
EU9	1	-	-	1
PU9	1	-	-	1
STU9	1	-	-	1

Škola UPJS Košice v roku 2009 bola neefektívna vo všetkých troch analýzach. V skupine dvakrát neefektívnych škôl boli v pedagogickom procese za rok 2009 TUZV Zvolen a ZU Žilina. V doktorandskom štúdiu a úrovni vedy a techniky sú neefektívne KU Ružomberok, UKF Nitra, UMB B. Bystrica. Všetky školy trikrát alebo dvakrát neefektívne sú z roku 2009.

Neefektívne školy za rok 2010 sú len v skupine jedenkrát neefektívnych škôl. To svedčí o zlepšujúcom sa stave verejných vysokých škôl na Slovensku z pohľadu efektívnosti - miery využívania vstupov na produkciu výstupov.

Na prvom a druhom stupni štúdia sú neefektívne: EU Bratislava, PU Prešov, SPU Nitra, TUKE Košice, TUZV Zvolen, TVU a UCM Trnava, UPJS Košice, ZU Žilina. Všetky v roku 2010. V doktorandskom štúdiu sú za rok 2010 neefektívne KU Ružomberok, UKF Nitra a UMB B. Bystrica. V oblasti vedy a techniky boli v roku 2009 neefektívne EU Bratislava, PU Prešov, STU Bratislava. Školy, ktoré nie sú v tabuľke, boli efektívne vo všetkých troch skúmaných oblastiach. Ide o nasledovné školy: SPU Nitra v roku 2009, STU Bratislava roku 2010, TUAD Trenčín, UJS Komárno a UK Bratislava za obidva roky, TUKE Košice a TVU Trnava v roku 2009. Ak zohľadníme aj počet období efektívnosti školy, tak najviac efektívne sú TUAD Trenčín, UJS Komárno a UK Bratislava. Boli efektívne dva po sebe idúce roky 2009 a 2010.

5. Záver

Na základe voľne dostupných údajov Ministerstva školstva SR sme uskutočnili FDH DEA analýzu vybraných ukazovateľov verejných vysokých škôl na Slovensku za rok 2009 a 2010 s cieľom posúdiť efektívnosť ich pedagogického procesu a úrovne vedy a techniky. Z dostupných vhodných ukazovateľov sme použili ako vstup počet všetkých pedagogických a vedecko-výskumných zamestnancov; do výstupov sme zaradili celkové objemy finančnej podpory na projekty (VEGA, KEGA, APVV a zahraničné); celkový počet vedeckých a celkový počet odborných publikácií. Pri analýze vzdelávania to boli odpovedajúce počty študentov a absolventov.

Z výsledkov analýzy vyplýva, že tri školy - TUAD Trenčín, UJS Komárno a UK Bratislava sú efektívne (relatívne najlepšie) z pohľadu využitia pedagogických a vedecko-výskumných zamestnancov na realizovanie pedagogického procesu a výstupov vedy a techniky.

Literatúra

[1] DEPRINS, D.- SIMAR, L. – TULKENS, H. 1984. Measuring Labor Efficiency in Post Offices. In: Marchand, M. – Pestieau, P – Tulkens, H. (eds) *The Performance of Public Enterprises*, 1984, p.243-267

[2] GAVUROVÁ, B. 2011. Determinanty hodnotenia kvality vysokého školstva na Slovensku. In: *Improving the quality of education at universities: Special edition of the collection of scientific papers*. No. 6 (2011), p. 137-159. - ISSN 2078-1431

[3] JABLONSKÝ, J. – DLOUHÝ, M. 2004. *Modely hodnocení efektivnosti produkčních jednotek*. Praha : Professional publishing, 2004. ISBN: 8086419495

Adresy autorov:

Samuel Korónyi, RNDr. PhD.
Centrum vedy a výskumu UMB
Cesta na amfiteáter 1
974 01 Banská Bystrica
Email: samuel.korony@umb.sk

Štefan Hronec, doc. Ing. PhD
Ekonomická fakulta UMB
Tajovského 10
974 01 Banská Bystrica
Email: stefan.hronec@umb.sk

Bariéry podnikania v MSP[1]

Trade barriers faced by SMEs

Matúš Kubák, Vladimír Gazda, Jozef Nemec, Jaroslav Korečko, Miroslava Rostášová

Abstract: Paper focuses on external and internal barriers for business and points out their severity. In order to define the severity of each of the barriers for SMEs we have chosen a method of questionnaires with specimen formed by small and medium size enterprises categorized according to recommendation of European commission 2003/361/EC. The results of survey are evaluated by an Index of barriers to doing business of SMEs. Index of barriers to doing business of SMEs is a tool of evaluation of barriers to doing business and in our study it expresses the rate of limitations to doing business within different regions of Slovakia.

Abstrakt: Článok sa zameriava na externé a interne bariéry podnikania a poukazuje na ich závažnosť. Na štúdium vplyvu vybraných bariér podnikania sme využili dotazníkový prieskum, ktorým sme oslovili malé a stredné podniky podľa klasifikácie Európskej komisie 2003/361/EC. Výsledky prieskumu sú interpretované prostredníctvom indexu bariér podnikania MSP. Index bariér podnikania je nástroj hodnotenia bariér podnikania a v našej štúdii vyjadruje mieru obmedzovania podnikania v rôznych regiónoch Slovenska.

Key words: SMEs, barriers to doing business, index of barriers to doing business.

Kľúčové slová: MSP, bariéry podnikania, index bariér podnikania.

JEL classification: C43

1. Introduction

Enterprises are the fundamentals of market economy. Just like it does not exist perfect competition it also does not exist perfect business environment. During its life cycle, every company faces problems no matter the firms´ size, legal form, business model, line of business etc. In this paper we focus on the barriers to doing business for SMEs. We distinguish two main types of barriers. First ones are external barriers. Among external barriers we file factors that inhibit start of business making and are determined by conditions out of firm which cannot be modified by entrepreneur. Here we distinguish administrative, legislative and financial barriers. On the other side are internal barriers. Internal barriers are those which are proper to potential entrepreneur. Here we class among others lack of motivation, excessive risk aversion, lack of self-confidence, lack of resources, non-acquaintance of business fundamentals etc. (Chapčákova 2012).

Academic research in field of trade barriers range wide scope. Many surveys supported by national and multinational funds schemes have been done with aim to boost SME´s development as SME´s are the largest employers and are seen as drivers of competition and innovation in market economy. Qureshi and Herani (Qureshi 2011) are studying role of small and medium-size enterprises in the socio-economic stability of economy. Besnik (Besnik 2007) identifies barriers to growth of small and medium-sized enterprises in Kosovo. Study is based on SME survey conducted by Riinvest Institute which identified critical business environment barriers perceived by entrepreneurs such as legal environment, administrative burden, external financing, tax burden and unfair competition. Here tax burden, unfair

[1] Paper was supported by Agency for PhD students and young researchers of University of Prešov in Prešov.

competition and inadequate financing are the main obstruction of firms´ growth. Marques Ibanez and Molyneux (Marques Ibanez 2002) are seeing reduction of barriers and promoting cross-border trade in financial services - especially for capital markets and retail / SME financial service areas as the main pillar of creation of a single financial services market in European Union. Dimitrov (Dimitrov 2002) analyses difficulties and barriers which meets SMEs in process of business starting and further development. Study also proposes recommendations for the governments, business organizations and international organizations for overcoming the barriers of trade. Sharmistha (Sharmistha 1999) is studying relationship between export orientation and innovation performance. Study shows that successful exporters are involved in both product and process innovation. Cizkowicz (Cizkowicz 2007) argue about the efficacy of public support for SMEs and its impact on long-term development and social welfare. Šebestová (Šebestová 2007) is identifying internal factors that drive SMEs behavior. Aidis (Aidis 2002) study formal, informal and environmental barriers to trade. Results suggest that SMEs trading options are negatively influenced by corruption, lack of information and inadequate business skills.

2. Material and methods

Study was focused on trade barriers faced by SMEs. To do so, we are using questionnaire and obtained data are presented in Index of doing business barriers of SMEs. Questionnaire was distributed either personally or via online Google docs. Questionnaire was distributed in all regions of Slovak republic in balanced count for purpose to observe doing business barriers perception in developed and less developed regions. Enquiry was divided into three areas of study. First one focused on basic information about firms as the legal form, age of the firm, region of activity etc. Second area of study was focused on barriers to doing business upon its significance. Five degrees Likert scale was proposed as an answer, where 1 means weak barrier, problem and 5 means serious barrier, problem. Questions in the second area of study were thematically divided upon sub-indexes of Index of barriers to doing business of SMEs. As every problem has a solution, third part of the questionnaire surveyed draft of solution which would make doing business for SMEs less complicated.

2.1 Questionnaire

As mentioned above, questionnaire consisted of three parts. In this chapter we present questionnaire more precisely. In total 121 enterprises answered the questionnaire, thus n=121. Respective counts by regions are presented in Figure 1.

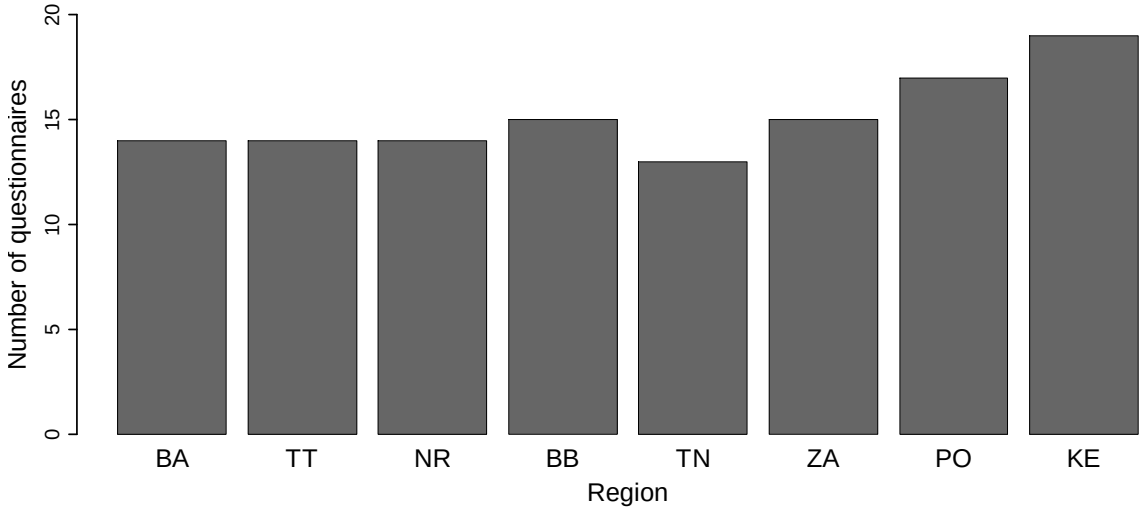


Figure 1: Distribution of questionnaires

2.1.1 Identification data

Questions aimed to classified firms upon its size (ordinal variable: less than 10 employees, between 10 and 50 employees, more than 50 employees up to 250), business sector (nominal variable: manufactory, transport, trade and services, construction), age of business (binary variable: less than one year, more than one year), geographical area of business activity (nominal variable consisting of 8 Slovak regions), experiences with doing business before started doing business on own (nominal variable: yes, no, partial).

2.1.2 Barriers to doing business

Here five degrees Likert scale was proposed and questions were divided in four groups (internal barriers, financial barriers, legislative barriers and administrative barriers). Here we denote chosen value of Likert scale as *L; L ∈ (1, 2, 3, 4, 5).*

Internal barriers: lack of motivation (a_1), business-anxiety (a_2), inadequate qualification for doing business (a_3), inadequate skills and experiences (a_4), non-acquaintance of concurrence and business area, lack of capital (a_5)

Financial barriers: access to financial resources (b_1), interest rate level (b_2), and taxation level (b_3).

Legislative barriers: endless changes in law (c_1), quality and clarity of law (c_2), law enforcement (c_3).

Administrative barriers: access to information (d_1), administrative robustness (d_2), bureaucracy (d_3), inadequate time response (d_4).

2.1.3 Improvements proposal

Five degrees Likert scale was proposed also in this part of questionnaire and questions aimed to reveal most requisite changes in business environment in Slovakia. Following improvements were proposed: creation of information center for entrepreneurs, user friendly brochure treating of starting business conditions, toll free courses improving business skills, better access to capital, diminution of banks fees, tax rebates, different tax levels in regions taking in account level of development of region, improvements in clarity of law, improvements in law enforcement, reduction of bureaucracy, possibility to fit administrative duties out of region where business is placed.

2.2 Indexes calculation

Index of barriers to doing business of SMEs was the tool used to explain intensity of barriers and expressed the differences among regions. In this analysis only data obtained in Barriers to doing business area of questionnaire entered. Index was calculated for every region and could reach score from 1 up to 5, where 1 means doing business with no obstacles and 5 means serious barriers to trade. Index of barriers to doing business of SMEs is composed of indicators which evaluated given areas of study. Every indicator could reach values 1 – 5. Indicators were classified according to sub-indexes. We distinguish four sub-indexes of barriers to doing business of SMEs:

- Sub-index of internal barriers (SIIB) $= \frac{1}{n}\sum_{i=1}^{n}\frac{\left(L_i^{a1}+L_i^{a2}+L_i^{a3}+L_i^{a4}+L_i^{a5}\right)}{5}$
- Sub- index of financial barriers (SIFB) $= \frac{1}{n}\sum_{i=1}^{n}\frac{\left(L_i^{b1}+L_i^{b2}+L_i^{b3}\right)}{3}$
- Sub-index of legislative barriers (SILB) $= \frac{1}{n}\sum_{i=1}^{n}\frac{\left(L_i^{c1}+L_i^{c2}+L_i^{c3}\right)}{3}$
- Sub-index of administrative barriers (SIAB$= \frac{1}{n}\sum_{i=1}^{n}\frac{\left(L_i^{d1}+L_i^{d2}+L_i^{d3}+L_i^{d4}\right)}{4}$

After calculation of four sub-indexes we are able to calculate Index of barriers to doing business of SMEs:

$$(\text{IBDB}) = \frac{(SIIB + SIFB + SILB + SIAB)}{4}.$$

3. Results

Values of given dub-indexes and indexes divided according to regions are shown in Table 1.

Table 1: Indexes

Region	IBDB	SIIB	SIFB	SILB	SIAB
Bratislava region	**3,45**	2,17	3,84	4,04	3,73
Trnava region	**3,63**	2,43	3,94	4,32	3,82
Nitra region	**3,66**	2,65	3,93	4,13	3,91
Banská Bystrica region	**3,67**	2,67	4,01	4,18	3,84
Trenčín region	**3,79**	2,54	4,63	4,13	3,86
Žilina region	**3,55**	2,49	3,85	4,03	3,83
Prešov region	**3,72**	2,62	4,05	4,27	3,95
Košice region	**3,72**	2,51	4,02	4,36	4,00

Eyeballing Table 1 let us conclude, that Slovak SMEs associate themselves with the uniform doing business barriers. According to our results, the worst situation is in region of Trenčín, where also the most serious financial barriers to doing business are. On the other hand in region of Bratislava the best area to doing business is. Here firms do not face serious internal barriers to doing business. Concerning internal barriers to doing business the worst situation is in region of Banská Bystrica.

Interesting are results concerning law barriers to trade. Obviously the most serious problem for Slovak SMEs is changings in legislative and law enforcement. Here we can speculate also about political influence on results and about skepticism of entrepreneurs. Surprisingly administrative barriers ranked on second place, what is interesting, because usually administrative barriers are on the first place in similar surveys.

As far as improvements proposal are concerned, respondents indicated, that creation of information center for entrepreneurs, tax rebates and reduction of bureaucracy are required.

4. Conclusion

In this paper we presented survey concerning barriers to doing business for SMEs. Using questionnaire we showed that Slovak firms consider the most serious barriers to trade to be law barriers, followed by financial barriers, administrative barriers and internal barriers. According to our query, managers would appreciate creation of information center for entrepreneurs, tax rebates and reduction of bureaucracy.

5. References

[1] AIDIS, RUTA. 2002. *Why Don't We See More Small- and Medium-sized Enterprises (SMEs) in Lithuania?: Institutional Impediments to SME Development*. Tinbergen Institute Working Paper No. 2002-038/2. Available at SSRN: http://ssrn.com/abstract=314201 or http://dx.doi.org/10.2139/ssrn.314201

[2] CHAPČÁKOVÁ, A. - HEČKOVÁ, J. - HUTTMANOVÁ, E. 2010. *Podnikanie v malých a stredných podnikoch*. Prešov: Prešovská univerzita v Prešove,. 338 s. ISBN 978-80-5550147-5.

[3] CIZKOWICZ, P. - RYBINSKI, K. 2008. *The role of banking and financial policies in promoting micro, small, and medium enterprises*. Working paper. Available at: http://ideas.repec.org/a/ris/jofitr/0021.html.

[4] KRASNIQI, B. 2007. Barriers to Entrepreneurship and SME Growth in Transition: The Case of Kosova. *Journal of Developmental Entrepreneurship*, Vol. 12, No. 1, pp. 71-94,. Available at SSRN: http://ssrn.com/abstract=1002983.

[5] MARQUÉS IBAÑEZ, D. – MOLYNEUX, P. 2002. *Integration of European Banking and Financial Markets.* No 14, EIFC - Technology and Finance Working Papers from United Nations University, Institute for New Technologies. Available at: http://EconPapers.repec.org/RePEc:dgr:unutaf:eifc02-14

[6] MITKO, D. 2002. *Opportunities and Barriers in Front of the Coss-Border Cooperation of the Enterprises in South Eastern Europe*. Economic Research Institute at Bulgarian Academy of Sciences. Available at: http://www.ceeol.com/aspx/issuedetails.aspx?issueid=5cb679f5-53c8-4bde-854f-bde373717dd2&articleid=6944d6c7-b640-4ddd-991a-7db9eda43404#a6944d6c7-b640-4ddd-991a-7db9eda43404

[7] QURESHI, J. – HERANI GOBIND, M. 2011. The role of small and medium-size enterprises (SMEs) in the socio-economic stability of Karachi. Published in: *Indus Journal of Management & Social Sciences.* No. 5(1): (30. June 2011): pp. 30-44.

[8] RIINVEST INSTITUTE FOR DEVELOPMENT RESEARCH (2002). SME survey database.

[9] SHARMISTHA BAGCHI-SEN. 1999. The Small and Medium Sized Exporters' Problems: An Empirical Analysis of Canadian Manufacturers. *Regional Studies*. Volume 33, Issue 3, pp. 231-245. DOI:10.1080

[10] ŠEBESTOVÁ, J. 2007. Aplikace VRIO metody a faktorové analýzy k nalezení bariér rozvoje malých a středních podniků v MS Kraji. Published in: *MEKON 2007 - Sborník abstraktů*. Vol. 4, pp. 83.

Authors:

Matúš Kubák, Ing. PhD.
Jozef Nemec, Ing. PhD.
Miroslava Rostášová, Mgr.
Jaroslav Korečko, Ing. PhD.

Faculty of Management
University of Prešov in Prešov
Konštantínova 16, 080 01 Prešov
matus.kubak@unipo.sk
jozef.nemec@unipo.sk
miroslava.rostasova@gmail.com
jaroslav.korecko@unipo.sk

Vladimír Gazda, doc. Ing. PhD.
Faculty of economics
Technical university in Košice
Nemcovej 32, 040 01, Košice
vladimir.gazda@mail.sk

Testing of 2-D signal separation statistical techniques for real and generated physical data sets

Testování 2-D separačních statistických metod pro reálná a generovaná fyzikální data

Václav Kůs, Jan Vejmola, Jiří Franc

Abstract: We deal with various classification methods in data sets originated from different physical experiments. It can be the case of accelerated particles data in proton-antiproton collisions (Fermilab) or applications of acoustic emission detection in nondestructive testing. The acoustic emission emerges due to the cracks, fatigues or possibly other nonlinear material effects. Signals of the acoustic emission may differ by types of materials and through these differences the signals can be assigned to material which they originate from. The classification of signals of acoustic emission can be done by means of different classification methods, in our case by means of Model-Based Clustering method (MBC). In this work we also test the suitability of chosen parameters which were used for identification of acoustic emission sources [1, 2]. The method is also successfully applied to real experimental data.

Abstrakt: Zabýváme se separačními technikami v datových souborech pocházejících z různých fyzikálních experimentů. Jde například o data z velkých urychlovačů elementárních částic pro proton-antiproton kolize (Fermilab), pro protonové srážky (Cern), případně o data pocházející z akustické emise v nedestruktivním testování materiálů. Akustická emise (AE) se objevuje v důsledku trhlin, únavy nebo jiných možných nelineárních materiálových efektů. Detekované signály se zpravidla liší podle typu materiálu, a tedy tyto rozdíly mohou být využity k identifikaci a přiřazení testovaného vzorku, ze kterého signály pocházejí. V našem případě jsme testovali MBC metodu (Model Based Clustering) založenou na statistických principech distribučních směsí a EM algoritmu. Bylo zjištěno, že významně závisí na podmnožině zvolených klasifikačních příznaků. Algoritmus je dále úspěšně aplikován i na reálná experimentální data AE.

Key words: Signal separation, Acoustic emission, Particle physics, Contaminated data, Model-Based clustering, Classification attributes

Klíčová slova: Separace signálů, Akustická emise, Částicová fyzika, Znečištěná data, Model-Based klastrování, klasifikační atributy

JEL classification: C61

1. Model-Based Clustering (MBC) principle

We are interested in testing Model-Based Clustering method (MBC) to separate data which originate from different types of materials. We also introduce a new attribute characterizing the signals of acoustic emission. The whole experiment and results of classification are described in the first part of the paper. In the second part we deal with heavily contaminated data representing the data sets obtained from particle accelerator in Fermilab. For this purpose we use pseudo-random generated data which are liable to normal and uniform distribution. Our goal here was to find contamination level in which MBC method fails to separate data. We also tested the success rate of MBC method in positive cases of classification for both experiments. We work with a finite convex combination of probability density functions which belongs to observation samples. Each component of the mixture represents one cluster. The best normal mixture fit is done by the maximum likelihood principle performed by the well-known iterative EM algorithm which allows us to work with incomplete data sets. To get

more information about MBC method and EM algorithm, see [1, 2, 3].

2. AE experiment & Signal attributes

As the source of acoustic emission (AE) signals we used pentest. These signals were detected by piezoelectric sensors situated on the surface of a measured sample. To store data we used measuring device DAKEL-XEDO with sampling rate 4 MHz. Onto each material sample we placed two AE sensors as well as we the source of acoustic emission was excited between the sensors. The device is able to distinguish the voltage between -2400 mV and 2400 mV. In the next Table 1 we describe specimens of materials used throughout the experiment.

Table 1: Description of experimental specimens and its structure

Type of material	*Shape and Dimensions of the Specimen*	*Surface*	*Number of measurements*
Spruce	cuboid: 56,0 x 3,8 x 5,9 cm	smooth	76
Cherry tree	cylinder: $r_p \approx 7$ cm, $v \approx 2$ cm	very coarse	67
Glass	bottle of wine $0{,}7$ l	smooth	95
Granite	shard of stone	coarse	60
Wall tile 1	slab: 80 x 80 x 4 mm	smooth	86
Wall tile 2	slab: 62 x 60 x 6 mm	smooth, coarse	101
Piece of metal	sheet of steel in T shape	smooth	106
Plastic	ruler 30cm long	worn with use	52

Our data analysis is performed through a different sets of the classification attributes which are computed directly from the monitored signals $\{x_t\}_{t=0}^{T-1}$ or are based on its normalized spectrums $\{\Xi(f)\}_{f=0}^{T-1}$ obtained by discrete Fourier transform.

- Attribute

$$W_a = argmin_{j\in\langle 0,T-1\rangle} \sum_{f=0}^{T-1} |\, j-f \,| \left| \Xi(f) - \overline{\Xi}_f \right|^a,$$

$$\overline{\Xi}_f = \frac{1}{T}\sum_{f=0}^{T-1}\Xi(f) = \frac{1}{T}, \qquad a \in \langle 1,+\infty).$$

- Attribute

$$Q_b = \min\{m \in \langle 0,T-1\rangle \,|\, \sum_{f=0}^{m}\Xi(f) \geq b\}, \quad b \in (0,1).$$

- Attribute

$$Z_c = \sum_{t=\min J}^{T-1} d(x_t), \quad \text{where } J = \{m \in \langle 0,T-1\rangle \,|\, x_m \geq c \max_{t\in\langle 0,T-1\rangle} |\, x_t \,|\},$$

$$d(x_t) = \begin{cases} 1, & sgn(x_t x_{t+1}) = -1, \\ 0, & sgn(x_t x_{t+1}) \neq -1. \end{cases}$$

- Attribute

$$M_g = \sum_{t=\min J}^{T-1} \Delta(x_t), \quad \text{where} J = \{m \in \langle 1,T-1\rangle \,|\, x_m \geq g \max_{t\in\langle 0,T-1\rangle} |\, x_t \,|\},$$

$$\Delta(x_t) = \begin{cases} 1, & x_t \text{ is a local extreme of signal}, \\ 0, & x_t \text{ isn't local extreme of signal}. \end{cases}$$

The first three attributes are taken from [1, 2]. Attribute $M\gamma$ is recently introduced.

3. Results of material structure classification

The classification was done by means of a different sets of attributes. We used couples, triplets and quartet of attributes. In each cases we used all the possible combinations of attributes. In the case of coupled attributes, the results are given in the following Table 2. For the best choice of the attributes couple we present also the schematic illustration in Figure 1.

Table 2: Classification successfulness for couples of the attributes

Attributes	$Z_{\frac{1}{20}}, Q_{0.15}$	$M_{\frac{1}{20}}, Q_{0.15}$	$M_{\frac{1}{20}}, W_2$	$M_{\frac{1}{20}}, Z_{\frac{1}{20}}$	$Q_{0.15}, W_2$	$Z_{\frac{1}{20}}, W_2$
Success rate	50.70%	46.97%	39.81%	**56.10%**	41.52%	50.39%

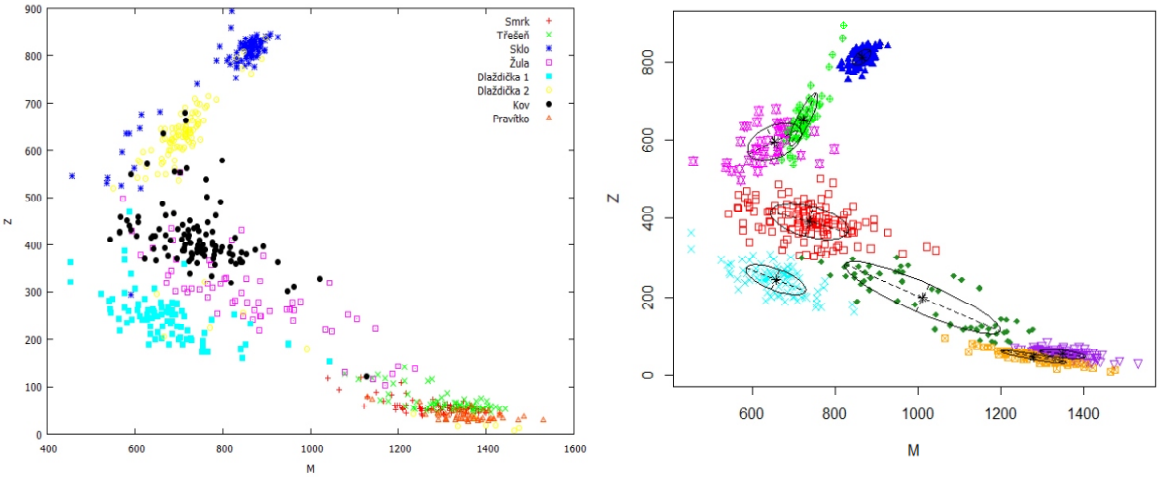


Figure 1: ***Results for the best couple of classification attributes M and Z (a) before classification, (b) after classification.***

Now we try to improve the achieved success rate by adding another attribute so we carry on the classification experiment using triplets of attributes. The corresponding results are shown in the following Table 3 accompanied by Figure 2 describing the situation for the best triplet of attributes.

Table 3: Results of material classification for different triplets of attributes

Attributes	$Z_{\frac{1}{20}}, Q_{0.15}, M_{\frac{1}{20}}$	$Z_{\frac{1}{20}}, W_2, M_{\frac{1}{20}}$	$Z_{\frac{1}{20}}, W_2, Q_{0.15}$	$Q_{0.15}, W_2, M_{\frac{1}{20}}$
Success rate	**65.63%**	63.92%	63.61 %	52.93%

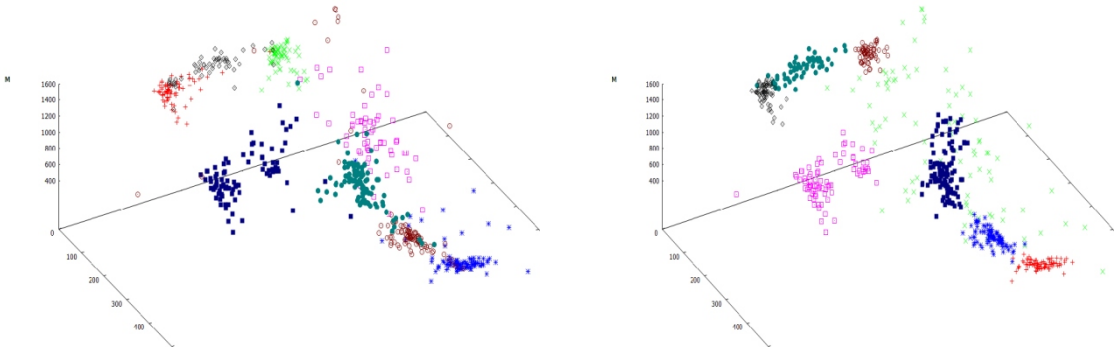


Figure 2: Results for the best triplet (Z, Q, M) of classification attributes (a) before classification, (b) after classification.

In the case of quartet of the attributes $Z_{1/20}$, $Q_{0.15}$, $M_{1/20}$, W_2, we achieved the success rate of 68.75%.

4. Heavily contaminated data (Fermilab)

A motivation for this experiment was the situation which occurs on huge particle accelerators, e.g. in Fermilab, where the majority of the observed data comes from a background with respect to a specific particle decay process. To test an efficiency of our separating algorithms, first, instead of using real data, we work with pseudo-random data sets generated. The whole situation is described by the next Table 4.

Table 4: Particle physics example with simulated contaminated data

Type of data	*Mean value*	*Covariance matrix*	*Number of occurrences*
1^{st} cluster	(5, 5)	$\begin{pmatrix} 1.0 & 0 \\ 0 & 0.5 \end{pmatrix}$	100
2^{nd} cluster	(35, 24)	$\begin{pmatrix} 0.2 & 0 \\ 0 & 1.0 \end{pmatrix}$	100
Normally distributed contamination	(5, 5)	$\begin{pmatrix} 30 & 0 \\ 0 & 30 \end{pmatrix}$	2000 - 20000
uniformly distributed contamination	uniform	uniform	2000 - 4000

As we can see, the mean values of normally distributed clusters and the contamination along with the corresponding covariance matrices are chosen in order to see what happens if the mean values will be overlapped and also if there occur another cluster in the middle part of contamination. For this setup we get very interesting separation results via MBC classification method. For illustration purposes we present the Figure 3 in order to see what are we trying to deal with. Consequently, the results of classification are summed up in Tables 5 and 6.

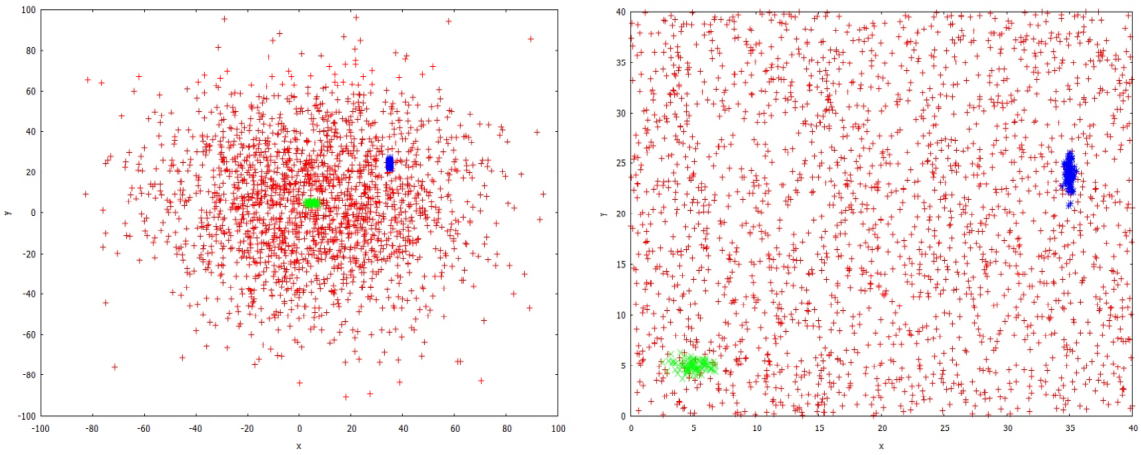


Figure 3: Illustration of heavily contaminated data for two small clusters (a) normal contamination, (b) uniform contamination

Table 5: Overall results for the normally distributed contamination

Level of contamination	*90%*	*90%*	*95%*	*95%*	*97%*	*97%*	*98%*
Number of clusters	3	5	5	6	8	9	8
Discovered clusters	1	1, 2	2	1, 2	1	1, 2	1
Success rate	99.73%	99.83%	99.54%	99.38%	99.97%	99.96%	98.76%

Table 6: Overall results for the uniformly distributed contamination

Level of contamination	90%	90%	90%	90%	95%	95%	95%
Number of clusters	4	6	7	9	5	8	9
Discovered clusters	0	0	1, 2	1, 2	0	0	1
Success rate	—	—	97,86%	99.68%	—	—	96.79%

In Table 5 we can see unexpected results for 97% and 98% level of contamination. The MBC method found the more contaminated cluster with lower number of expected clusters in our data. In the case of uniform contamination we can see that it is needed to expect much more clusters to be successful. We got very high success rates in all positive cases of the cluster separation.

5. Conclusions

The paper was experimentally oriented and we tried to test the efficiency of Model Based Method of cluster separation or signal classification. Our signals originated from the Acoustic Emission environment or from Fermilab huge particle accelerator and detector. We succeed in AE signal classification for different choices of materials under consideration only partially since the AE signals are not so strongly influenced by the passing through the given material structure as we expected. For the Fermilab inspired data sets we achieved considerably good results in the cluster separation corresponding to some specific decay elementary particle. Thus the particle would be precisely detected even if the background contamination form over 98% of the overall data sample at the disposal.

Acknowledgements. This work was supported by the grants SGS12/197/OHK4/3T/14, MSMT INGO-II LG12020, and by the research program of the Ministry of Education of Czech Republic under the contract MSM 6840770039.

References

[1] ZUZANA FAROVÁ. *Statistické metody odhadu hustot a klasifikace signálu*. Diplomová práce FJFI ČVUT v Praze, Praha, 2010.

[2] JAN TLÁSKAL. *Statistické klasifikační metody v akustické emisi*. Diplomová práce FJFI ČVUT v Praze, Praha, 2008.

[3] TANNER, M. A.; *Tools for Statistical Inference, Third Edition*. Springer-Verlag New-York, Inc., 1996.

[4] WITOLD PEDRYCZ. *Knowledge-Based Clustering: From Data to Information Granules.* A Wiley-Interscience publication, ISBN 0-471-46966-1, USA, 2005.

[5] VÁCLAV KŮS, DOMINGO MORALES, IGOR VAJDA. Extension of the Parametric Families of Divergences Used in Statistical Inference. *Kybernetika*, 44/1, 95-112, 2008.

Authors addresses:

Václav Kůs, Ing, PhD.
Katedra matematiky
FJFI ČVUT
Trojanova 13, 120 00 Praha 2
vaclav.kus@fjfi.cvut.cz

Jan Vejmola, Bc.
Katedra matematiky
FJFI ČVUT
Trojanova 13, 120 00 Praha 2
vejmojan@fjfi.cvut.cz

Jiří Franc, Ing.
Katedra matematiky
FJFI ČVUT
Trojanova 13, 120 00 Praha 2
francjir@fjfi.cvut.cz

Deriváty na počasie - oceňovanie basket call opcií na HDD index s použitím stochastickej volatility

Weather derivatives – pricing of basket call options on HDD index with stochastic volatility

Marko Lalić, Zuzana Gordiaková, Martina Rusnáková

Abstract: The paper is focused on pricing of basket options in group of weather derivatives, specifically for call options on HDD index. Pricing is applied on average temperatures of 3 Slovak cities – Košice, Bratislava and Poprad. The paper includes brief description of basic stochastic processes for pricing of this type of derivatives. These processes have been applied in simulations to determine differences between the pricing with deterministic and stochastic volatility.

Abstrakt: Práca sa zaoberá oceňovaním basket opcií na deriváty v skupine derivátov na počasie, konkrétne na HDD call opcie. Oceňovanie je aplikované na priemerné hodnoty teplôt 3 slovenských miest – Košice, Bratislava a Poprad. Práca obsahuje stručný popis základných stochastických procesov určených na oceňovanie tohto typu derivátov. Tieto procesy boli aplikované pri simuláciách, ktoré určili rozdiely pri oceňovaní týchto derivátov prostredníctvom deterministickej a stochastickej volatility.

Key words: HDD index, basket options pricing, stochastic volatility, risk premium, volatility risk premium

Kľúčové slová: HDD index, oceňovanie basket opcií, stochastická volatilita, riziková prémia, riziková prémia volatility

JEL classification: G12

Úvod

Deriváty na počasie predstavujú istú formu zaistenia proti rizikám plynúcim z vývoja počasia. Podkladovým inštrumentom derivátov zameriavajúcich sa na teploty sú väčšinou indexy teplôt zachytávajúce určité obdobie. Najpopulárnejšie indexy sú HDD (heating degree day) alebo CDD (cooling degree day). Tieto indexy odrážajú nakoľko bolo sledované obdobie náročné napr. na spotrebu energií alebo ako nepriamo ovplyvňovali vývoj ekonomickej aktivity (napr. turistický ruch, stavebný priemysel alebo v neposlednom rade aj poľnohospodárstvo). Basket opcie predstavujú možnosť zaistiť sa celoplošne, resp. použiť portfólio HDD indexov.

1. Deriváty na počasie

Pre indexy teplôt HDD a CDD platí:

$$HDD(t_s, t_e, HDD_t) = \sum_{i=t_s}^{n=t_e} (HDD_t - T_A)^+ \quad (1)$$

$$CDD(t_s, t_e, CDD_t) = \sum_{i=t_s}^{n=t_e} (CDD_t - T_A)^+ \quad (2)$$

kde t_s predstavuje čas kedy sa začína počítať index, t_e čas kedy sa index prestáva počítať a HDD_t limitnú teplotu, pri ktorej sa do indexu započítavajú rozdiely. V našej práci budeme používať najčastejšiu teplotu $HDD_t = 18$ °C.

Deriváty, najmä opcie (call a put) sú charakteristické svojimi výplatnými funkciami, pre ktoré platí:

$$p(call) = (HDD - K)^+$$
$$p(put) = (K - HDD)^+ \quad (3)$$

Táto práca je zameraná na basket opcie, ktoré predstavujú alternatívu zaistenia portfólia. Pre výplatné funkcie basket opcie (call a put) platí:

$$p(call) = (P - K)^+$$
$$p(put) = (K - P)^+ \quad (4)$$

kde P je:

$$P = \sum_{i=1}^{N} w_i HDD_i(t_s, t_e, HDD_t), \qquad \forall\, i:\ w_i > 0. \quad (5)$$

2. Stochastické procesy pri oceňovaní derivátov na počasie

Pre potreby modelovania teploty využijeme model, ktorý sa využíva aj na modelovanie úrokových sadzieb. Ide o Hull Whiteov model, resp. stochastickú diferenciálnu funkciu, ktorá popisuje tento model:

$$dT_i(t) = \big(\alpha_i(t) - \beta_i(t)T(t)\big)dt + \sigma_i(t)dW_i(t) \quad (6)$$

kde podľa Alatona (2002) platí:

$$\alpha_i(t) = \frac{dT^m{}_i(t)}{dt} + \theta_i T^m{}_i(t), \qquad \beta(t) = \theta_i \quad (7)$$

Z rovníc (6) a (7) je zrejmé, že sa jedná o mean reverting proces, ktorý má tendenciu sa vracať ku hodnote $T^m{}_i(t)$, ktorá je funkciou času. Podľa Dornier & Queruela je k modelovaniu teploty potrebné dodať aj člen $\frac{dT^m{}_i(t)}{dt}$ (viď rovnicu (7)). $\sigma_i(t)$ predstavuje smerodajnú odchýlku teplôt. Riešením tejto rovnice podľa Shrevea (1997) je:

$$T(\tau).\exp(K(\tau - h)) = \mathrm{T_i(h)} + \int_{\mathrm{h}}^{\tau} (\exp(K_i(s)).\alpha(s))ds + \int_{\mathrm{h}}^{\tau} \sigma_i(s).\exp(K(s))dW_i(s)$$
$$\mathrm{K_i}(\tau) = \int_h^{\tau} \beta_i(t)dt = \theta_i(\tau - h)$$
$$T_i(\tau) = (-T_i^m(h) + T(h)).\exp\big(-\theta_i(\tau - h)\big) + T^m{}_i(\tau) + \int_h^{\tau} \Big(\exp\big(\theta_i(s - \tau)\big)\Big)\sigma_i(s)dW_i(s) \quad (8)$$

pričom deterministickú teplotu môžeme definovať ako

$$T_i^m(t) = A_i + B_i t + C_i.\sin(\omega t + \varphi_i) \quad (9)$$

Pre basket opcie a pre možnosť stochastickej volatility je však potrebné uvažovať aj o korelovanom Brownovom pohybe, pre ktorý platí:

$$W(t) = HZ(t) \quad (10)$$

kde W(t) predstavuje n-dimenzionálny korelovaný Brownov pohyb, ku ktorému existuje $Z(t) = (Z_1(t), Z_2(t), \ldots \ldots Z_n(t))$ t.j. n-dimezionálny nezávislý Brownov pohyb a H predstavuje maticu z Choleského dekompozície korelačnej matice P, pre ktorú platí:

$$P = HH^T \quad (11)$$

Možností, ako modelovať volatilitu, je viacero. Prvou z nich je určiť volatilitu ako deterministickú, t.j. ako konštantu alebo ako periodickú funkciu času. Podľa Mraoua & Bari je však vhodné aby volatilita bola stochastická, vypočítaná zo smerodajných odchýlok:

$$d\sigma_i(t) = \kappa_i\big(\bar{\sigma}_i - \sigma_i(t)\big)dt + \phi_i dW_j(t) \tag{12}$$

kde, κ_i predstavuje speed of reversion parameter a ϕ_i je volatilita volatility. Je potrebné pripomenúť, že medzi dW_j a dW_i je korelácia a tieto Brownové pohyby sú zostrojené podľa rovníc (10) a (11).

Na odhad parametrov sme použili metódu najmenších štvorcov, pričom model a ostané odhady vychádzajú z Alatona (2001) a Mraoua & Bari**:**

Rovnica modelu a úprava parametrov	Parametre stochastickej volatility
$T(t) = a_{i1} + a_{i2}t + a_{i3}\sin(\omega t) + a_{i4}\cos(\omega t)$ $A_i = a_{i1},\ B_i = a_{i2},$ $C = \sqrt{a_{i3}^2 + a_{i4}^2},\ \varphi_i = \arctan\left(\frac{a_{i4}}{a_{i3}}\right) - \pi$ $\hat{\sigma}^2_{t-l}(t) = \frac{365}{l}\sum_{j=t-l}^{t}(T(j+1) - T(j))^2$	$\bar{\sigma}_i = \frac{1}{N}\sum \sigma_i(j)$ $\kappa_i = 365\left[-log\left(\frac{\sum_{i=1}^{n} Y_{i,j-1}\{\sigma_i(j) - \bar{\sigma}_i\}}{\sum_{j=1}^{n} Y_{i,j-1}\{\sigma_i(j-1) - \bar{\sigma}_i\}}\right)\right]$ $\phi_i = \sqrt{\frac{1}{N}\sum\big(\sigma_i(j) - \sigma_i(j-1)\big)^2}$
$\hat{\theta}_i = 365\left[-log\left(\frac{\sum_{i=1}^{n} Y_{i,j-1}\{T_i(j) - T_i{}^m(j)\}}{\sum_{j=1}^{n} Y_{i,j-1}\{T_i(j-1) - T_i{}^m(j-1)\}}\right)\right],\quad Y_{i,j-1} = \frac{-T_i(j-1) + T_i{}^m(j-1)}{\hat{\sigma}^2_{i,j-l}(j)}$	

(13)

3. Oceňovanie opcií

Pre cenu opčných prémií platí:

$$c(call) = \mathbb{E}^{Q(\lambda,\xi)}[(HDD - K)^+]$$
$$c(put) = \mathbb{E}^{Q(\lambda,\xi)}[(K - HDD)^+] \tag{14}$$

kde $Q(\lambda, \xi)$ predstavuje mieru s rizikovými prémiami λ a ξ, ktoré upravujú rovnice (6) a (12) na tvar:

$$dT_i^Q(t) = \big(\alpha_i(t) - \beta_i(t)T(t)\big)dt + \sigma_i(t)dW_i^Q(t) - \lambda\sigma_i(t)dt$$
$$d\sigma_i^Q(t) = \kappa_i\big(\bar{\sigma}_i - \sigma_i(t)\big)dt + \phi_i dW_j^Q(t) - \xi\phi_i dt \tag{15}$$

resp. vzťahy v (15) môžeme napísať ako:

$$\mathbb{E}^{Q(\lambda,\xi)}\big[T_\tau | T_h, \sigma_h\big] = \mathbb{E}^P\big[T_\tau | T_h, \sigma_h\big] - \int_h^\tau \lambda\,\sigma_i(t)\exp\left(-\theta_i(\tau - t)\right)dt$$
$$\mathbb{E}^{Q(\lambda,\xi)}[\sigma_\tau | \sigma_h] = \mathbb{E}^P[\sigma_\tau | \sigma_h] - \int_h^\tau \xi\,\phi_i(t)\exp\left(-\kappa_i(\tau - t)\right)dt \tag{16}$$

Pomocou odhadnutých parametrov (Tab.1) sme prostredníctvom rovníc (15 resp. 16), ktoré sme nasimulovali, odhadli hodnoty opčných prémií basket opcií pomocou deterministickej volatility, pričom portfólio je definované ako $P = \sum_{i=1}^{N} w_i HDD_i(t_s, t_e, HDD_t)$.

Parametre opcií:

- použité mestá/teploty KE_A, BA_A, PP_A,
- $w_1 = w_2 = w_3 = 1$,

- čas začatia indexu : t_start=250/365,
- čas ukončenia indexu (čas splatnosti opcie): t_end=(260/365, 265/365,.....450/365),
- počet iterácií 10000.

Pre rozdiely, ktoré sú načrtnuté v grafoch platí:

$$R(\lambda, \xi, K, t_{start,} \dots\dots) = c_{SV}(call, K, w, T_0, \sigma_0, t_{start}, t_{end}, \lambda, \xi) - c_{DV}(call, K, w, T_0, t_{start}, t_{end}, \lambda)$$

Tab. 1: Odhadnuté parametre a korelačné koeficienty

	KE_A	BA_A	PP_A
A	9,6613	10,8978	6,577
C	11,6676	11,1537	10,9735
φ	-1,8141	-1,8128	-1,8397
θ	78,526	85,892	92,205
$\bar{\sigma}$	41,911	44,036	49,434
ϕ	1,5881	1,733	1,9306
κ	8,3821	7,2199	5,5275

determinsitická volatilita			
	KE_A	BA_A	PP_A
$\bar{\sigma}$	77,852	83,76	87,188
θ	42,65	42,65	42,65

korelačná matica P			
	KE_A	BA_A	PP_A
KE_A	1,00	0,81	0,87
BA_A	0,81	1,00	0,84
PP_A	0,87	0,84	1,00
V_KE_A	-0,05	-0,06	-0,05
V_BA_A	-0,01	0,00	-0,01
V_PP_A	-0,10	-0,09	-0,09

	V_KE_A	V_BA_A	V_PP_A
V_KE_A	1,00	0,63	0,67
V_BA_A	0,63	1,00	0,65
V_PP_A	0,67	0,65	1,00

Zdroj: Vlastné spracovanie

4. Rozdiely v ohodnocovaní

Hlavným predmetom práce je znázorniť rozdiely medzi dvomi spôsobmi oceňovania – oceňovanie s deterministickou alebo so stochastickou volatilitou. Nasledujúce grafy znázorňujú rozdiely v cenách, pre ktoré platí:

$R(\lambda, \xi, K, t_{start,} \dots\dots) = c_{SV}(call, K, w, T_0, \sigma_0, t_{start}, t_{end}, \lambda, \xi) - c_{DV}(call, K, w, T_0, t_{start}, t_{end}, \lambda)$,

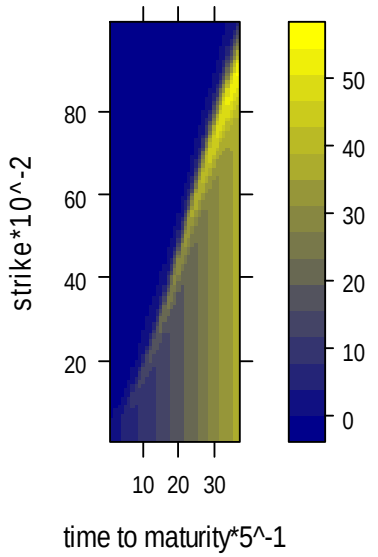
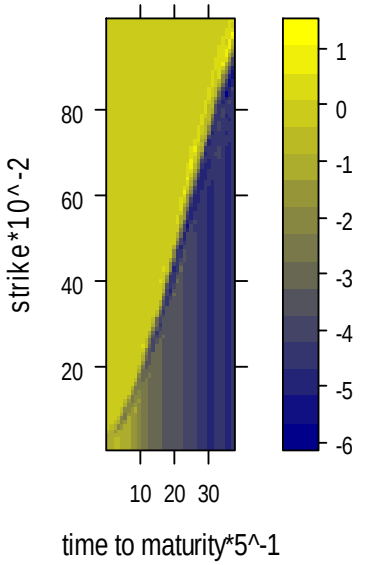
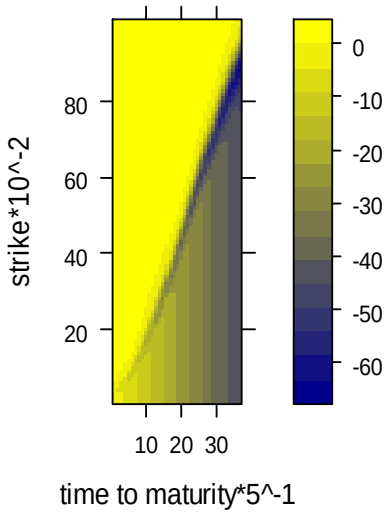
pričom ceny sú počítané ako rozdiely opčných prémií basket opcii, ktoré spĺňajú parametre popísané v Tab. 1.

Najvýraznejšie hodnoty R vznikajú pri realizačných cenách, ktoré sú zhodné s očakávanou hodnotou indexu. V prípade, že $\xi = 0$, nevznikajú výrazné rozdiely v hodnotách opčných prémií. Opčné prémie so stochastickou volatilitou majú nižšiu hodnotu. V prípade, že $\lambda = -10$, je hodnota opčných prémií nižšia pre opcie vypočítané so stochastickou volatilitou s realizačnou cenou nižšou ako stredná hodnota indexu. V opačnom prípade, keď $\lambda = 10$, sú tieto opčné prémie pri realizačných cenách nižších ako stredná hodnota indexu vyššie, a to hlavne v prípadoch dlhších splatností.

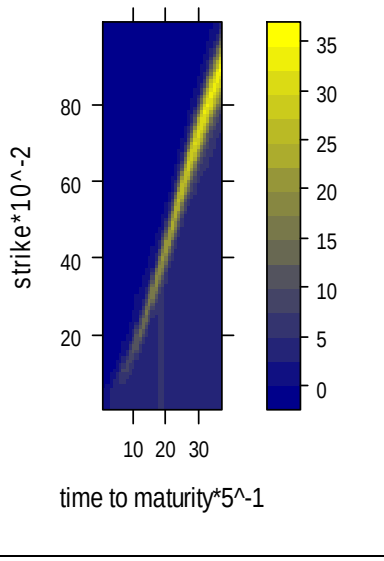
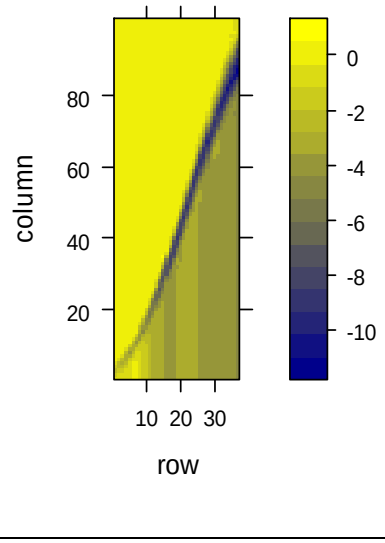
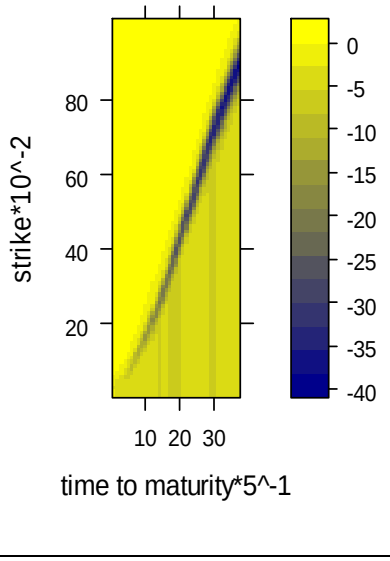
V prípade kladnej prémie volatility dochádza naopak k výrazným kladným hodnotám $R(\lambda, \xi \dots\dots)$ pri realizačných cenách, ktoré sú zhodné so strednou hodnotou indexu. Pri kladnej rizikovej prémii dochádza aj k navýšeniu hodnoty opčných prémií počítaných so stochastickou volatilitou pri nižších realizačných cenách ako je stredná hodnota indexu. Tieto kladné rozdiely rastú hlavne s rastom doby do splatnosti. V prípade nulovej rizikovej prémie a $\xi = 2{,}5$ sú výrazné rozdiely len v okolí realizačných cien zhodných so strednou hodnotou indexu. V prípade kladnej rizikovej prémie dochádza k záporným rozdielom pri nižších realizačný cenách medzi spôsobom počítania so stochastickou volatilitou a deterministickou, avšak pri realizačných cenách zhodných so strednou hodnotou indexu ostávajú rozdiely kladné.

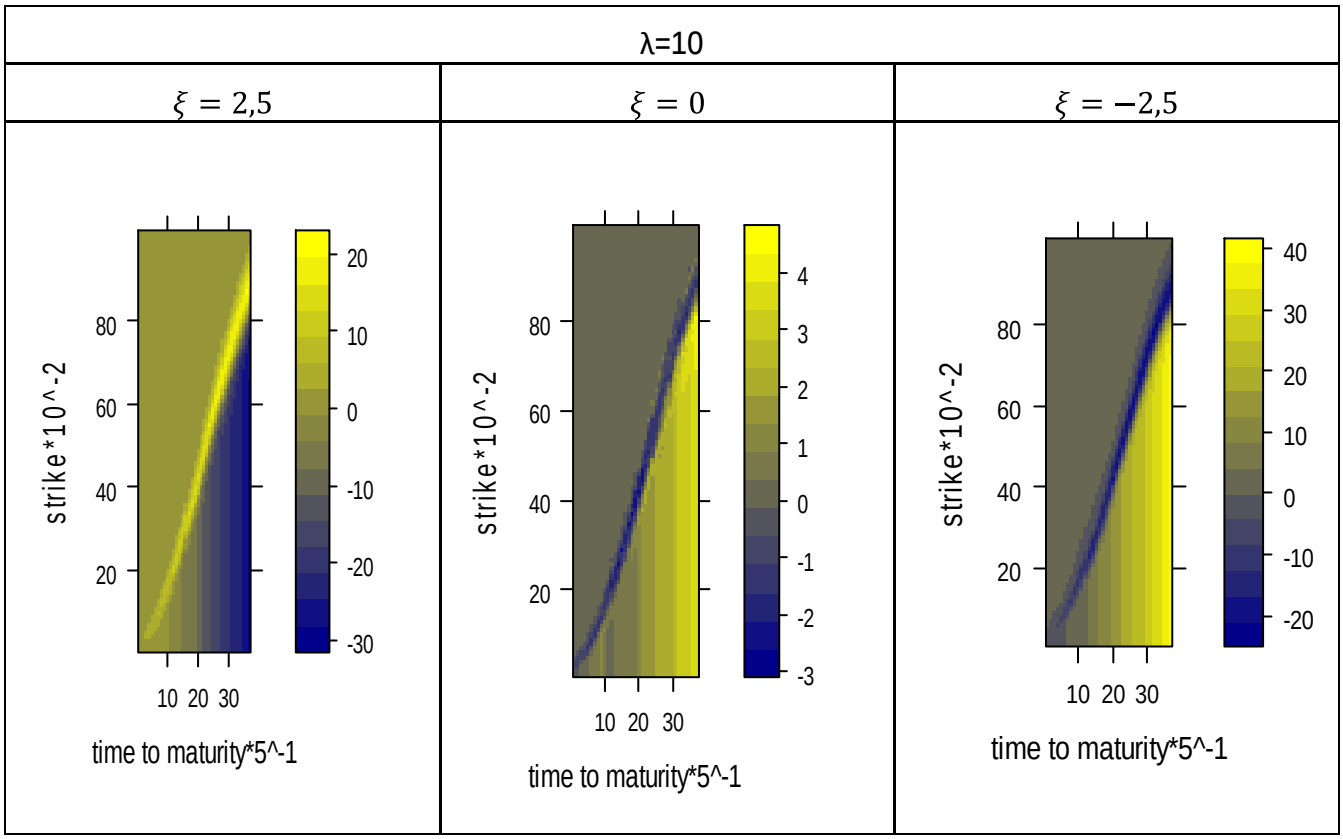
V prípade zápornej rizikovej prémie volatility dochádza ku záporným rozdielom medzi hodnotami opčných prémií počítaných so stochastickou a s deterministickou volatilitou. Vplyv rizikovej prémie je opačný ako v prípade, keď $\xi = 2{,}5$.

λ=-10

$\xi = 2{,}5$	$\xi = 0$	$\xi = -2{,}5$
		

λ=0

$\xi = 2{,}5$	$\xi = 0$	$\xi = -2{,}5$
		



Obr. 1: parameter R – porovnanie cien opčných prémií
Zdroj: Vlastné spracovanie

5. Záver

Oceňovanie pomocou stochastickej volatility predstavuje presnejší spôsob oceňovania opcií ako pri oceňovaní s deterministickou volatilitou. Kým pri nulových rizikových prémiách nie sú rozdiely výrazné (najvýraznejšie pri realizačných cenách zhodných so strednou hodnotou indexu), pri výraznejšej zmene miery (väčšia zmena rizikových prémií) hlavne v oblasti rizikovej prémie volatility sa objavujú veľmi výrazné zmeny. Riziková prémia volatility mení nielen očakávanú volatilitu, ktorá priamo vstupuje do oceňovania, ale taktiež sa mení aj vplyv rizikovej prémie teploty. Pri kladnej rizikovej prémii volatility sa ceny opcií nadhodnocujú hlavne pri realizačných cenách zhodných so strednou hodnotou HDD indexu. Naopak pri zápornej rizikovej prémii volatility sme pozorovali výrazný záporný rozdiel medzi cenami opcií vypočítanými so stochastickou volatilitou a deterministickou volatilitou opäť hlavne pri realizačných cenách zhodných so strednou hodnotou HDD indexu. Riziková prémia teploty má vplyv hlavne na opčné prémie pri realizačných cenách pod očakávanou hodnotou indexu. Rizikové prémie ako také definujú aj možnosť predaja/nákupu opcií na trhu a možnosť znášať špecifické riziko počasia. Keďže neexistuje žiaden likvidný trh s týmito derivátmi, je posudzovanie hodnoty derivátov vzhľadom na použitie stochastickej volatility pri rôznych rizikových prémiách dôležitou súčasťou pri riadení rizika a prípadne pri tvorbe štruktúrovaných produktov.

Literatúra

[1] ALATON, P. – DJEHICHE, B. – STILLBERGER, D: *On modeling and pricing weather derivatives.* [online]. [cit. 2012-22-10] Dostupné na internete: <http://www.treasury.nl/files/2007/10/treasury_301.pdf>

[2] BEAUMONT, P. H:: *Financial Ebgineering Principes*. Chichester: John Wiley & Sons, Ltd, 2004. ISBN 0-471-46358-2.

[3] BENTH, F.E. – BENTH, J.S. The volatility of temperature and pricing of weather derivatives, *Quantitative Finance*, 2007. [online]. [cit. 2012-20-10]. Dostupné na internete: < http://www.tandfonline.com/doi/abs/10.1080/14697680601155334 >

[4] DORNIER, F. - QUERUEL, M. Caution to the Wind, Weather Risk Special Report 2000, *Energy & Power Risk Management/Risk Magazine*

[5] JÄCKEL, P.: *Monte Carlo Methods in Finance*. Chichester: John Wiley and sons, 2002, ISBN 0 471 49741 X

[6] TALEB, N.: *Dynamic Hedging: Managing vanilla and exotic options.* New York: John Wiley and Sons Inc., 1997. ISBN 0-471-15280-3.

[7] SHREVE, S. – CHALASANI, P.- SOMESH, J: *Stochastic Calculus and Finance*, 1997, [online]. [cit. 2012-20-10]. Dostupné na internete:
<http://www.google.sk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&sqi=2&ved=0CHgQFjAD&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.137.6951%26rep%3Drep1%26type%3Dpdf&ei=_-GoT7u9AsrE4gTrvtTQCQ&usg=AFQjCNFnXvEwtfwS81tAL5K_EJag6pB6nQ&sig2=qPo96RWfWyN-HWaZqRKfig>

[8] WEERT F.: *Exotic Options Trading*. John Wiley and Sons Ltd, 2008. ISBN 978-0-471-51790-1.

[9] WILLMOTT P. – HOWINSON S. – DEWYNNE J.: *The Matematics of Financial Derivates*. Cambridge: Cambridge University Press, 1996. ISBN 0-521-49789-2.

[10] GREEN, W. H. 1997. *Econometric Analyses*. London: Prentice – Hall, 1997. 1076 s. ISBN 0-13-7246659-5.

[11] MRAOUA M. - BARI D.: *Temperature stochastic modeling and weather derivatives pricing: empirical study withMoroccan data*, 2005. [cit. 2012-25-10]. Dostupné na internete: <http://www.univ-rouen.fr/LMRS/JSEF06/JSEF1_fichiers/mraoua.pdf>

Adresa autorov:

Ing. Marko Lalić
Ekonomická Fakulta TUKE
Katedra Financií
Boženy Nemcovej 32, 040 01 Košice
marko.lalic@tuke.sk

Ing. Zuzana Gordiaková
Ekonomická Fakulta TUKE
Katedra Financií
Boženy Nemcovej 32, 040 01 Košice
zuzana.gordiakova@tuke.sk

Ing. Martina Rusnáková, PhD.
Ekonomická Fakulta TUKE
Katedra Financií
Boženy Nemcovej 32, 040 01 Košice
martina.rusnakova@tuke.sk

Spotrebiteľské správanie a preferencia nákupu bioproduktov

Consumer behavior and preference of organic products

Vanda Lieskovská, Silvia Megyesiová, Katarína Petrovčíková

Abstract: The aim of this paper is to present partial results, which are part of the project VEGA 1/0906/11. It builds on the importance of linking the issue of civilizational challenges, which is part of the increasing importance of health and quality of life in the context of more active use of domestic raw materials for current use, protection and reproduction of biological resources and the use of domestic raw materials. The line of the connection of these two planes evoked the need to examine the interactions between the market with a range of organic food and products, trade and individual consumption. The main idea is to analyze consumer behavior.

Abstrakt: Cieľom príspevku je prezentovať čiastkové výsledky, ktoré sú súčasťou riešenia projektu VEGA 1/0906/11. Nadväzuje na dôležitosť prepojenia problematiky civilizačných výziev, súčasťou ktorých je zvyšovanie významu zdravia a kvality života v kontexte aktívnejšieho využívania domácich surovinových zdrojov za súčasného využívania, ochrany a reprodukcie biologických zdrojov a využitia domácich surovinových zdrojov. Línia prepojenia uvedených rovín evokovala potrebu skúmania vzájomných interakcií medzi trhom so sortimentom biopotravín a bioproduktov, obchodom a individuálnou spotrebou. Nosnou myšlienkou je analýza spotrebiteľského správania.

Key words: organic, consumer behavior, consumer behavior research.

Kľúčové slová: bioprodukty, spotrebiteľské správanie, prieskum spotrebiteľského správania.

JEL classification: M 32, P 46, Q 57

Úvod

Revízia európskej „Biovyhlášky“ (Nariadenie Rady č. 2092/91) spôsobila vznik početných iniciatív usilujúcich sa presadiť určujúce smerovanie ekologického poľnohospodárstva do praxe. Na uvedené podnety reagoval aj Program rozvoja vidieka SR 2007 – 2013, prostredníctvom ktorého bolo možné podporiť rozširovanie pestovania bioproduktov, vrátane zaostalých regiónov Slovenska. Ekologický konzumerismus v západnej Európe, Severnej Amerike ale aj v nových členských krajinách EÚ postupne narastá. Spotrebitelia sa stávajú náročnejšími na kvalitu produktov, pričom kvalita už nestojí oddelene od ekologických požiadaviek. Mení sa spotrebiteľské správanie a vnímanie kvalitného produktu. S rastúcim počtom environmentálne citlivých spotrebiteľov sa environmentálny marketing a trhy s ekologickými produktmi stávajú úspešnými.

Ukazovateľom ekologickej uvedomelosti spotrebiteľov sa stáva spotreba produktov, ktoré sú vyrobené ekologickými postupmi. Alternatívou je aj preferovanie nákupu domácich produktov. Špecifickou alternatívou je aj nákup z dvora. Projekt na podporu aktivít súvisiacich s predajom z dvora bol uvedený do života v roku 2009. Získal podporu z blokového grantu EMVO TUR nadácie EKOPOLIS, bol spolufinancovaný z Finančného mechanizmu Európskeho hospodárskeho priestoru, Nórskeho finančného mechanizmu, štátneho rozpočtu SR a rozpočtu EKOTREND Slovakia. V súčasnosti združuje prvovýrobcov vyrábajúcich z vlastných surovín podľa vlastných receptúr prevažne ručne, ktorí predávajú priamo konečným spotrebiteľom pod hlavičkou **predaj z dvora**.

1. Spotrebiteľské správanie pri nákupe bioproduktov

Pri sledovaní spotrebiteľského správania týkajúceho sa nákupu bioproduktov je potrebné prehodnotiť ako motívy, tak aj bariéry nákupu biopotravín. Pri motívoch je rozhodujúcim faktorom kúpy produktu jeho potreba. Konkretizácia potreby je odrazom motívu, ktorý kupujúceho stimuluje. Zmena vnímania hodnôt v spoločnosti a vývoj spoločenského vedomia profilujú i smer kúpnych motívov. Tak je tomu aj u komodity biopotraviny, kde v ostatnom čase prevládajú motívy pre ich kúpu konzumentmi ako sú: snaha vyskúšať nový trend v stravovaní, uplatňovanie zdravého životného štýlu, zdravotná indikácia, podpora ekologického poľnohospodárstva. K základným požiadavkám na kvalitu biopotravín patrí ich vysoká kvalita, zdravotná a ekologická istota ako aj ekologický pôvod. Medzi hlavné príčiny bariér kúpy bioproduktov patria:

a) kvalitatívna bariéra – časť biopotravín je charakteristická nezvyklou chuťou alebo vzhľadom a konzument má malú možnosť výberu v porovnaní so širokou ponukou konvenčných potravín,
b) cenová bariéra – predstavuje jednu z najdôležitejších bariér pri tvorbe trhu eko-potravín. Vysoká cena prekračuje pripravenosť konzumenta tovar kúpiť,
c) situačná bariéra – biopotraviny nie je spravidla dostať na všetkých miestach, kde sa predávajú potraviny,
d) zvyková bariéra – konzumenti zostávajú často naviazaní na nákupy v predajniach, kde už dlhšie obdobie nakupujú a nie sú pripravení vyhľadať ďalšie obchody, kde už je i ponuka ekologických potravín,
e) motivačná bariéra – ide o skutočný nezáujem o biopotraviny a ochranu životného prostredia. Určitou skupinou ľudí nie je objavenie hodnôt ekológie pozitívne vnímané,
f) informačná bariéra – vzniká vtedy, keď má konzument málo vedomostí o výhodách a prednostiach bioproduktov a ich úžitku pre neho,
g) bariéra dôvery – deficit informácií o ekologických potravinách spôsobuje nedôveru konzumentov v tieto produkty.

Spotrebiteľské správanie je významným spôsobom ovplyvňované prostredím, v ktorom sa spotrebiteľ nachádza a ktoré na neho vplýva. Preto je nevyhnutným predpokladom realizovania úspešných retail marketingových aktivít spoznanie aj marketingového prostredia. Predpokladáme, že spotrebiteľ budúcnosti bude požadovať stále väčší „nadštandard“ v kvalite konzumovaných potravín, bude očakávať vyššiu a vyváženú biologickú hodnotu, pričom príťažlivá bude tak kvalita produktu, ako aj jeho forma predaja, prezentácie (obal, atraktívny vzhľad). Jednotlivec sa v budúcnosti bude viac riadiť heslom, že menej je niekedy viac a bude hľadať produkty biologicky cenné, ktorých konzumáciou aj v menšej miere zabezpečí príjem biologicky cenných látok, ktoré sú nevyhnuté pre zdravie a pohodu.

2. Bioprodukty a ich vnímanie spotrebiteľské verejnosťou

Za účelom zistenia informácií o znalosti bioproduktov spotrebiteľskou verejnosťou, ich dostupnosťou a celkovým vnímaním sme v roku 2011 zrealizovali výskum na vzorke 1302 respondentov Košického a Prešovského kraja. Zastúpenie z hľadiska pohlavia bolo nasledovné. Vzorku tvorilo 60 % žien a 40 % mužov. Pre potreby ďalšieho spracovania bolo použiteľných 1246 dotazníkov. Spracovanie bolo uskutočnené prostredníctvom štatistického softwaru SAS. Pri otázke koľkí z nich si už kúpili produkt s označením „bio“ bolo zistené, že 55 % respondentov udávalo kladnú odpoveď a 45 % zápornú. Väčšina má teda skúsenosti s nákupmi bioproduktov. Že nakupujú hlavne ženy a pravdepodobne sa vo zvýšenej miere venujú starostlivosti o zdravie svoje a svojej rodiny, môže byť dôležité pri sledovaní odpovedí na tú istú otázku z hľadiska pohlaví: 60% žien uviedlo že už niekedy bio výrobok zakúpilo, u mužov to bolo len 47 %.

Pri snahe zistiť vnímanie bioproduktov respondentmi sme použili v rámci piatich výrokov päťstupňovú bipolárnu škálu, ktorá má na krajných póloch protikladné výrazy. Respondenti hodnotili vnímanie cenovej hladiny bioproduktov, ich užitočnosť, vplyv na zdravie, posudzovali módnosť trendu týkajúceho sa konzumácie bioproduktov, ale rovnako aj dostatočnosť ich propagácie.

Otázka vnímania bioproduktov bola konštruovaná nasledovne: *Zakrúžkujte na škále políčko, ktoré vyjadruje Vaše vnímanie: „Bioprodukty sú“:*

	1	*2*	*3*	*4*	*5*	
• *lacné*	□	□	□	□	□	*drahé*
• *zbytočné*	□	□	□	□	□	*užitočné*
• *nemajú vplyv na zdravie*	□	□	□	□	□	*zdraviu prospešné*
• *nie sú módnou záležitosťou*	□	□	□	□	□	*trendové*
• *málo propagované*	□	□	□	□	□	*propagované vhodne*

Priemerné hodnoty odpovedí respondentov sú uvedené v tabuľke 1. Z prezentovaných hodnôt vyplýva, že vysoká priemerná hodnota vnímania ceny (4,25) bio produktov svedčí o tom, že tieto produkty sú považované za drahé. Vysoká hodnota priemeru 4,11 je znakom toho, že respondenti považujú dané produkty za zdraviu prospešné. Nasledujúca priemerná hodnota 3,82 súvisí s ich názorom, že bio produkty sú pre nás užitočné. Priemernou hodnotu 3,26 hodnotili respondenti názor na ich trendovosť a teda ich považujú za módnu záležitosť. Najnižšiu priemernú hodnotu sme zaznamenali pri hodnotení vnímania propagácie (2,56). Z tejto priemernej hodnoty je teda zrejmé, že respondenti považujú propagáciu bio produktov za málo dostačujúcu.

Tab. 1 Priemerné hodnoty vnímania dôležitosti bioproduktov

Variable	Mean
07A_Vnimanie_lacne	4.25
07B_Vnimanie_zbytocne	3.82
07C_Vnimanie_zdravie	4.11
07D_Vnimanie_moda	3.26
07E_Vnimanie_propagacia	2.56

Z postojov respondentov teda zaznieval najmä argument, že bioprodukty sú zdraviu prospešné, užitočné, ale sú drahé a trendové. Hodnotenie úrovne propagácie bolo vnímané ako málo dostatočné. V nasledujúcej tabuľke sme sledovali postoje respondentov podľa pohlavia (viď tab. 2) a už na prvý pohľad je zrejmé, že vnímanie dôležitosti niektorých faktorov je rozdielne u mužov a žien a preto sme sa rozhodli možnú existenciu rozdielov podrobiť ďalšej analýze.

Tab. 2 Priemerné hodnoty vnímania dôležitosti bioproduktov podľa pohlavia

17_Pohlavie	N Obs	Variable	Mean
ženy	740	07A_Vnimanie_lacne	4.30
		07B_Vnimanie_zbytocne	3.90
		07C_Vnimanie_zdravie	4.21
		07D_Vnimanie_moda	3.26
		07E_Vnimanie_propagacia	2.55

Tab. 2 - dokončenie

17_Pohlavie	N Obs	Variable	Mean
muži	501	07A_Vnimanie_lacne	4.18
		07B_Vnimanie_zbytocne	3.72
		07C_Vnimanie_zdravie	3.97
		07D_Vnimanie_moda	3.27
		07E_Vnimanie_propagacia	2.57

Analýzu zhody stredných hodnôt pri vnímaní dôležitosti bioproduktov podľa pohlavia sme uskutočnili pomocou neparametrickej analýzy rozptylu. Na základe procedúry *Nonparametric One-Way ANOVA* sme dospeli k nasledovným záverom. Na bežne používanej hladine významnosti 0,05 bol potvrdený štatisticky významný rozdiel vo vnímaní ceny bioproduktov medzi oboma pohlaviami (P-hodnota Kruskal-Wallis testu je P=0,0333). Respondenti mužského a ženského pohlavia odlišne vnímali aj hodnotenie toho, či bioprodukty sú alebo nie sú pre nich užitočné (P-hodnota Kruskal-Wallis testu bola P=0,0341). Najvýraznejšie rozdiely stredných hodnôt boli zaznamenané vo vnímaní vplyvu bioproduktov na zdravie. Kým v skupine mužov bol dosiahnutý priemerný výsledok hodnotenia tejto otázky na úrovni 3,97, ženy hodnotili vplyv bioproduktov na zdravie priemernou hodnotou 4,21 (P-hodnota Kruskal-Wallis testu je P=0,0001). Vnímanie bioproduktov ako módnej záležitosti bola v priemerne rovnako hodnotená oboma pohlaviami (P-hodnota Kruskal-Wallis testu je P=0,866). Podobne nebol dokázaný štatisticky významný rozdiel strednej hodnoty vnímanie propagácie bioproduktov (P=0,8008).

Rovnako nás zaujímal názor, či respondenti považujú tuzemské biopotraviny za porovnateľné v oblasti kvality so zahraničnými. Priemerná hodnota ich odpovedí na otázku, či sú tuzemské biopotraviny kvalitatívne porovnateľné so zahraničnými na 5 stupňovej škále (1 – určite áno, ..., 5 – určite nie), dosiahla hodnotu 2,70.

Priemerné hodnotenie respondentov na úrovni 3,45 na otázku, či je v obchodoch dostatočný výber slovenských biopotravín opäť na 5 stupňovej škále (1 – určite áno, ..., 5 – určite nie) poukazuje na to, že respondenti majú výhrady voči nedostatočnému zastúpeniu bioproduktov v predajnej sieti na Slovensku.

Najhoršie hodnotili respondenti na 5 stupňovej škále propagáciu slovenských bioproduktov. Priemerná hodnota 3,99 napovedá, že tieto produkty sú z pohľadu respondentov nedostatočne propagované.

3. Predaj z dvora v odraze spotrebiteľských preferencií

V rovnakom dotazníku sme zisťovali záujem týkajúci sa alternatívy preferencie nákupu z dvora u piatich tovarových komodít. Išlo o zeleninu, ovocie, mäso, mlieko a mliečne výrobky. Hodnotenie sme uskutočnili na základe priemerných hodnôt 5 bodovej škály (1 – určite áno, 2 – skôr áno, 3 – neviem, 4 – skôr nie, 5 – určite nie). Najviac preferovanou komoditou bolo ovocie a zelenina s celkovou priemernou hodnotou preferencií nákupu z dvora na úrovni 1,97. Nasledovalo mlieko s priemerom 2,50. Preferencia nákupu mliečnych výrobkov z dvora dosiahla 2,70 a najmenej atraktívnou sledovanou tovarovou komoditou pri nákupe z dvora bolo mäso, ktorého priemerná hodnota preferencií dosiahla len 2,89.

Tab. 3: Preferencia nákupu jednotlivých tovarových komodít pri nákupe z dvora

Variable	Mean
11A_PZD_zelenina	1.97
11B_PZD_ovocie	1.97
11C_PZD_maso	2.89
11D_PZD_mlieko	2.50
11E_PZD_mliecne_vyrobky	2.70

Následne sme sledovali, či existujú štatisticky významné rozdiely preferencií nákupu z dvora podľa pohlavia. Pri komodite zelenina bola priemerná hodnota preferencií nákupu z dvora u žien 1,88 a u mužov len 2,10. Štatisticky významný rozdiel na bežne používanej hladine významnosti 0,05 bol preukázaný pomocou neparametrickej analýzy rozptylu (P-hodnota Kruskal-Wallis testu P=0,0023). Podobne to bolo aj v prípade analýzy rozdielov stredných hodnôt preferencií nákupu ovocia z dvora podľa pohlavia. Aj v tomto prípade bol preukázaný štatisticky významný rozdiel v odpovediach podľa pohlavia (P-hodnota Kruskal-Wallis testu bola P=0,0027). Štatisticky významné rozdiely preferencií podľa pohlavia však neboli preukázané pri nasledovných tovarových komoditách: mäso (P-hodnota Kruskal-Wallis testu dosiahla hodnotu P=0,9870), mlieko (P=0,7621), mliečne výrobky (P=0,8738).

Medzi hlavné bariéry týkajúce sa nákupu z dvora patril argument vysokej ceny (42 %), nízkej miery informovanosti (25 %), ale aj zvykové správanie súvisiace s doterajšou spokojnosťou s tradičným nakupovaním v obchodoch (18 %).

Tab. 4 Priemerné hodnoty preferencií tovarových komodít pri nákupe z dvora podľa pohlavia

17_Pohlavie	N Obs	Variable	Mean
ženy	740	11A_PZD_zelenina	1.88
		11B_PZD_ovocie	1.89
		11C_PZD_maso	2.88
		11D_PZD_mlieko	2.50
		11E_PZD_mliecne_vyrobky	2.69
muži	501	11A_PZD_zelenina	2.10
		11B_PZD_ovocie	2.07
		11C_PZD_maso	2.89
		11D_PZD_mlieko	2.48
		11E_PZD_mliecne_vyrobky	2.69

4. Záver

Problematika týkajúca sa spotreby, ale aj dopytu po potravinách produkovaných šetrným spôsobom voči životnému prostrediu je aktuálna a stáva sa súčasťou akceptácie spoločensky zodpovedného konania nielen jednotlivých organizácií, ale aj individuálnych osôb. V spoločnosti sa čoraz výraznejšie vyčleňuje spotrebiteľská skupina, pre ktorú sa stáva kvalita výrazne preferenčným faktorom pri výbere a nákupe potravín. Na strane spotrebiteľov je však cítiť deficit žiaducej informačnej osvety vo všetkých rovinách, od ekologického poľnohospodárstva až po aktivity v poslednom článku distribučného reťazca, teda v rovine retail manažmentu. Pozitívne ovplyvňovanie dopytu po bioproduktoch je vo výraznej miere závislé od jednotlivých aktivít nástrojov komunikačného mixu v prieze komplexu

uzavretého potravinového reťazca. Správne pochopenie marketingu a jeho implementácia je jedným zo základných krokov smerujúcich k dosiahnutiu úspechu bioproduktov na trhu.

Literatúra

[1] BRAY, J.- JOHNS, N. – KILBURN, D. 2011. An Exploratory Study into the Factors Impeding Ethical Consumption. In: *Journal of Business Ethics*, roč. 98, 2011, č. 4, s. 597-608, ISSN 1573-0697.

[2] CHAJDIAK, J. – KRIŠKOVÁ, A. 2012. Usporiadanie otázok dotazníka Správanie podporujúce zravie podľa intenzity celkového hodnotenia zdravotníckych asistentov. In: *Forum Statisticum Slovacum* 2/2012. SŠDS Bratislava. 2012. ISSN 1336-7420.

[3] CHAJDIAK, J.: *Štatistika jednoducho*. Bratislava, STATIS 2010. ISBN 978-80-85659-60-3.

[4] LIESKOVSKA,V a kol.: *Zelený marketing*. Bratislava, EKONÓM 2010, ISBN 978-80-225-3047-7.

[5] LUHA, J. 2010. Metodologické zásady záznamu dát z rozličných oblastí medicíny a zásady ich kontroly. In: FORUM STATISTICUM SLOVACUM 1/2010. SŠDS Bratislava. 2010. ISSN 1336-7420.

[6] LUHA, J. 2009. Matematicko-štatistické aspekty spracovania dotazníkových výskumov. In: FORUM STATISTICUM SLOVACUM 3/2009. SŠDS Bratislava. 2009. ISSN 1336-7420.

[7] LUHA, J. 2006. Štatistické metódy analýzy kvalitatívnych znakov. In: *Forum Statisticum Slovacum* 2/2006. SŠDS Bratislava. 2006. ISSN 1336-7420.

[8] LÖSTER, T. – ŘEZANKOVÁ, H. – LANGHAMROVÁ, J. 2009. *Statistické metody a demografie*. 1. vydanie. VŠEM, Praha. s. 297. ISBN 978-80-86730-43-1.

[9] STANKOVIČOVÁ I. – VOJTKOVÁ, M. 2007. *Viacrozmerné štatistické metódy s aplikáciami*. IURA EDITION, Bratislava 2007, ISBN 978-80-8078-152-1.

Adresa autorov:

Vanda Lieskovská, prof. Ing., PhD.
EU Bratislava, Podnikovohospodárska fakulta v Košiciach
Tajovského 11, 040 00 Košice
Vanda.lieskovska@euke.sk

Silvia Megyesiová, Ing. PhD.
EU Bratislava, Podnikovohospodárska fakulta v Košiciach
Tajovského 11, 040 00 Košice
silvia.megyesiova@euke.sk

Katarína Petrovčíková, Ing., PhD.
EU Bratislava, Podnikovohospodárska fakulta v Košiciach
Tajovského 11, 040 00 Košice
Katarina.petrovcikova@euke.sk

Počty zahraničných študentov na českých vysokých školách
Numbers of foreign students at the Czech universities

Bohdan Linda, Jana Kubanová

Abstract: The number of prospective candidates for university education is declining in the Czech Republic due to a long-term demographic development. Existence of universities, from an economic perspective, still depends largely on the number of students that can study at the university. Financial situation of the school can be also improved by providing of education to the applicants from foreigner countries. This article deals with the mapping of the current situation in the university education in terms of the number of foreign students.

Abstrakt: V Českej republike v dôsledku demografického vývoja dlhodobo klesá počet perspektívnych uchádzačov o vysokoškolské vzdelanie. Existencia škôl z ekonomického hľadiska však stále závisí z väčšej časti na počte študentov, ktoré vysoká škola má. Jedna z možností, ako si školy môžu zlepšiť finančnú situáciu je poskytovanie štúdia uchádzačom zo zahraničia. Tento článok sa zaoberá mapovaním súčasnej situácie na vysokých školách z pohľadu počtu zahraničných študentov, ktorí na nich študujú.

Key words: number of foreign students, public universities, private universities

Kľúčové slová: počty študujúcich cudzích štátnych príslušníkov, verejné vysoké školy, súkromné vysoké školy

JEL classification: I20

Úvod

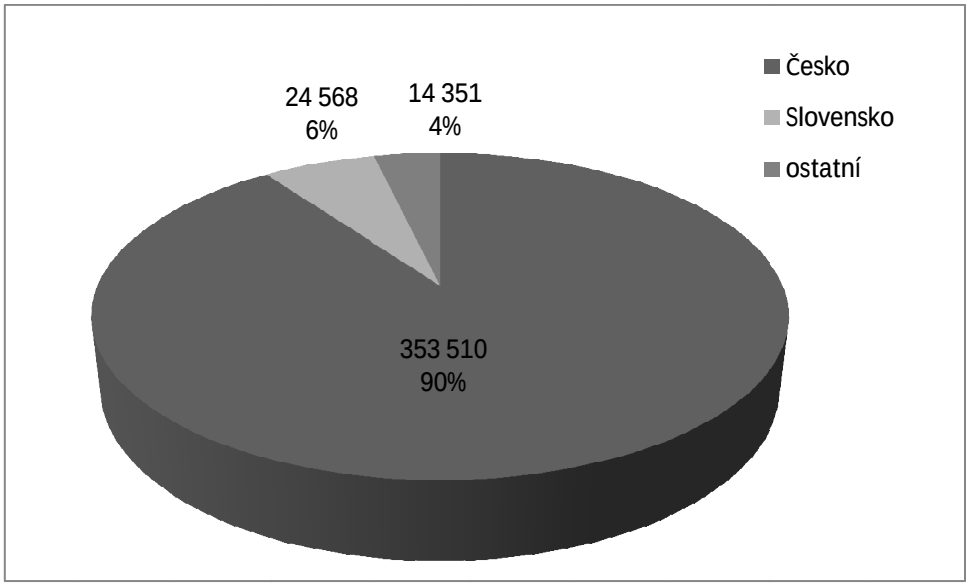
Je známým faktem, že v České republice od roku 1974 nastoupil trend klesající porodnosti s výjimkou krátkého období cca 2001 – 2008. Co se týče počtu míst na vysokých školách, je trend od roku 1991 právě opačný. V současné době v České republice je počet poskytovaných studijních míst na všech vysokých přibližně stejný, jako počet absolventů všech typů středních škol. To znamená, že už dnes každý absolvent střední školy i s tím nejhorším prospěchem má možnost studovat na některé vysoké škole. S probíhajícím procesem internacionalizace i vzhledem k úrovni některých absolventů středních škol přistupují stále častěji univerzity k tomu, že nabízejí studium uchazečům ze zahraničí.

V České republice mohou uchazeči ze zahraniční studovat vysokou školu dvojí formou. Pokud se student přihlásí do studijního programu v českém jazyce, studuje dle stejných pravidel jako český student. Hlavní výhodou je bezplatné studium. To vysvětluje, že největší počet zahraničních studentů je ze Slovenska (viz obr. 1), protože odpadá jazyková bariéra. Pokud student není schopen studovat v českém studijním programu – zásadní důvod je neznalost českého (slovenského) jazyka, může studovat ve studijním programu vyučovaném v anglickém jazyce. Pak by student měl být v roli samoplátce, pokud není toto studium ošetřeno bilaterální či vícestrannou smlouvou. Z ekonomického hlediska jsou studenti - samoplátci pro vysoké školy nejžádanější.

Současná situace

V současné době na českých vysokých školách studuje 392 249, z toho 38 919 studentů je zahraničních. Největší podíl zahraničních studentů pochází se Slovenska, celkem 24 568. Podobnost českého a slovenského jazyka je zde nepopíratelnou výhodou, tito studenti jsou především ve studijních programech vyučovaných v českém jazyce, pouze 19 slovenských

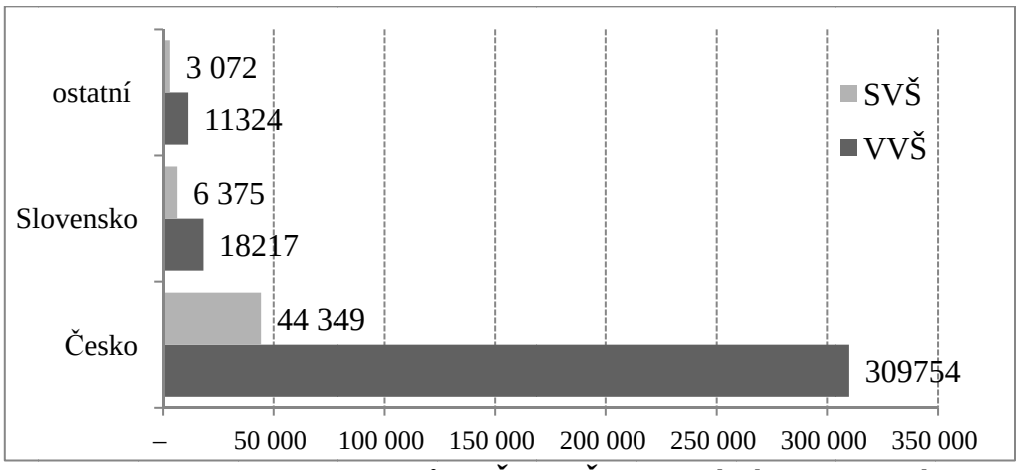
studentů je uváděno jako samoplátci. Na českých vysokých školách studují cizinci z více než 150 zemí, tvoří však jen 4% z celkového počtu (tj. 14 351 studentů).



,

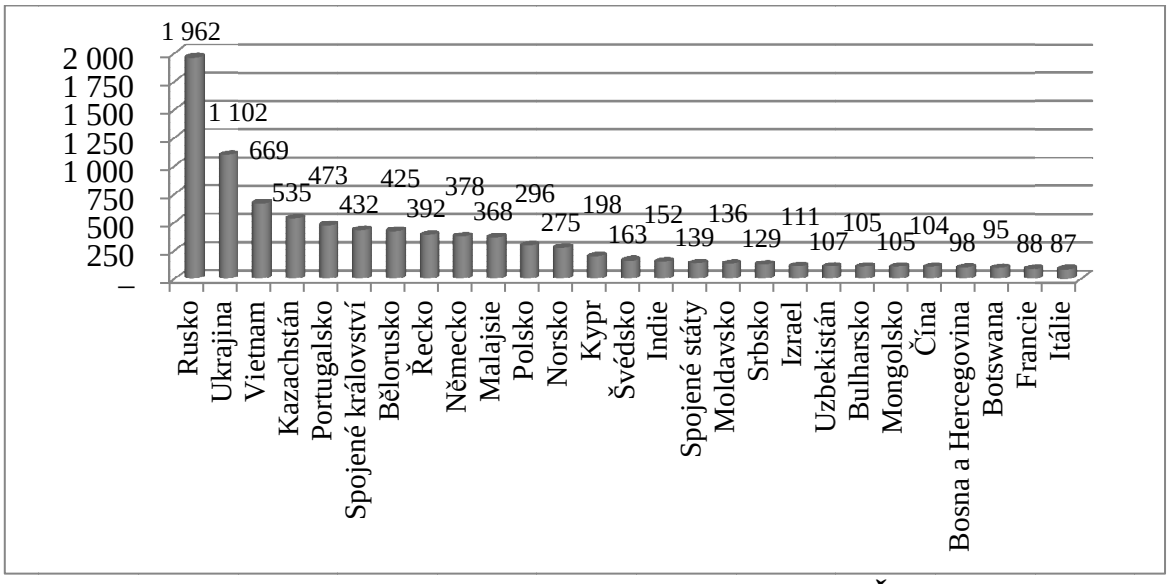
Obr. 1: Studenti VŠ podle státního občanství

Na obrázku 2 můžeme vidět počty studentů veřejných vysokých škol (dále VVŠ) a soukromých vysokých škol (dále SVŠ) z pohledu členění dle státního občanství.
Na SVŠ studuje 44 349 českých studentů, což je 13% ze všech studentů české národnosti, ze studentů slovenské národnosti studuje na SVŠ 6 375, což je 26%. U studentů ostatních národností je na SVŠ 3 072 studentů, tj. 21%.



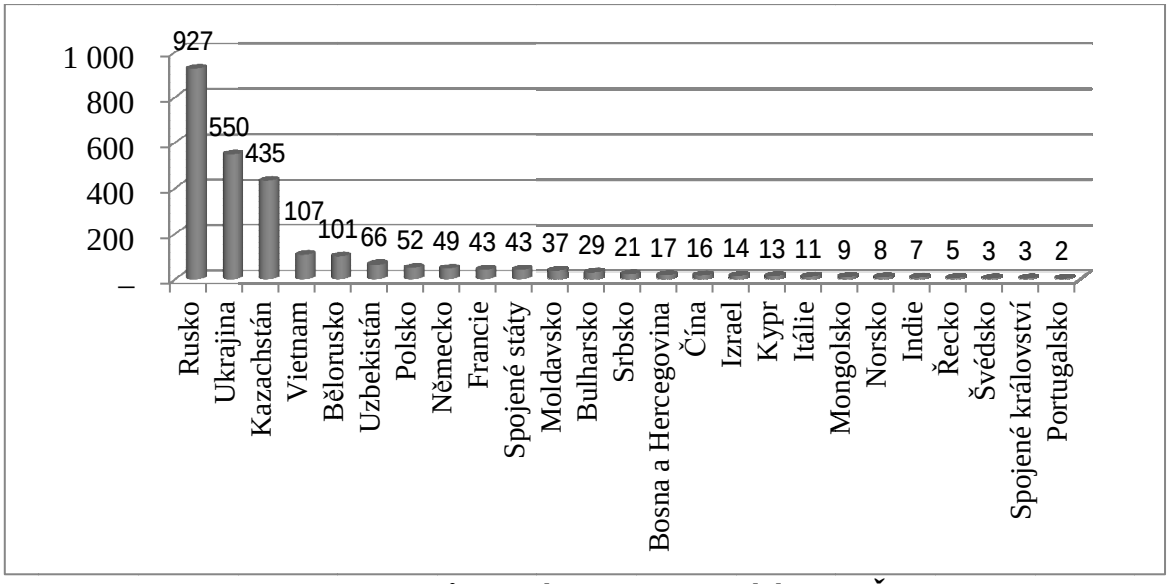
Obr. 2: Počty studentů VVŠ a SVŠ podle státního občanství

Počty cizinců dle národnosti studujících na VVŠ s výjimkou Slovenska jsou uvedeny na obr. 3. Je zřejmé, že nejvíce zahraničních studentů přichází na VVŠ z Ruska, 1962 studentů, dále z Ukrajiny (1102), Vietnamu (669), Kazachstánu (535). Za příčinu této skladby zahraničních studentů můžeme považovat především podobnost jazyka (u prvních dvou), ale i historické kořeny mezinárodní spolupráce (Vietnam).



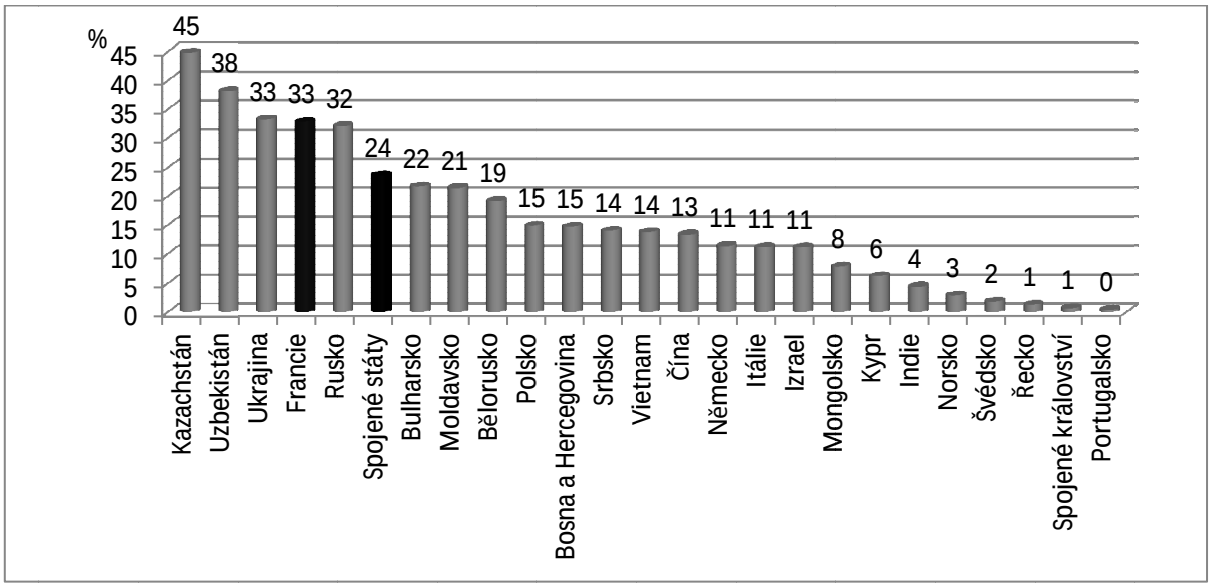
Obr. 3: Počty cizinců dle národnosti studující na VVŠ

Podobně, jako jsme analyzovali počty cizích státních příslušníků na VVŠ, jsme se zaměřili na stav na SVŠ. Ani zde nebyl zaznamenán jiný trend, opět je nejvíce studentů z Ruska (927) a Ukrajiny (550), dále pak z Kazachstánu (435) a Vietnamu (107) – viz obr. 4. Zde hrají jistě významnou úlohu i manažerské schopnosti jednotlivých SVŠ, neboť skupina cizinců jedné země směřuje často na jednu SVŠ.



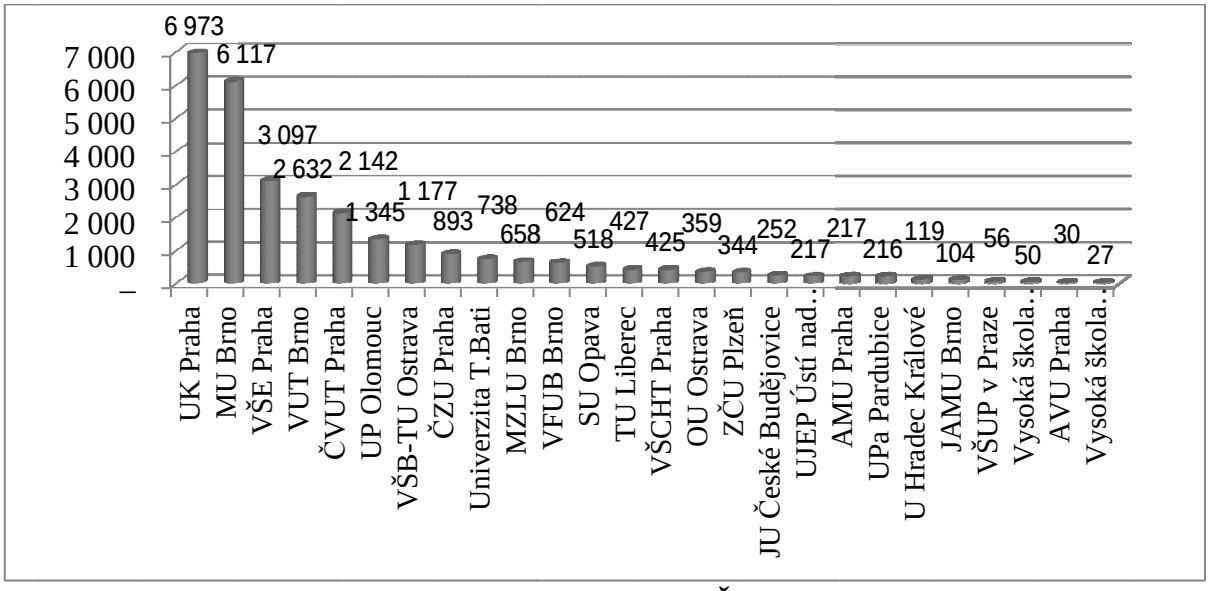
Obr. 3: Počty cizinců dle národnosti studující na SVŠ

Na obr. 4 jsou uvedena procenta cizinců dané národnosti studující v České republice na SVŠ. To znamená například, že 45% ze všech Kazachů studujících v Č je na SVŠ. Podobně je na SVŠ 38% Uzbeků, 33% Ukrajinců a stejné procento Francouzů, 32% Rusů apod. Jsou to již významné počty studentů, kteří studují na tomto typu škol. Často se koncentrují na jedné ze SVŠ, která určitý studijní program zajišťuje a která věnuje určité úsilí náboru zahraničních studentů. Jistě zaujme vysoký počet studentů z Francie (33) a USA (24).

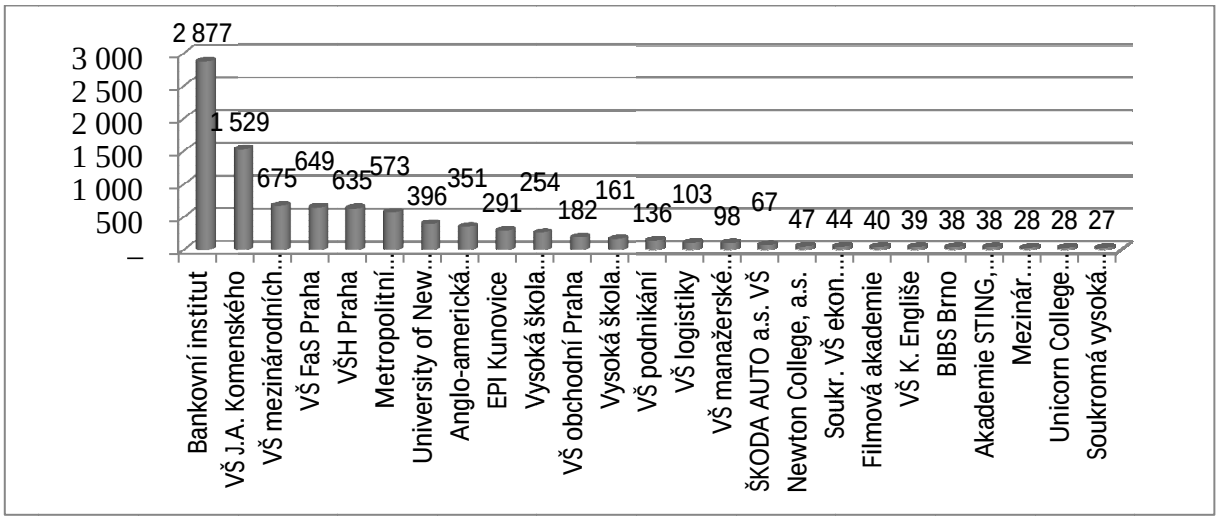


Obr. 4: Procento cizinců dle národnosti studující na SVŠ
(100% je počet cizinců dané národnosti studující na VŠ v ČR)

Na následujícím obrázku 5 jsou demonstrovány počty zahraničních studentů jednotlivých VVŠ. Nejvyšší počet je dle předpokladu na největší české univerzitě, na Univerzitě Karlově (6 973), z toho na 1. lékařské fakultě 1 417 studentů, na MFF 707 studentů a na FF 700 studentů. Na Masarykově univerzitě v Brně studuje 6 117 zahraničních studentů, z toho 1 438 na Lékařské fakultě, 1 263 na FF, 1014 na FI (Fakultě informatiky). Na VŠE Praha studuje 3 097 cizinců, nejvíce na Fakultě mezinárodních vztahů (956) a na Fakultě podnikohospodářské (754). Na VUT Brno je v celkového počtu 2 632 zahraničních studentů 533 na Fakultě informačních technologií a 551 na Fakultě elektrotechniky a informatiky.



Obr. 5: Počty cizinců dle VVŠ



Obr. 6: Počty cizinců dle SVŠ

Co se týká cizinců na SVŠ, nejvíce zahraničních studentů je na Bankovním institutu (2 877), dále na VŠ J. A. Komenského (1 529). Počty na dalších SVŠ jsou patrné z obr. 6.

Závěr

Disproporce mezi počtem nabízených míst a počtem absolventů středních škol vytváří konkurenční boj mezi fakultami, což se projevuje, bohužel negativně, především v kvalitě absolventů. Tato skutečnost poznamenává i studium cizinců na našich vysokých školách. Ve snaze udržet se nad vodou z finančního hlediska vytvářejí vysoké školy, především soukromé, různá detašovaná pracoviště buď v zahraničí (hlavně v krajinách bývalé RVHP), anebo na území České republiky ve spolupráci s nějakou zahraniční vysokou školou. Studenti pak zde získávají za ne zcela standardních situací vysokoškolské tituly. V nedávné minulosti v této souvislosti proběhly veřejnými médii zprávy o návrzích akreditační komise na odebrání akreditace některým takovýmto vysokým školám, které však byly bývalým ministrem školství Dobešem odmítnuty.

Tím, co bylo výše řečeno, nechtějí autoři vyvolat dojem, že by se nemělo podporovat studium zahraničních studentů na našich vysokých školách. Právě naopak. Avšak toto studium by měly nabízet především ty vysoké školy a fakulty, které dokáží zabezpečit vysokou úroveň studia a tak propagovat nejen samotné vysoké školství, ale skrze něj i Českou republiku v zahraničí. Studium zahraničních studentů by se nemělo zredukovat pouze na zdroj chybějících peněz ve školství bez ohledu na jeho kvalitu. Zde by měla sehrát důslednou kontrolní roli právě akreditační komise.

Literatura

[1] HTTP://DSIA.UIV.CZ/VYSTUPY/VU_VS.HTML

Adresa autora (-ov):

Bohdan Linda
Fakulta ekonomicko-správní
Univerzita Pardubice
Studentská 95, 53210 Pardubice
bohdan.linda@upce.cz

Jana Kubanová
Fakulta ekonomicko-správní
Univerzita Pardubice
Studentská 95, 53210 Pardubice
jana.kubanova@upce.cz

Srovnání podnikatelské a nepodnikatelské sféry v regionech ČR z hlediska trhu práce

Comparison of Business and Non-business sphere in regions of the Czech republic in the view of Labour market

Tomáš Löster, Jana Langhamrová

Abstract: The aim of this paper is to describe differences between Business and Non-business sphere (the subdivision MLSA) in regions of the Czech Republic in the view of Labour market. In terms of labour cost are analysed for individual regions the average gross wage, the median of gross wage, the average hourly wage, median hourly wages, average hourly wages of men and women. Furthermore, in both periods is observed proportion of workers with below average labour costs. In addition to comparisons in each region is observed among change between periods from 2008 to 2011. For example in terms of the non-business sphere was revealed that in all regions there was a decrease in the difference between the average hourly wages of men and women. The highest decrease in difference was recorded in the Central Region, where the difference was reduced to 9.76%.

Key words: Labour Market, Business sphere, Non-business sphere, average gross wage, median of gross wage.

Klíčová slova: Trh práce, podnikatelská sféra, nepodnikatelská sféra, průměrná hrubá měsíční mzda, medián hrubé měsíční mzdy.

JEL classification: C40, E24

1. Úvod

Problematice příjmových rozdělení, nerovnosti, chudoby, ale také nezaměstnanosti a jejich regionálním analýzám je věnována řada výzkumných prací a článků a to nejen v České republice, na Slovensku ale také v dalších zemích EU. Problematika nezaměstnanosti je také závažný ekonomický problém s řadou aspektů na celý ekonomický proces. Svědčí o tom řada prací, jako například [7], [8] či [9]. Z řad ekonomů je dále pak speciálně analyzována dlouhodobá nezaměstnanost vzhledem k jejím důsledkům, viz [10]. Modelováním a analýzou chudoby a příjmovým rozdělením se pak věnují další práce, viz např. [1], [3], [4] či [11].

2. Analýza charakteristik trhu práce u podnikatelské sféry

Mezi základní charakteristiky ceny práce, které jsou v rámci tohoto článku v jednotlivých krajích sledovány, patří průměrná hrubá mzda (v Kč), medián mzdy (v Kč), podíl zaměstnanců s podprůměrným hodinovým výdělkem, průměrná hodinová mzda (v Kč), medián hodinového výdělku (v Kč), průměrná hodinová mzda žen (v Kč) a průměrná hodinová mzda mužů (v Kč). (Poznámka: „*Hodinový výdělek se zjišťuje jako průměrný hodinový výdělek definovaný v § 351 až § 362 zákona č. 262/2006 Sb., zákoníku práce, ve znění pozdějších předpisů*“, viz MPSV).

V tabulce 1 jsou uvedeny rozdíly průměrné hodinové sazby mužů a žen (v Kč) mezi 4. čtvrtletím roku 2008 a 4. čtvrtletím roku 2011. Je zřejmé, že v Hlavním městě Praze došlo k nárůstu průměrného hodinového výdělku, a to o 25,76 Kč u žen a 44,51 Kč u mužů. Z tabulky 1 je také patrné, že ve většině krajů došlo k poklesu průměrného hodinového výdělku a to jak u mužů, tak u žen. K největšímu propadu průměrného hodinového výdělku u žen došlo v kraji Vysočina (o 18,91 Kč) a u mužů v Královéhradeckém kraji (o 20,23 Kč).

Tabulka 1: Změny průměrného hodinového výdělku mezi roky 2008 a 2011 (v Kč)

Kraj	Průměr. hod. ženy	Průměr. hod. muži
Jihočeský	12,80	9,31
Jihomoravský	2,77	-5,02
Karlovarský	-17,92	-19,89
Královéhradecký	-18,81	-20,23
Liberecký	-7,00	-0,90
Moravskoslezský	-9,02	0,33
Olomoucký	-14,31	-5,58
Pardubický	-18,88	-9,04
Plzeňský	-8,26	-1,92
Hl. město Praha	25,76	44,51
Středočeský	-1,77	10,00
Ústecký	-13,02	-1,76
Vysočina	-18,91	-6,39
Zlínský	-14,47	-10,04

ZDROJ: VLASTNÍ VÝPOČET

Tabulka 2: Srovnání ukazatelů (v Kč) v roce 2008

Kraj	Rozdíl pr. hod. v. M - Ž	Rozdíl hod. prům. - medián	Rozdíl PHM-MM
Jihočeský	36,72	17,29	2887
Jihomoravský	47,92	28,14	4830
Karlovarský	31,54	5,33	764
Královéhradecký	34,21	8,82	1129
Liberecký	22,55	6,69	993
Moravskoslezský	33,28	6,19	1116
Olomoucký	20,46	6,86	1068
Pardubický	21,37	5,35	733
Plzeňský	28,84	7,61	1280
Hl. město Praha	37,66	14,46	2371
Středočeský	30,16	6,69	671
Ústecký	26,31	7,03	994
Vysočina	30,83	6,06	913
Zlínský	33,73	5,31	558

ZDROJ: VLASTNÍ VÝPOČET

Zajímavé je i srovnání ukazatelů, jako je rozdíl průměrného hodinového výdělku mužů a žen (v Kč), rozdíl průměrného hodinového výdělku a mediánu hrubého výdělku, stejně tak rozdíl průměrného hrubého výdělku a mediánu hrubého výdělku. Hodnoty těchto ukazatelů jsou zachyceny v tabulce 2 (pro rok 2008) a v tabulce 3 (pro rok 2011). Z tabulek je zřejmé, že průměrný hodinový výdělek je ve všech krajích v obou letech vyšší, než medián hodinového výdělku. Znamená to tedy, že více než 50 % v podnikatelské sféře nedosahuje na průměrný hodinový výdělek.

Kromě analýzy nerovnoměrnosti rozdělení výdělků je zajímavé sledovat, k jaké změně v této nerovnoměrnosti došlo mezi 4. čtvrtletím roku 2008 a 4. čtvrtletím roku 2011. Tyto údaje jsou zachyceny v tabulce 4. Je zřejmé, že kromě několika krajů (Jihočeský, Jihomoravský, Karlovarský a Královéhradecký) docházelo k prohlubování rozdílů mezi průměrným hodinovým výdělkem mužů a žen. K nejvyššímu prohloubení došlo v Hlavním městě Praze, kde se rozdíl v průměrném hodinovém výdělku zvýšil o 49,79 %. Zároveň je z tabulky 4 zřejmé, že docházelo (kromě Jihomoravského kraje) ke zvyšování rozdílu mezi průměrným hodinovým výdělkem a mediánem hodinového výdělku, což znatelně prohlubovalo nerovnoměrnost rozdělení výdělků v podnikatelské sféře.

Tabulka 3: Srovnání ukazatelů (v Kč) v roce 2011

Kraj	Rozdíl pr. hod.v. M - Ž	Rozdíl hod. prům. - medián	Rozdíl PHM - MM
Jihočeský	33,23	18,25	3262
Jihomoravský	40,13	23,35	4100
Karlovarský	29,57	19,26	2925
Královéhradecký	32,79	17,11	2780
Liberecký	28,64	15,89	2938
Moravskoslezský	42,63	18,85	2778
Olomoucký	29,20	13,49	2256
Pardubický	31,22	16,85	2786
Plzeňský	35,17	17,73	2763
Hl. město Praha	56,41	51,24	7894
Středočeský	41,93	24,61	3590
Ústecký	37,58	18,46	3024
Vysočina	43,34	16,89	2965
Zlínský	38,16	16,15	2587

ZDROJ: VLASTNÍ VÝPOČET

Tabulka 4: Změny ukazatelů (v %) mezi roky 2008 a 2011

Kraj	Rozdíl pr. hod. v. M - Ž	Rozdíl hod. prům. - medián	Rozdíl PHM-MM
Jihočeský	-9,50	5,57	13,01
Jihomoravský	-16,25	-17,02	-15,13
Karlovarský	-6,25	261,33	283,09
Královéhradecký	-4,17	94,06	146,26
Liberecký	27,04	137,44	195,78
Moravskoslezský	28,11	204,42	148,83
Olomoucký	42,70	96,77	111,21
Pardubický	46,08	215,12	280,16
Plzeňský	21,97	132,90	115,84
Hl. město Praha	49,79	254,28	232,89
Středočeský	39,05	267,92	434,65
Ústecký	42,81	162,50	204,20
Vysočina	40,61	178,59	224,84
Zlínský	13,14	204,08	364,06

ZDROJ: VLASTNÍ VÝPOČET

3. Analýza charakteristik trhu práce u nepodnikatelské sféry

V tabulce 5 jsou uvedeny rozdíly průměrné hodinové sazby mužů a žen (v Kč) mezi 4. čtvrtletím roku 2008 a 4. čtvrtletím roku 2011 u nepodnikatelské sféry. Je zřejmé, že ve všech krajích došlo k nárůstu průměrného hodinového výdělku ve všech sledovaných krajích a to u obou pohlaví. K největšímu nárůstu průměrného hodinového výdělku došlo v Plzeňském kraji a to o 22,45 Kč u žen a o 26,06 Kč u mužů. Z hlediska srovnání nárůstů průměrných hodinových výdělku jsou zajímavé dvě skutečnosti. Jednak již zmíněný fakt, že ve všech krajích došlo oproti podnikatelské sféře k nárůstu průměrných hodinových výdělku a jednak, že nárůst průměrného hodinového výdělku (kromě Hlavního města Prahy) je vyšší u žen, než u mužů.

Zajímavé je i srovnání ukazatelů, jako je rozdíl průměrného hodinového výdělku mužů a žen (v Kč), rozdíl průměrného hodinového výdělku a mediánu hrubého výdělku, stejně tak rozdíl průměrného hrubého výdělku a mediánu hrubého výdělku. Hodnoty těchto ukazatelů jsou zachyceny v tabulce 6 (pro rok 2008) a v tabulce 7 (pro rok 2011). Z tabulek je zřejmé, že průměrný hodinový výdělek je ve všech krajích v obou letech vyšší, než medián hodinového výdělku. Znamená to tedy, že více než 50 % v nepodnikatelské sféře nedosahuje na průměrný hodinový výdělek.

Tabulka 5: Změny průměrného hodinového výdělku mezi roky 2008 a 2011 (v Kč)

Kraj	Průměr. hod. ženy	Průměr. hod. muži
Jihočeský	12,46	4,80
Jihomoravský	16,86	8,56
Karlovarský	11,11	6,90
Královéhradecký	12,17	3,74
Liberecký	15,17	12,58
Moravskoslezský	16,72	13,89
Olomoucký	13,45	6,00
Pardubický	13,87	9,64
Plzeňský	22,45	26,06
Hl. město Praha	14,68	9,94
Středočeský	13,65	3,14
Ústecký	12,19	4,63
Vysočina	10,08	8,27
Zlínský	8,79	3,17

ZDROJ: VLASTNÍ VÝPOČET

Tabulka 6: Srovnání ukazatelů (v Kč) v roce 2008

Kraj	Rozdíl pr. hod. v. M - Ž	Rozdíl hod. prům. - medián	Rozdíl PHM-MM
Jihočeský	26,10	5,75	835
Jihomoravský	29,86	7,33	1202
Karlovarský	31,54	5,33	764
Královéhradecký	34,21	8,82	1129
Liberecký	22,55	6,69	993
Moravskoslezský	33,28	6,19	1116
Olomoucký	20,46	6,86	1068
Pardubický	21,37	5,35	733
Plzeňský	28,84	7,61	1280
Hl. město Praha	37,66	14,46	2371
Středočeský	30,16	6,69	671
Ústecký	26,31	7,03	994
Vysočina	30,83	6,06	913
Zlínský	33,73	5,31	558

ZDROJ: VLASTNÍ VÝPOČET

Kromě analýzy nerovnoměrnosti rozdělení výdělků je zajímavé sledovat, k jaké změně v této nerovnoměrnosti u nepodnikatelské sféry došlo mezi 4. čtvrtletím roku 2008 a 4. čtvrtletím roku 2011. Tyto údaje jsou zachyceny v tabulce 8. Z tabulky je zřejmé, že (na rozdíl od podnikatelské sféry) došlo ve všech krajích ke snižování rozdílu v průměrném hodinovém výdělku mužů a žen. K nejzásadnější změně došlo v Olomouckém kraji, kde se rozdíl mezi průměrným hodinovým výdělkem mužů a žen snížil o 36,40 % a v kraji Středočeském, kde se rozdíl snížil o 34,84 %, což z hlediska pohlaví vedlo ke zlepšení v rovnosti mezi pohlavím z hlediska výdělku. Zároveň je z tabulky 8 zřejmé, že docházelo u většiny krajů ke zvyšování rozdílu mezi průměrným hodinovým výdělkem a mediánem hodinového výdělku, což mnohdy znatelně prohlubovalo nerovnoměrnost rozdělení výdělků v nepodnikatelské sféře. Například v Plzeňském kraji se rozdíl mezi hodinovým průměrem a mediánem hodinového výdělku zvýšil o 88,02 %. Graficky jsou změny rozdílů v průměrném hodinovém výdělku mužů a žen v jednotlivých krajích v nepodnikatelské sféře (a pro srovnání i v podnikatelské sféře) zachyceny na obrázku 1. Jak již bylo uvedeno výše, rozdíly mezi průměrným hodinovým výdělkem u mužů a žen se snižovaly ve všech krajích u nepodnikatelské sféry, na rozdíl od podnikatelské sféry, kde kromě Jihočeského a Jihomoravského kraje docházelo k prohlubování nerovnosti.

Tabulka 7: Srovnání ukazatelů (v Kč) v roce 2011

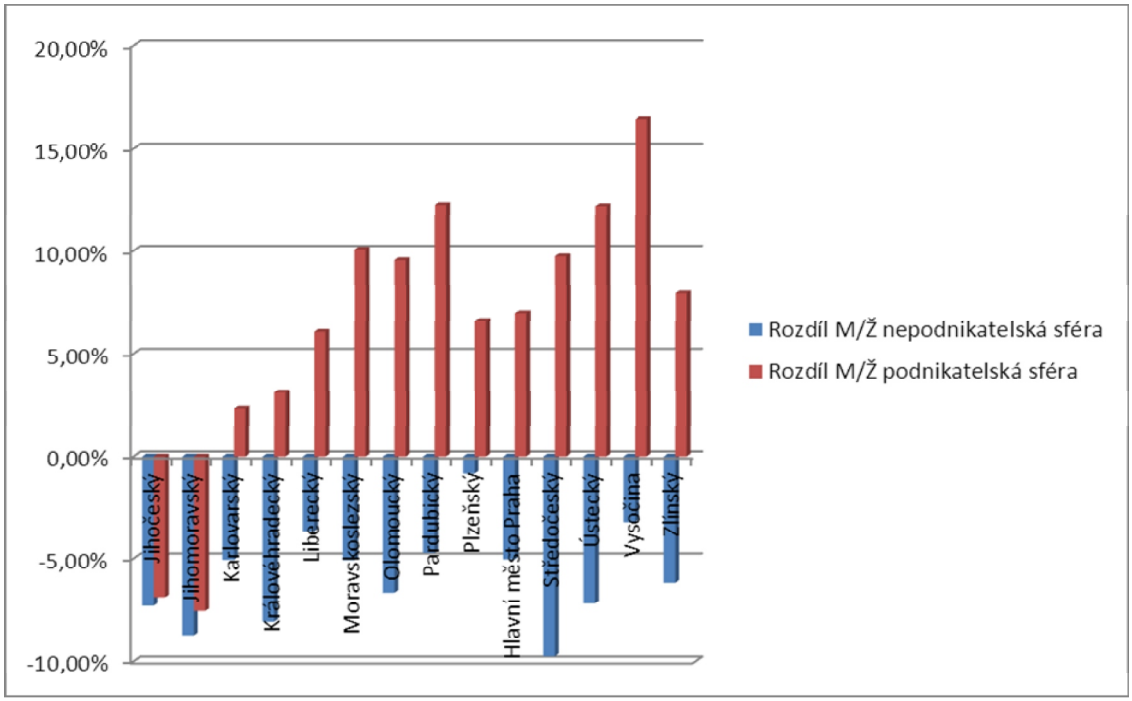
Kraj	Rozdíl pr. hod. v. M - Ž	Rozdíl hod. prům. - medián	Rozdíl PHM-MM
Jihočeský	18,44	5,39	611
Jihomoravský	21,56	9,04	1318
Karlovarský	27,33	6,43	730
Královéhradecký	25,78	8,26	1335
Liberecký	19,95	8,38	867
Moravskoslezský	30,45	7,75	1241
Olomoucký	13,01	8,47	1338
Pardubický	17,14	7,08	611
Plzeňský	32,45	14,31	1625
Hl. město Praha	32,92	14,33	2485
Středočeský	19,65	5,83	645
Ústecký	18,75	7,43	791
Vysočina	29,02	8,20	1063
Zlínský	28,11	3,77	454

ZDROJ: VLASTNÍ VÝPOČET

Tabulka 8: Změny ukazatelů (v %) mezi roky 2008 a 2011

Kraj	Rozdíl pr. hod. v. M - Ž	Rozdíl hod. prům. - medián	Rozdíl PHM-MM
Jihočeský	-29,37	-6,14	-26,86
Jihomoravský	-27,81	23,32	9,58
Karlovarský	-13,35	20,56	-4,40
Královéhradecký	-24,65	-6,28	18,24
Liberecký	-11,50	25,23	-12,68
Moravskoslezský	-8,49	25,23	11,18
Olomoucký	-36,40	23,48	25,27
Pardubický	-19,80	32,52	-16,68
Plzeňský	12,53	88,02	26,92
Hl. město Praha	-12,58	-0,92	4,78
Středočeský	-34,84	-12,86	-4,01
Ústecký	-28,73	5,59	-20,48
Vysočina	-5,86	35,23	16,45
Zlínský	-16,67	-29,08	-18,51

ZDROJ: VLASTNÍ VÝPOČET



Obrázek 3: Srovnání změn ukazatelů u podnikatelské a nepodnikatelské sféry

ZDROJ: VLASTNÍ VÝPOČET

4. Závěr

Výše výdělku, stejně tak míra nezaměstnanosti v jednotlivých regionech ovlivňují celý ekonomický proces a život obyvatelstva a proto je analýza těchto ukazatelů velmi důležitá pro získání komplexní představy o jednotlivých trzích práce. Při analýzách jednotlivých krajů u podnikatelské sféry bylo například zjištěno, že nejvyšší nárůst rozdílu mezi průměrným hrubým výdělkem a mediánem hrubého výdělku mezi rokem 2008 a 2011 v Hlavním městě Praze a to o 434 %, což prohloubilo nerovnost v rámci rozdělení výdělků. Naopak opačná situace byla identifikována v Jihomoravském kraji, kde došlo ke snížení rozdílu mezi průměrným hrubým výdělkem a mediánem hrubého výdělku a to o 15,13 %, čímž došlo ke snížení nerovnoměrnosti v rámci rozdělení. Z hlediska pohlaví u podnikatelské sféry bylo zjištěno, že ve všech krajích v obou sledovaných obdobích byl průměrný hodinový výdělek mužů vždy větší (minimálně o 17,01 % v Pardubickém kraji, maximálně o 41,59 % v Jihomoravském kraji). Během sledovaného období se kromě Jihočeského a Jihomoravského kraje rozdíly mezi průměrným hodinovým výdělkem mužů a žen ještě více prohloubily a to na maximální hodnotu 41,27 % v kraji Vysočina.

Při analýzách jednotlivých krajů u nepodnikatelské sféry bylo například zjištěno, že průměrný hodinový výdělek je v roce 2008 u mužů ve všech krajích vyšší než u žen. Nejvyšší průměrný hodinový výdělek byl, jak u mužů, tak u žen v Praze, kde dosáhl hodnoty 189,27 Kč (ve srovnání s podnikatelskou sférou 233,77 Kč) u mužů a 151, 61 Kč (ve srovnání s podnikatelskou sférou 177,37 Kč) u žen. Co se týká podílu zaměstnanců s podprůměrným hodinovým výdělkem, ve všech krajích je tato hodnota vyšší než 50 %, což znamená, že více než 50 % zaměstnanců nedosahuje na průměrný výdělek. Další analýzou bylo zjištěno, že ve všech krajích došlo k nárůstu průměrného hodinového výdělku a to u obou pohlaví. K největšímu nárůstu průměrného hodinového výdělku došlo v Plzeňském kraji a to o 22,45 Kč u žen a o 26,06 Kč u mužů. Z hlediska srovnání nárůstů průměrných hodinových výdělku je zajímavé, že oproti podnikatelské sféře došlo k nárůstu průměrných hodinových výdělku ve všech krajích, a že nárůst průměrného hodinového výdělku (kromě Hlavního města Prahy) je vyšší u žen, než u mužů. Zajímavá je i analýza vývoje rozdílů mezi průměrným hodinovým výdělkem mužů a žen. K nejzásadnější změně došlo v Olomouckém kraji, kde se rozdíl mezi průměrným hodinovým výdělkem mužů a žen snížil o 36,40 % a v kraji Středočeském, kde se rozdíl snížil o 34,84 %, což z hlediska pohlaví vedlo ke zlepšení v rovnosti mezi pohlavím z hlediska výdělku. Jak již bylo uvedeno výše, rozdíly mezi průměrným hodinovým výdělkem u mužů a žen se snižovaly ve všech krajích u nepodnikatelské sféry, na rozdíl od podnikatelské sféry, kde kromě Jihočeského a Jihomoravského kraje docházelo k prohlubování nerovnosti.

PODĚKOVÁNÍ:

Článek vznikl za podpory projektu Interní grantové agentury VŠE v Praze č. 19/2012 pod názvem Flexibilita trhu práce České republiky (IG 307042).

5. Literatura

[1] BARTOŠOVÁ, JITKA, FORBELSKÁ, MARIE. Comparison of Regional Monetary Poverty in the Czech and Slovak Republics. *Conference on Social Capital, Human Capital and Poverty in the Regions of Slovakia*. Herlany, Slovakia, October 13, 2010. ISBN 978-80-553-0573-8, pp. 76–84.

[2] BARTOŠOVÁ, JITKA. Analysis and Modelling of Financial Power of Czech Households. Bratislava 03.02.2009 – 06.02.2009. In: *8th International Conference APLIMAT 2009*.

Bratislava : Slovak University of Technology, 2009. ISBN 978-80-89313-31-0. pp. 717-722.

[3] BÍLKOVÁ, DIANA . Pareto Distribution and Wage Models. Bratislava 03.02.2009 – 06.02.2009. In: *Aplimat 2009* [CD-ROM]. Bratislava : Slovak university of technology, 2009, s. 723–732. ISBN 978-80-89313-31-0.

[4] BÍLKOVÁ, DIANA . Recent Development of the Wage and Income Distribution in the Czech Republic. *Prague Economic Papers* , 2/2012, (roč. 21, č. 2), s. 233–250. ISSN 1210-0455

[5] ČADIL, JAN, PAVELKA, TOMÁŠ, KAŇKOVÁ, EVA, VORLÍČEK, JAN. Odhad nákladů nezaměstnanosti z pohledu veřejných rozpočtů. *Politická ekonomie*, 2011, roč. 59, č. 5, s. 618–637. ISSN 0032-3233.

[6] MEGYESIOVÁ, S. – HUDÁK, M. Regionálne rozdiely mier nezamestnanosti a miezd na Slovensku a v Českej republike. In *Forum Statisticum Slovacum*. ISSN 1336-7420. Roč. 6, č.5 (2010), s. 155-160.

[7] MEGYESIOVÁ, SILVIA. Nezamestnanosť na Slovensku a v okolitých krajinách. In *Acta oeconomica Cassoviensia* No 3. Košice : Podnikovohospodárksa fakulta EU so sídlom v Košiciach, 1999. ISBN 80-88964-15-6. s. 303-308.

[8] MISKOLCZI, MARTINA, LANGHAMROVÁ, JITKA, FIALA, TOMÁŠ. Unemployment and GDP. Prague 22.09.2011 – 23.09.2011. In: *International Days of Statistics and Economics at VŠE, Prague [CD-ROM]*. Prague : VŠE, 2011, s. 1–9. ISBN 978-80-86175-72-0.

[9] MISKOLCZI, MARTINA, LANGHAMROVÁ, JITKA, LANGHAMROVÁ, JANA. Recognition of Differentiation in Unemployment Trends among Regions in the Czech Republic. Jindřichův Hradec 07.09.2011 – 09.09.2011. In: *IDIMT-2011*. Linz : Trauner Verlag universitat, 2011, s. 387–388. ISBN 978-3-85499-873-0.

[10] PAVELKA, TOMÁŠ. Long-term unemployment in the Czech republic. Praha 22.09.2011 – 23.09.2011. In: PAVELKA, Tomáš (ed.). *International Days of Statistic and Economics at VŠE [CD-ROM]*. Slaný : Melandrium, 2011. 9 s. ISBN 978-80-86175-72-0.

[11] ŽELINSKÝ, TOMÁŠ. Regions of Slovakia from the View of Poverty. In: *Conference on Social Capital, Human Capital and Poverty in the Regions of Slovakia.* Herlany, Slovakia, October 13, 2010. ISBN 978-80-553-0573-8, pp. 37-50.

Adresa autorů

Tomáš Löster, Ing., Ph. D.
Jana Langhamrová, Bc.
Katedra statistiky a pravděpodobnosti
Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
Nám. W. Churchilla 4, 130 67 Praha 3
Česká republika
Tel.: +420 2 24095 484
E-mail: tomas.loster@vse.cz, xlanj18@vse.cz

Názory verejnosti na migrantov a ich integráciu v SR: IV. čo by Vám prekážalo, keby?

Public opinion on migrants and their integration in SR: IV. what would hinder You, if?

Ján Luha, Lenka Berová, Martina Žáková

Abstract: We present results from public opinion research on migrants and their integration in Slovak republic, fourth part – public opinion on what would hinder You, if?

Abstrakt: V príspevku prezentujeme výsledky výskumu verejnej mienky názorov migrantov na ich integráciu v Slovenskej republike, štvrtá časť – názory verejnosti na to čo by Vám prekážalo, keby?

Key words: public opinion, immigrants, integration in SR, hinders You.

Kľúčové slová: názory verejnosti, imigranti, integrácia v SR, prekáža Vám.

JEL Classification: C1, C12.

1. Úvod

V príspevku prezentujeme štvrtú časť výsledkov vlastného výskumu názorov dospelej populácie Slovenskej republiky na aktuálne otázky a problémy spojené s problematikou migrácie, ktorý bol realizovaný v rámci vypracovania PhD. dizertačnej práce Berová L. (2012). Dotazník obsahuje niekoľko oblastí, ktoré špeciálne prezentujeme v príspevkoch - prvá časť výsledkov bola uverejnená v práci Berová L., Luha J., Žáková M. (2012a) - I. postoje k imigrantom prichádzajúcim do SR , druhá v práci Luha J., Berová L., Žáková M. (2012) II. začleňovanie imigrantov do spoločnosti - a tretia v práci Berová L., Luha J., Žáková M. (2012b) - III. čím nás môžu imigranti obohatiť.

Základná chrakterizácia výskumu je samozrejme rovnaká: Terénna fáza celoslovenského reprezentatívneho výskumu bola realizovaná v období od polovice novembra 2011 do konca januára 2012 poučenými dobrovoľnými anketármi.

Základný súbor tvorilo 4 405 673 dospelých obyvateľov SR, t.j. 81,06% z 5 435 273 všetkých obyvateľov SR k 31.12.2010, podľa údajov Štatistického úradu SR (Vekové zloženie obyvateľstva SR v roku 2010. Demografická a sociálna štatistika. ŠÚ SR Bratislava).

Výberový súbor o rozsahu 1120 respondentov bol reprezentatívny podľa kontrolovaných znakov pohlavie, vek, kraj a aj podľa nekontrolovaného znaku vzdelanie.

V tomto príspevku prezentujeme názory dospelej populácie SR na ďalšiu časť otázok o názoroch na cudzincov/migrantov, ktorí prichádzajú na Slovensko – aká situácia by respondentom prekážala. Zisťovali sme názory respondentov na 6 situácií, ktoré boli kladené respondentom ako podotázka. S touto tematikou súvisí aj otázka:

Prikláňali by ste sa k tomu, aby takto postupovala aj Slovenská republika - aby zaviedla kvóty pre migrantov?

Výsledky vo frekvenčnej tabuľke ukazujú na prevahu súhlasu so zavedením kvót pre imigrantov – spolu to bolo až 70,8% respondentov a tých čo by neboli za zavedenie kvót pre imigrantov bolo 29,2%. Na otázku neodpovedalo iba 1% respondentov.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 áno	411	36,7	37,1	37,1
	2 skôr áno	374	33,4	33,7	70,8
	3 skôr nie	148	13,2	13,3	84,1
	4 nie	74	6,6	6,7	90,8
	5 neviem	102	9,1	9,2	100,0
	Total	1109	99,0	100,0	
Missing	System	11	1,0		
Total		1120	100,0		

Téme bolo venovaných 6 podotázok:
Základná otázka bola: **Prekážalo by Vám, keby nastala nasledovná situácia?**
Podotázky zneli:

- cudzinci žili vo Vašej obci,
- cudzinci žili vo vašom susedstve,
- cudzinci boli vašimi blízkymi spolupracovníkmi,
- cudzinci sa stali súčasťou Vašej rodiny,
- ste pri transfúzii dostali krv cudzinca,
- bol v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov.

Škála odpovedí bola: 1=áno, prekážalo; 2=nie, neprekážalo. Kvôli jednoduchšie prezentovateľným výsledkom sme rekódovali odpovede na batériu otázok áno=1 a nie=0. Podiel odpovedí neviem bol pri uvedených podotázkach malý, takže má zmysel prezentovať ich podiel pomocou priemeru, resp. v percentách.

Výsledky za celý súbor dospelej populácie za skúmané otázky prezentujeme v percentách odpovede áno v jednoduchej tabuľke.

Prekážalo by Vám, keby?: situácia:

cudzinci žili vo Vašej obci	cudzinci žili vo vašom susedstve	cudzinci boli vašimi blízkymi spolupracovníkmi	cudzinci sa stali súčasťou Vašej rodiny	ste pri transfúzii dostali krv cudzinca	bol v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
17,24%	30,34%	16,13%	45,25%	43,20%	71,53%

Najvyšší podiel kladných odpovedí – až 71,53% bolo zistených na podotázku situáciu „bol v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov“. Nasledujú dve podotázky s „blizkym“ podielom kladných odpovedí 45,25% „cudzinci sa stali súčasťou Vašej rodiny“ a 43,20% „ste pri transfúzii dostali krv cudzinca“. Pomerne veľký podiel 30,34% je aj tých respondentov, ktorým by prekážalo keby „cudzinci žili vo vašom susedstve“. ***Keď je situácia „vzdialená“ od respondenta***, tak je podiel tých, ktorým by to prekážalo nižší, čo dokumentujú dve podotázky: 17,24% „cudzinci žili vo Vašej obci“ a 16,13% „cudzinci boli vašimi blízkymi spolupracovníkmi“. Profily podľa demografických znakov, ktoré skúmame v druhej kapitole vykazujú podobný „priebeh“ ako je hore uvedený profil za celý skúmaný súbor dospelej populácie.

2. Špecifiká názorov na imigrantov z pohľadu čo by im prekážalo, keby? - podľa demografických znakov

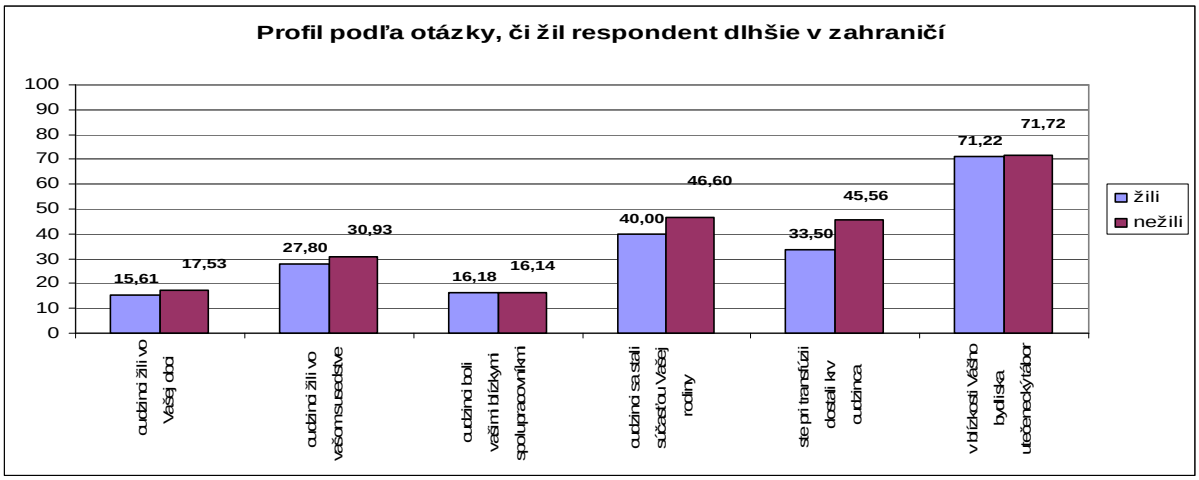
Ako sme už skôr uviedli, rekódovaním pôvodnej škály môžeme výsledky prezentovať podielom kladných odpovedí. Diferenciáciu názorov respondentov podľa demografických znakov prezentujeme v grafoch profilov za 6 skúmaných podotázok podľa demografických znakov, doplnených o dve otázky „Žili ste niekedy viac ako tri mesiace v zahraničí?“ a „Poznáte vo svojom blízkom okolí migranta z iného štátu, ktorý sa rozhodol žiť v SR?“. Prezentácia pomocou profilov je názorná a zrozumiteľná ako vyplýva z nasledovných 8 grafov príslušných profilov. Overovanie signifikantnosti rozdielov vo výsledkoch podľa demografických znakov sme robili pomocou Chí-kvadrát testu kontingenčnej tabuľky. Rozdiely považujeme za signifikantne odlišné keď P-hodnota Chí-kvadrát testu je menšia ako 0,05.

Na otázku ***„Žili ste niekedy viac ako tri mesiace v zahraničí***?“ neodpovedalo iba 7 respondentov, čo je 0,6%. Z 1113 platných odpovedí sme zistili, že v zahraničí viac ako 3 mesiace žilo 18,4% a nežilo 81,6% respondentov. Zaujímalo nás, ako takáto skúsenosť ovplyvňuje ich postoje ku imigrantom.

Výsledky prezentujeme v grafe 1. Vo všetkých skúmaných otázkach sme zaznamenali nižší podiel kladných odpovedí u respondentov, ktorí dlhšie žili v zahraničí ako u respondentov, ktorí takú skúsenosť nemajú. P-hodnoty Chí-kvadrát testu v tabuľke ukazujú signifikantnú menlivosť pri podotázke „Prekážalo by Vám, keby ‚situácia-ste pri transfúzii dostali krv cudzinca“ P=0,002, keď z respondentov čo žili dlhšie v zahraničí 33,50% respondentov uviedlo kladnú odpoveď a viac 45,56% kladných odpovedí uviedli respondenti čo nežili dlhšie v zahraničí. Na hranici signifikantnosti P=0,087 sú diferencované odpovede pri otázke „Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny“. Pri ostatných skúmaných podotázkach je diferencácia štatisticky nesignifikantná.

Žili ste niekedy viac ako tri mesiace v zahraničí?	Prekážalo by Vám, keby,situácia-cudzinci žili vo Vašej obci	Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve	Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi	Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny	Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca	Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
P-hodnoty	0,511	0,380	0,990	0,087	0,002	0,886

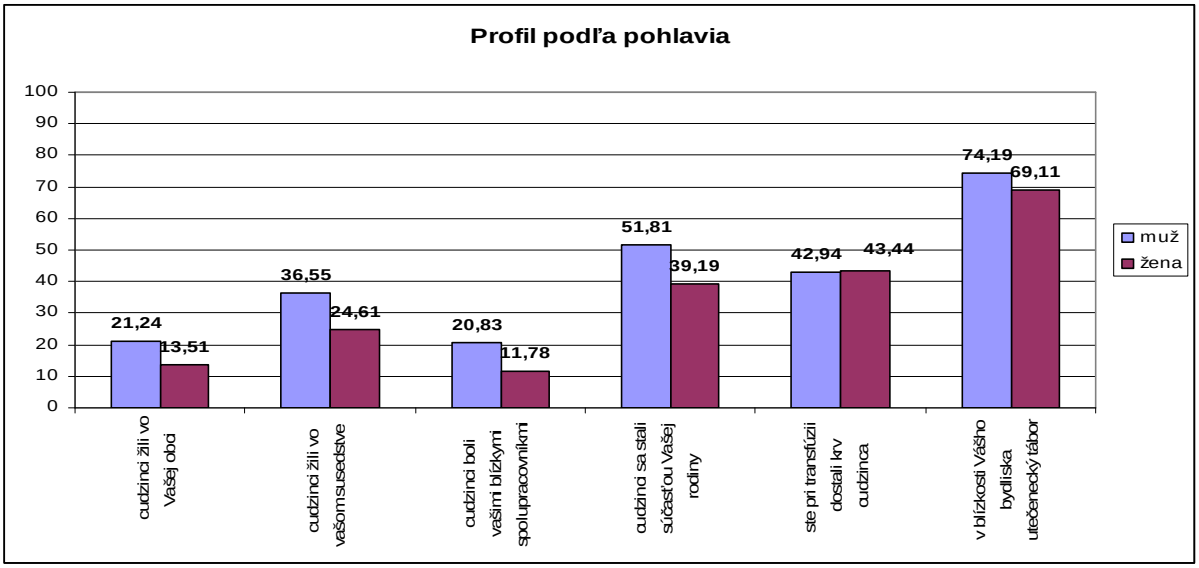
Graf 1. Profil kladných odpovedí na skúmané otázky podľa otázky „Žili ste niekedy viac ako tri mesiace v zahraničí?“



Profil odpovedí na skúmané podotázky ***podľa pohlavia*** prezentujeme v grafe 2. P-hodnoty Chí-kvadrát testu uvedené v tabuľke ukazujú signifikantnú menlivosť podľa pohlavia u všetkých podotázok s výnimkou podotázky „Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca“ , kde je P=0,868 a pri podotázke „Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov“ je P-hodnota P=0,063, čo je na hranici štatistickej signifikantnosti.

pohlavie	Prekážalo by Vám, keby,situácia-cudzinci žili vo Vašej obci	Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve	Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi	Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny	Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca	Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
P-hodnoty	0,001	0,000	0,000	0,000	0,868	0,063

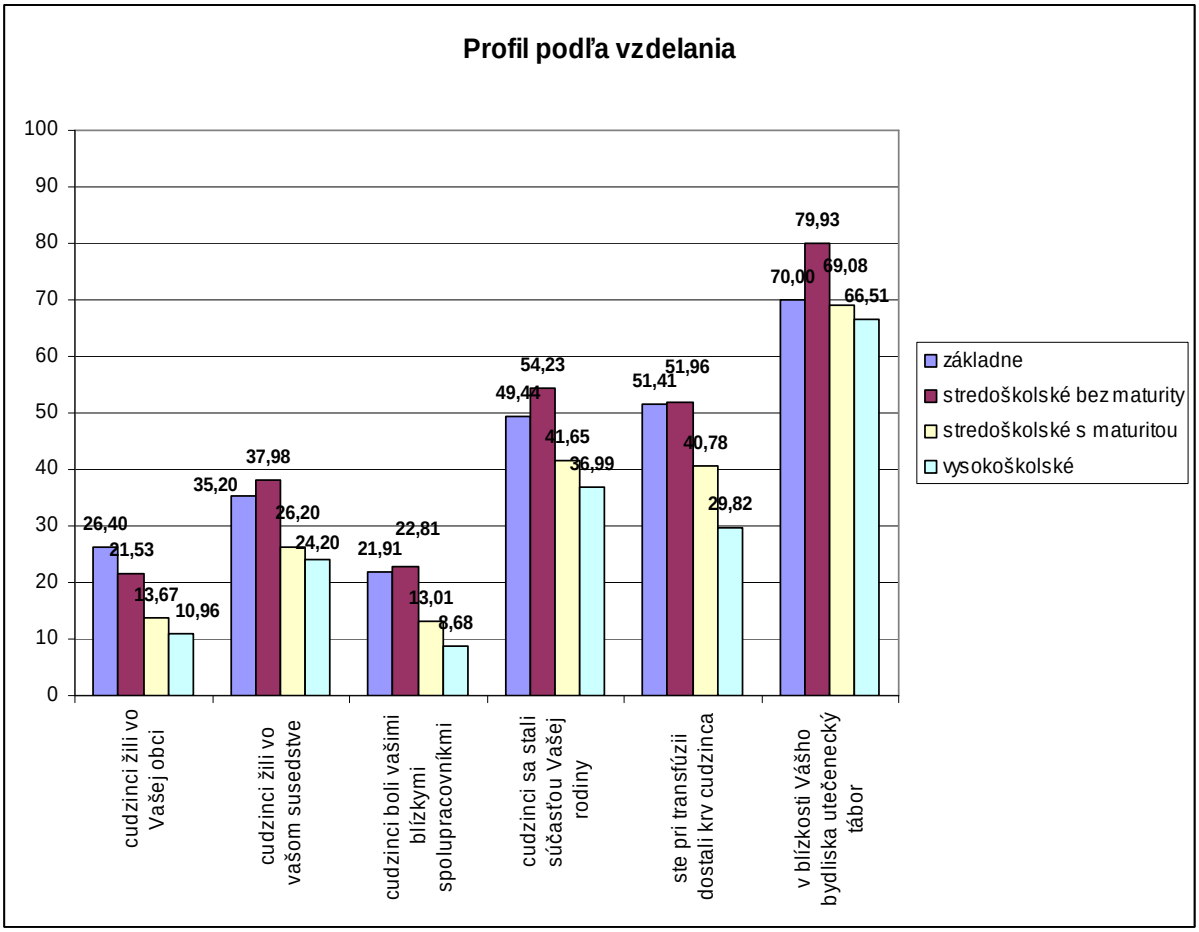
Graf 2. Profil kladných odpovedí na skúmané otázky podľa „pohlavia“



Menlivosť odpovedí ***podľa vzdelania*** charakterizujeme v grafe 3. P-hodnoty Chí-kvadrát testu uvedené v tabuľke ukazujú signifikantnú menlivosť podľa vzdelania pri všetkých skúmaných podotázkach. Ako vyplýva z podielu kladných odpovedí je podiel súhlasných odpovedí vyšší u respondentov s nižším vzdelaním (základné, stredoškolské bez maturity) ako u respondentov s vyšším vzdelaním (stredoškolské s maturitou, vysokoškolské).

vzdelanie	Prekážalo by Vám, keby,situácia-cudzinci žili vo Vašej obci	Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve	Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi	Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny	Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca	Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
P-hodnoty	0,000	0,001	0,000	0,000	0,000	0,003

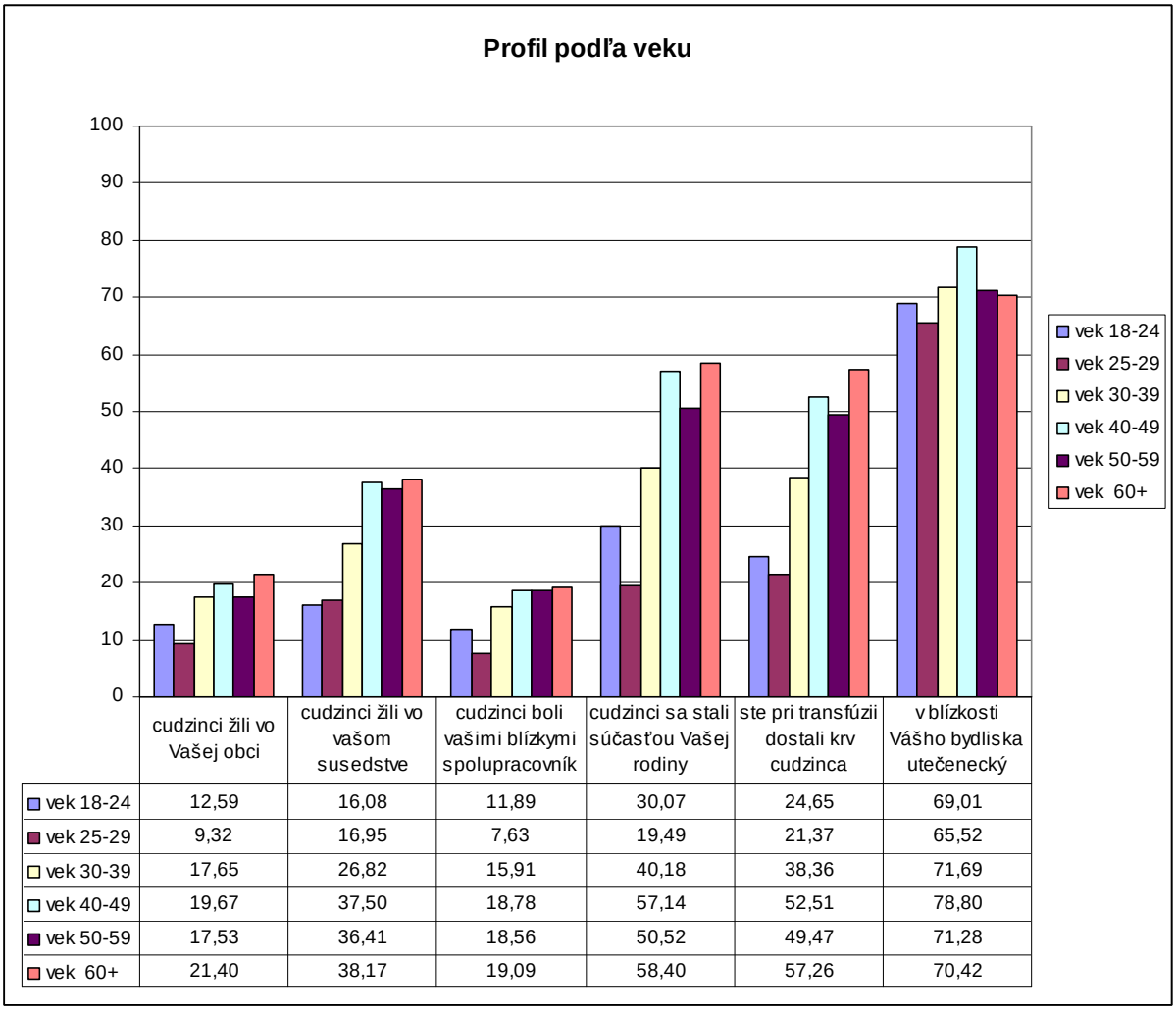
Graf 3. Profil kladných odpovedí na skúmané otázky podľa „vzdelania“



Aj menlivosť odpovedí ***podľa vzdelania*** je pre všetky skúmané podotázky štatisticky signifikantne odlišné, s výnimkou podotázky „Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov“, kde je P=0,190. Menlivosť podľa kategórií veku pri jednotlivých podotázkach, ako ukazuje graf 4, kolíše. Analýzou údajov možno zistiť, že tolerantnejší sú mladší respondenti (vekové kategórie 18-24r., 25-29r.) ako v strednom veku (vekové kategórie 30-39r., 40-49r.) a starší respondenti (vekové kategórie 50-59r. a 60r. a viac).

vek	Prekážalo by Vám, keby,situácia-cudzinci žili vo Vašej obci	Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve	Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi	Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny	Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca	Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
P-hodnoty	0,050	0,000	0,041	0,000	0,000	0,190

Graf 4. Profil kladných odpovedí na skúmané otázky podľa „veku“

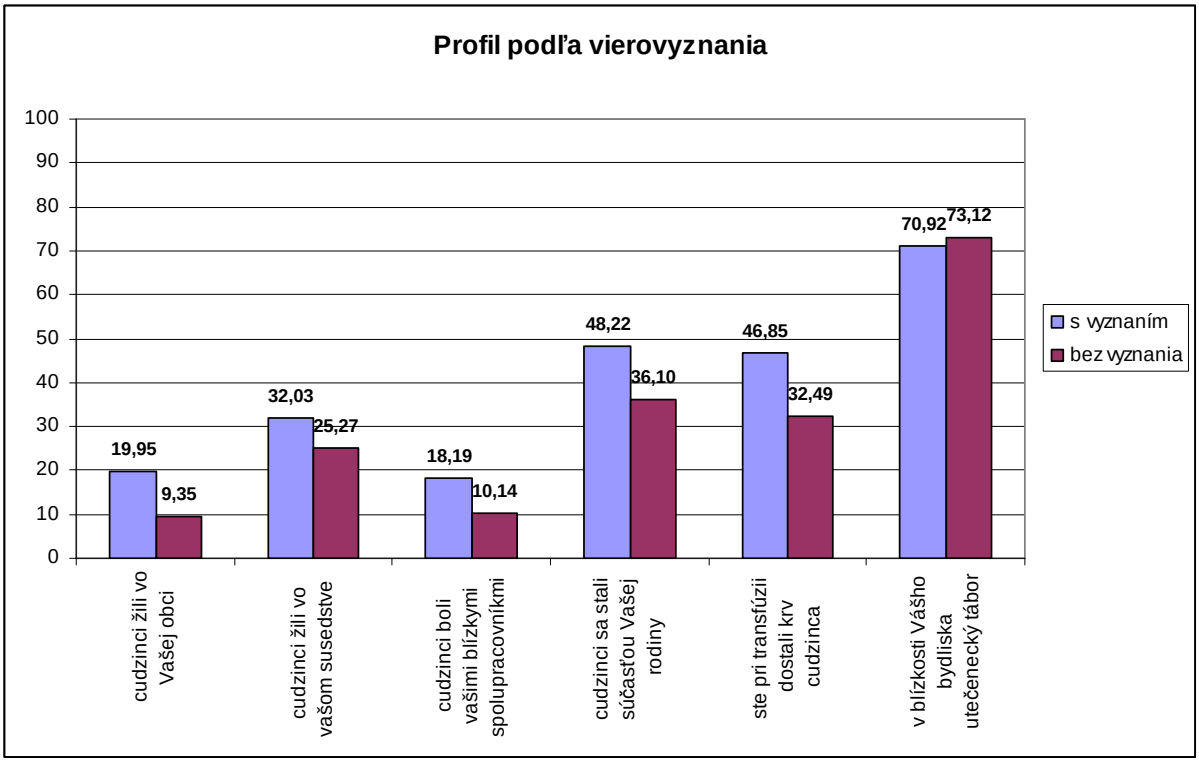


	cudzinci žili vo Vašej obci	cudzinci žili vo vašom susedstve	cudzinci boli vašimi blízkymi spolupracovník	cudzinci sa stali súčasťou Vašej rodiny	ste pri transfúzii dostali krv cudzinca	v blízkosti Vášho bydliska utečenecký
vek 18-24	12,59	16,08	11,89	30,07	24,65	69,01
vek 25-29	9,32	16,95	7,63	19,49	21,37	65,52
vek 30-39	17,65	26,82	15,91	40,18	38,36	71,69
vek 40-49	19,67	37,50	18,78	57,14	52,51	78,80
vek 50-59	17,53	36,41	18,56	50,52	49,47	71,28
vek 60+	21,40	38,17	19,09	58,40	57,26	70,42

Profil odpovedí na skúmané podotázky ***podľa vierovyznania*** prezentujeme v grafe 5. Z tabuľky P-hodnôt vyplýva, že okrem podotázky „Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov“ (P=0,483) sú odpovede u všetkých štatisticky signifikantne odlišné. Podľa úrovne podielu kladných odpovedí vidíme, že respondenti bez vyznania sú tolerantnejší ako respondenti s vyznaním. Pri poslednej podotázke je tento podiel opačný, ale ako sme už uviedli, nie je štatisticky signifikantne rozdielny.

vierovyznanie	Prekážalo by Vám, keby,situácia-cudzinci žili vo Vašej obci	Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve	Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi	Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny	Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca	Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
P-hodnoty	0,000	0,033	0,002	0,000	0,000	0,483

Graf 5. Profil kladných odpovedí na skúmané otázky podľa „vierovyznania“



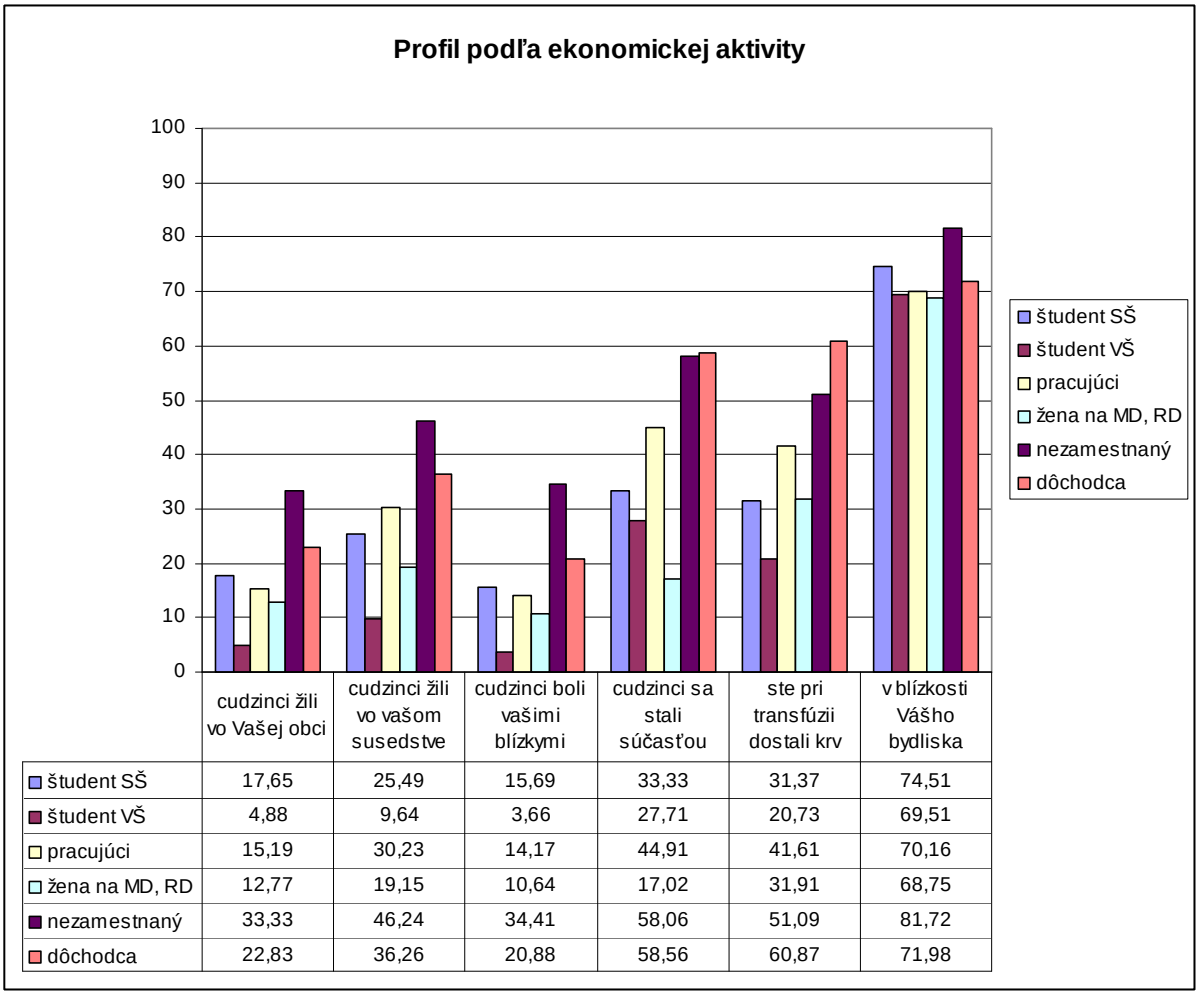
Profil odpovedí na skúmané podotázky ***podľa ekonomickej aktivity*** je uvedený grafe 6.

P-hodnoty Chí-kvadrát testu uvedené v tabuľke ukazujú signifikantnú menlivosť pri všetkých podotázkach s výnimkou poslednej „Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov“.

Menlivosť odpovedí u jednotlivých podotázok podľa kategórií ekonomickej aktivity kolíše. Pri všetkých 6-tich podotázkach sme zistili najvyšší podiel súhlasných odpovedí (najmenšiu tolerantnosť) u nezamestnaných. Pri prvých troch podotázkach a aj pri piatej podotázke boli najtolerantnejší študenti VŠ, pri štvrtej a šiestej ženy na MD a RD.

ekonomická aktivita	Prekážalo by Vám, keby,situácia-cudzinci žili vo Vašej obci	Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve	Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi	Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny	Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca	Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
P-hodnoty	0,000	0,000	0,000	0,000	0,000	0,314

Graf 6. Profil kladných odpovedí na skúmané otázky podľa „ekonomickej aktivity“

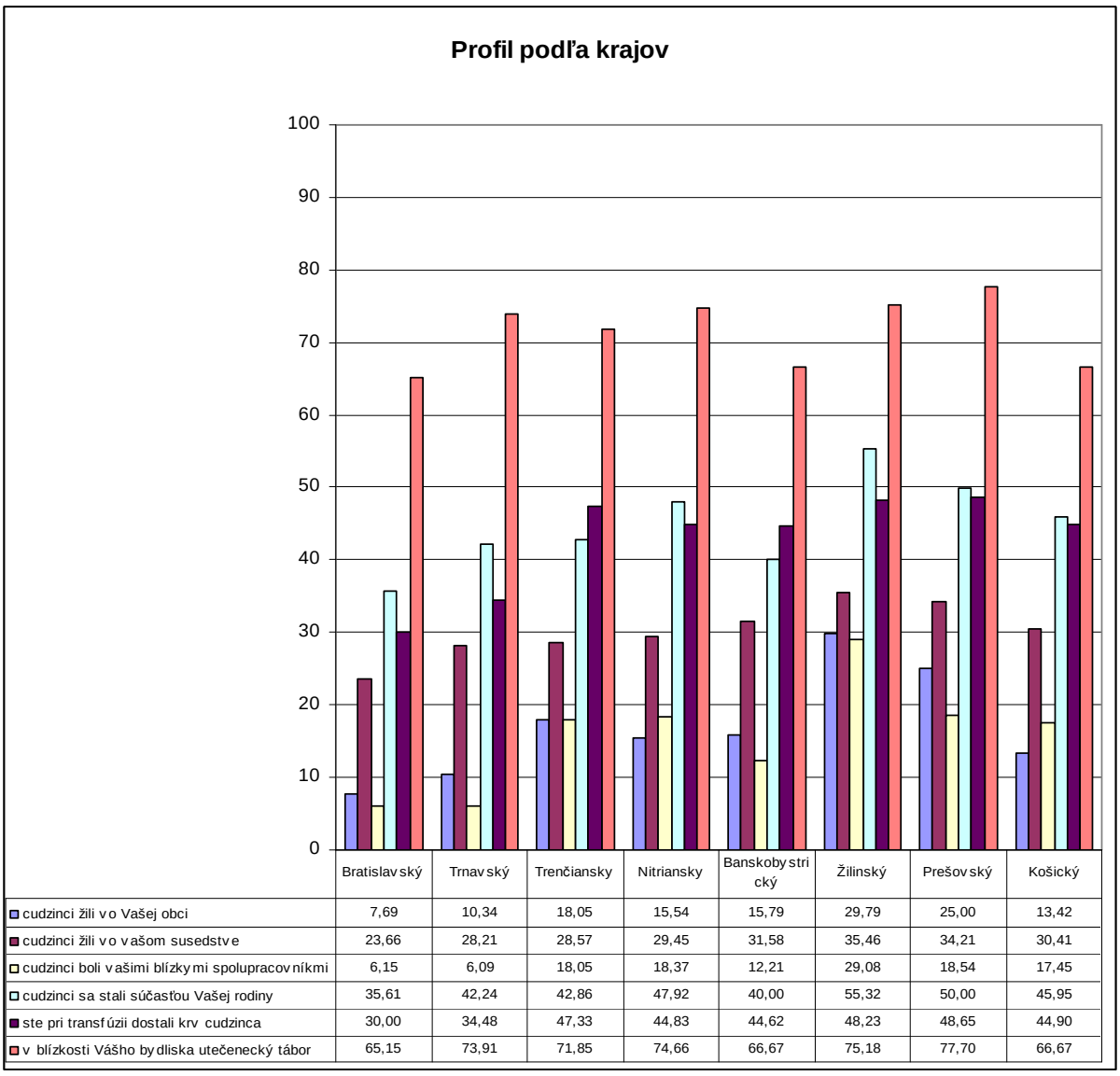


	cudzinci žili vo Vašej obci	cudzinci žili vo vašom susedstve	cudzinci boli vašimi blízkymi	cudzinci sa stali súčasťou	ste pri transfúzii dostali krv	v blízkosti Vášho bydliska
študent SŠ	17,65	25,49	15,69	33,33	31,37	74,51
študent VŠ	4,88	9,64	3,66	27,71	20,73	69,51
pracujúci	15,19	30,23	14,17	44,91	41,61	70,16
žena na MD, RD	12,77	19,15	10,64	17,02	31,91	68,75
nezamestnaný	33,33	46,24	34,41	58,06	51,09	81,72
dôchodca	22,83	36,26	20,88	58,56	60,87	71,98

Diferenciácia odpovedí na skúmané podotázky ***podľa krajov*** je zobrazená v profile v grafe 7. Vzhľadom na väčší počet kategórií znaku kraj je prezentovaná tabuľka dát „otočená“. Pri dvoch podotázkach „Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve“ a „Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov“ sme nezaznamenali štatisticky signifikantné rozdiely odpovedí podľa krajov

kraj	Prekážalo by Vám, keby,situácia-cudzinci žili vo Vašej obci	Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve	Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi	Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny	Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca	Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
P-hodnoty	0,000	0,518	0,000	0,040	0,015	0,157

Graf 7. Profil kladných odpovedí na skúmané otázky podľa „krajov“



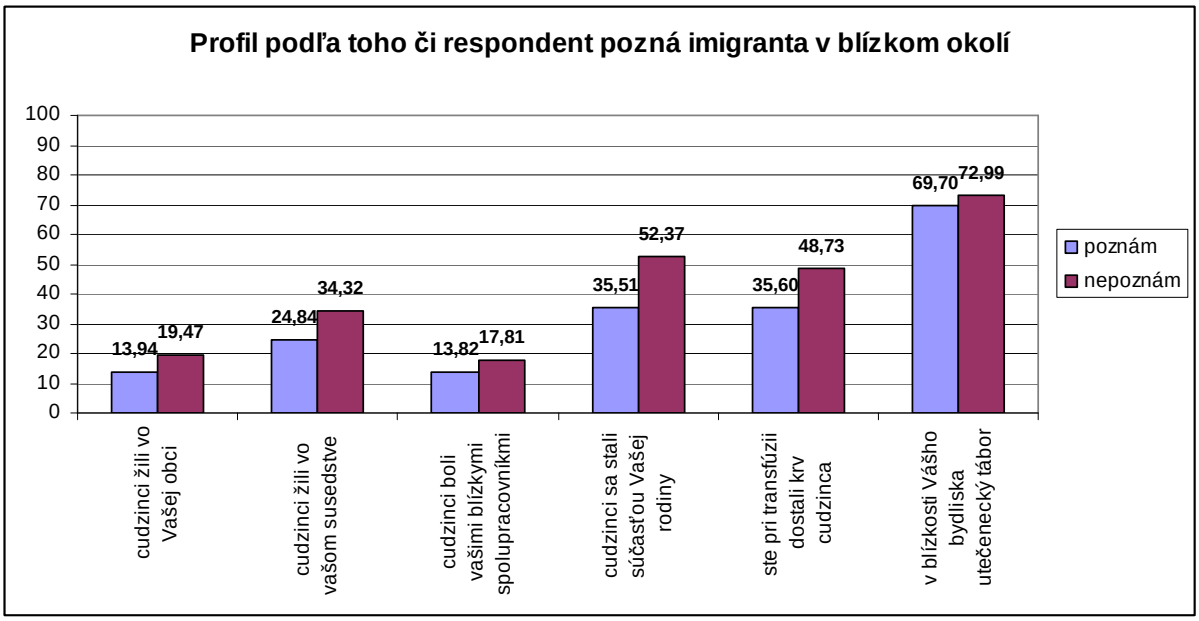
	Bratislavský	Trnavský	Trenčiansky	Nitriansky	Banskobystrický	Žilinský	Prešovský	Košický
cudzinci žili vo Vašej obci	7,69	10,34	18,05	15,54	15,79	29,79	25,00	13,42
cudzinci žili vo vašom susedstve	23,66	28,21	28,57	29,45	31,58	35,46	34,21	30,41
cudzinci boli vašimi blízkymi spolupracovníkmi	6,15	6,09	18,05	18,37	12,21	29,08	18,54	17,45
cudzinci sa stali súčasťou Vašej rodiny	35,61	42,24	42,86	47,92	40,00	55,32	50,00	45,95
ste pri transfúzii dostali krv cudzinca	30,00	34,48	47,33	44,83	44,62	48,23	48,65	44,90
v blízkosti Vášho bydliska utečenecký tábor	65,15	73,91	71,85	74,66	66,67	75,18	77,70	66,67

Na otázku „***Poznáte vo svojom blízkom okolí migranta z iného štátu, ktorý sa rozhodol žiť v SR?***“ neodpovedali iba 2 respondenti. Respondenti čo poznajú vo svojom okolí imigranta tvoria až 41,3% percenta dospelej populácie SR a tých čo imigranta vo svojom okolí nepoznajú je 57,7%. Je zaujímavé ako táto okolnosť diferencuje názory respondentov, Výsledky prezentujeme pomocou profilu kladných odpovedí na skúmané podotázky podľa toho či pozná alebo nepozná imigranta vo svojom blízkom okolí, výsledky sú v grafe 8. Zaujalo nás, že tolerantnejší postoj sme zistili u respondentov, ktorí poznajú imigranta vo svojom blízkom okolí.

Výsledky testov štatistickej signifikantkosti sú v prehľadnej tabuľke. Štatisticky nesignifikantné sú rozdiely pri odpovedi na podotázku „Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov“, na hranici štatistickej signifikantnosti sú diferencie odpovedí pri podotázke „Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi“, diferencie pri ostatných podotázkach sú štatisticky signifikantné.

Poznáte vo svojom blízkom okolí migranta z iného štátu, ktorý sa rozhodol žiť v SR?	Prekážalo by Vám, keby,situácia-cudzinci žili vo Vašej obci	Prekážalo by Vám, keby,situácia-cudzinci žili vo vašom susedstve	Prekážalo by Vám, keby,situácia-cudzinci boli vašimi blízkymi spolupracovníkmi	Prekážalo by Vám, keby,situácia-cudzinci sa stali súčasťou Vašej rodiny	Prekážalo by Vám, keby,situácia-ste pri transfúzii dostali krv cudzinca	Prekážalo by Vám, keby,situácia-keby bolo v blízkosti Vášho bydliska zriadený utečenecký tábor resp. iné zariadenie pre utečencov
P-hodnoty	0,016	0,001	0,076	0,000	0,000	0,233

Graf 8. Profil kladných odpovedí na skúmané otázky podľa otázky „Poznáte vo svojom blízkom okolí migranta z iného štátu, ktorý sa rozhodol žiť v SR?“



3. Závery

Na základe výsledkov nášho výskumu vplýva, že imigranti v podstate väčšine respondentov neprekážajú, len ich nechcú vo svojej blízkosti. Všeobecne možno konštatovať, že keď je situácia „vzdialená“ od respondenta, tak je podiel tých, ktorým by to prekážalo nižší, čo vidno už z profilu odpovedí za celý skúmaný súbor. Respondentom by najviac prekážal v blízkosti ich bydliska utečenecký tábor, túto odpoveď uviedlo 71,53% respondentov, ako je uvedené v prvej kapitole. Najmenší podiel respondentov 16,13% je takých, čo by im prekážalo mať cudzinca za blízkeho spolupracovníka, podobné je aj percento respondentov, ktorým by prekážalo keby cudzinci žili v ich obci. Profily podľa demografických otázok vykazujú podobný „priebeh“ ako profil za celý súbor dospelej populácie, sú však „vnútorne“ diferencované v závislosti od demografického znaku a skúmanej podotázky. Podrobná analýza výsledkov profilov za demografické znaky je v druhej kapitole. Spomenieme ešte zaujímavý poznatok, že respondenti, ktorí poznajú imigranta vo svojom blízkom okolí sú tolerantnejší ako tí, čo imigrantov vo svojom blízkom okolí nepoznajú.

Literatúra

[1] BEROVÁ L. (2012): *Názory verejnosti na migrantov a ich integráciu do spoločnosti*. PhD. dizertačná práca. Trnavská univerzita v Trnave, Fakulta zdravotníctva a sociálnej práce. Trnava 2012.

[2] BEROVÁ L., LUHA J., ŽÁKOVÁ M. (2012a): Názory verejnosti na migrantov a ich integráciu v SR: I. postoje k imigrantom prichádzajúcim do SR. *FORUM STATISTICUM SLOVACUM* 3/2012. SŠDS Bratislava 2012. ISSN 1336-7420.

[3] BEROVÁ L., LUHA J., ŽÁKOVÁ M. (2012b): Názory verejnosti na migrantov a ich integráciu v SR: III. čím nás môžu imigranti obohatiť. *FORUM STATISTICUM SLOVACUM* 6/2012. SŠDS Bratislava 2012. ISSN 1336-7420.

[4] KUBANOVÁ, J.: *Statistické metody pro ekonomickou a technickou praxi*. Statis, Bratislava 2008. Vydání třetí – doplněné. ISBN 978- 80-85659-47-4. pp 245.

[5] LINDA, B.: *Pravděpodobnost*. Monografie. Univerzita Pardubice, Pardubice 2010. ISBN 978-80-7395-303-4. pp 168.

[6] LUHA, J. (1985): *Testovanie štatistických hypotéz pri analýze súborov charakterizovaných kvalitatívnymi znakmi*. STV Bratislava 1985.

[7] LUHA J., BEROVÁ L., ŽÁKOVÁ M. (2012): Názory verejnosti na migrantov a ich integráciu v SR: II. začleňovanie imigrantov do spoločnosti. *FORUM STATISTICUM SLOVACUM* 4/2012. SŠDS Bratislava 2012. ISSN 1336-7420.

[8] *LUHA J. (2007): Kvótový výber. FORUM* STATISTICUM SLOVACUM 1/2007. SŠDS Bratislava 2007. ISSN 1336-7420.

[9] LUHA J. (2009): Matematicko-štatistické aspekty spracovania dotazníkových výskumov. *FORUM STATISTICUM SLOVACUM* 3/2009. SŠDS Bratislava 2009. ISSN 1336-7420.

[10] LUHA J. (2010): Metodologické zásady záznamu dát z rozličných oblastí výskumu. *FORUM STATISTICUM SLOVACUM* 3/2010. SŠDS Bratislava 2010. ISSN 1336-7420.

[11] PECÁKOVÁ I.(2008): *Statistika v terénních průzkumech*. Proffessional Publishing, Praha 2008. ISBN 978-80-86946-74-0.

[12] ŘEZANKOVÁ A.(2007): *Analýza dat z dotazníkových šetření*. Proffessional Publishing, Praha 2007. ISBN 978-80-86946-49-8.

[13] STANKOVIČOVÁ I., VOJTKOVÁ M.(2007): *Viacrozmerné štatistické metódy s aplikáciami*. IURA EDITION, Bratislava 2007, ISBN 978-80-8078-152-1.

Adresy autorov:

Ján Luha, RNDr., CSc.
Ústav lekárskej biológie, genetiky a klinickej genetiky LF UK a UN Bratislava
jan.luha@fmed.uniba.sk

Lenka Berová, Ing.,PhD.
Katedra sociálnej práce
FZaSP, Trnavská univerzita
lenka.berova@gmail.com

Martina Žáková, doc. PhDr., PhD.
Katedra sociálnej práce
FZaSP, Trnavská univerzita
martina.zakova@truni.sk

Štúdium cudzích jazykov na Podnikovohospodárskej fakulte EU
Study of foreign languages at the Faculty of Business Economics

Silvia Megyesiová, Lucia Tóthová, Silvia Kokošková

Abstract: The European Union is built on the principle of free movement of its citizens, capital, goods and services, bringing together more than half a billion people with a different cultural and linguistic backgrounds. The Commission's objective is to promote language learning and support of individual multilingualism to all citizens of the EU. Development of foreign language skills is important to encourage mobility within the Union and to the creation of a truly European labour market by allowing citizens to take full advantage of the freedom to work or study in another EU Member State. About 78.3 % of the students of the Faculty of Business Economics studies at longest the English language. Interest in learning Russian language comes gradually to an end, only 7.8% of the students take their living examination in Russian language.

Abstrakt: Európska únia budovaná na princípe voľného pohybu osôb, kapitálu, tovaru a služieb zjednocuje viac než pol miliardy obyvateľov s rozličným kultúrnym a jazykovým zázemím. Cieľom Komisie je podporovať jazykové vzdelávanie Európanov a podpora individuálnej viacjazyčnosti, aby všetci obyvatelia EÚ mohli v plnom rozsahu využívať slobodu študovať v niektorom z členských štátov resp. využívať európsky pracovný trh v plnom rozsahu. Študenti Podnikovohospodárskej fakulty Ekonomickej univerzity sa počas štúdií najdlhšie venovali štúdiu anglického jazyka a to až 78,3 % respondentov. Postupne zaniká záujem o štúdium ruského jazyka, keďže iba 7,8 % respondetov maturovalo z tohto jazyka. Vysokoškolskí študenti sa počas štúdia venujú štúdiu dvoch cudzích jazykov, čím spĺňajú základné princípy jazykového vzdelávania a viacjazyčnosti Európanov.

Key words: study of foreign languages, Eurobarometer, communication barriers, Fisher´s exact test, Nonparametric One-Way ANOVA.

Kľúčové slová: štúdium cudzích jazykov, Eurobarometer, bariéry v komunikácii, Fisherov exaktný test, neparametrická analýza rozptylu.

JEL classification: I21, C14, I29.

Úvod

Štúdium cudzích jazykov je v súčasnej dobe nevyhnutnosťou a zároveň otvára nové možnosti obyvateľov pre ich uplatnenie na pracovnom trhu tak domácom, ako aj zahraničnom. Slovenská republika je jednou z členských krajín Európskej únie. V súčasnosti EÚ tvorí zoskupenie 27 krajín, celkový počet obyvateľov dosiahol 503,7 milióna. Hlavnými cieľmi EÚ je zabezpečiť voľný pohyb občanov, kapitálu, tovarov a služieb pre všetkých obyvateľov krajín EÚ, pričom musíme mať na pamäti ich kultúrne, etnické a taktiež jazykové rozdiely. EÚ má v súčasnosti 23 úradných a pracovných jazykov. Počet úradných jazykov bude naďalej stúpať v dôsledku predpokladaného rozširovania EÚ. Väčšina dokumentov sa však z časových a finančných dôvodov neprekladá do všetkých 23 jazykov. Hlavnými pracovnými jazykmi Európskej komisie sú nasledovné jazyky: anglický, nemecký a francúzsky. Komisia podporuje schopnosť porozumieť a komunikovať v iných jazykoch než v materinskom ako jednu zo základných schopností, ktorú by mali mať všetci občania EÚ. Cieľom jazykovej politiky EÚ je teda podpora jazykového vzdelávania s predpokladom lepšieho zapojenia jazykovo zdatných Európanov do integrácie v rámci zjednoteného priestoru Únie.

Cieľ rozvoja znalostí cudzích jazykov je dôležitý z pohľadu mobility v rámci EÚ, prispeje k vytvoreniu skutočného európskeho pracovného trhu, pretože občanom umožní plne využívať slobody pracovať alebo študovať v inom členskom štáte. Pracovná sila s praktickými jazykovými a medzikultúrnymi znalosťami umožní európskym podnikom účinne konkurovať na svetovom trhu[1]. Cieľom Komisie je pritom zabezpečiť také systémy vzdelávania a odbornej prípravy, ktoré by umožnili rast individuálnej viacjazyčnosti do takej miery, že každý občan EÚ bude mať dostatočné praktické zručnosti aspoň v dvoch jazykoch okrem materinského jazyka.

1. Eurobarometer 386 v porovnaní so štúdiom cudzích jazykov študentov PHF EU

Špeciálny prieskum Eurobarometer 386 sa uskutočnil na jar 2012 na vzorke 27000 respondentov z rôznych sociálnych a vekových skupín vo všetkých 27 členských štátoch EÚ. Respondenti pri osobnom rozhovore odpovedali vo svojom materinskom jazyku. Z prieskumu vyplývajú nasledovné skutočnosti[2]:

- Takmer každý Európan (98 %) považuje znalosť cudzích jazykov za dôležitú pre budúcnosť vlastných detí a takmer 88 % je presvedčených, že znalosť iného ako materinského jazyka je pre nich veľmi užitočná.
- Takmer tri štvrtiny Európanov (72 %) súhlasia s cieľom EÚ celoplošne zaviesť výučbu aspoň dvoch cudzích jazykov od útleho veku.
- Dve tretiny (67 %) občanov EÚ radia angličtinu medzi dva najužitočnejšie jazyky. Medzi ďalšie najčastejšie uvádzané jazyky patria nemčina (17 %), francúzština (16 %), španielčina (14 %) a čínština (6).

Kvôli snahe o čo najlepšie porozumie názorov a problémov pri štúdiu cudzích jazykov sa uskutočnil anonymný prieskum medzi študentmi denného, ako aj externého štúdia na Podnikovohospodárskej fakulte Ekonomickej univerzity v Bratislave so sídlom v Košiciach v letnom semestri akademického roka 2011/2012. Prieskumu sa zúčastnilo 143 respondentov, z toho 52 mužov (36,4 %) a 91 žien (63,6 %). Priemerný vek respondentov bol 21,2, mediánový ako aj modálny vek respondentov dosiahol 21 rokov. V príspevku sa budeme venovať spracovaniu len niektorých otázok zisťovaných daným prieskumom.

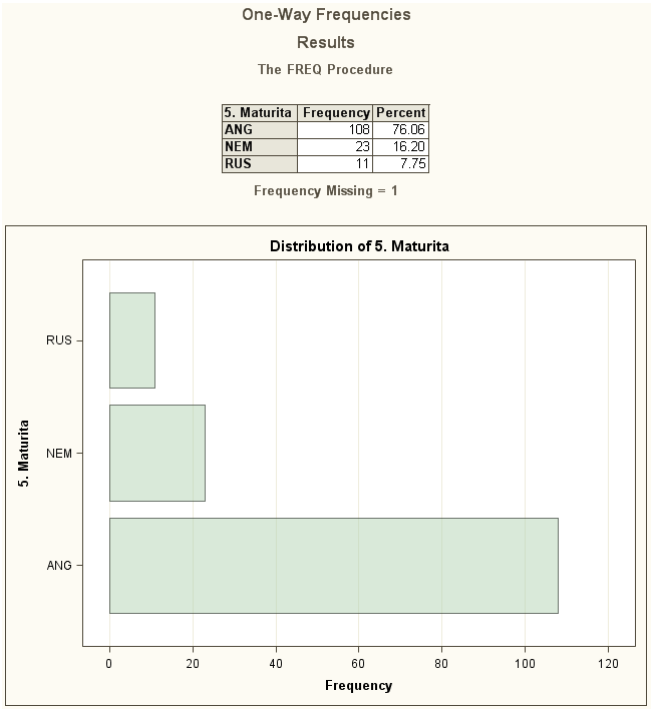
Štúdium cudzích jazykov má na Slovensku dlhoročnú tradíciu, kým v predchádzajúcom spoločenskom zriadení sa začínalo so štúdiom ruštiny ako povinného jazyka už na základnej škole, ďalším preferovaným jazykom bývala nemčina. Mladí ľudia v súčasnosti však majú pri výbere štúdia cudzích jazykov podstatne širšiu ponuku. V istej skupine otázok sme preto zisťovali, z ktorého cudzieho jazyka robili maturitnú skúšku, prijímacie skúšky na vysokú školu, a ktorý cudzí jazyk študovali najdlhšie. Výsledky ich odpovedí sú uvedené v nasledovných výstupoch.

Študenti PHF EU najčastejšie maturovali z anglického jazyka (76,05 %), z nemeckého jazyka maturovalo 16,2 % študentov a iba 7,75 % študentov maturovalo z ruského jazyka. Odpovede na otázku, z ktorého jazyka študenti maturovali vykazoval silnú asociáciu s dvoma ďalšími otázkami, a to s otázkou, z ktorého jazyka robili prijímacie pohovory na vysokú školu, a ktorý cudzí jazyk študujú najdlhšie. Asociáciu medzi týmito kvalitatívnymi odpoveďami sme hodnotili Fischerovým exaktným testom, pretože Chí-kvadrát test nie je vhodným v prípade tak nerovnomerného rozloženia početností jednotlivých skupín odpovedí. Samozrejme dalo sa očakávať, že medzi týmito otázkami bude štatisticky významná asociácia, pretože študenti maturujú a robia prijímacie pohovory najčastejšie práve z toho jazyka, ktorý študujú najdlhšie. Možno trochu na škodu je skutočnosť, že naša mladá

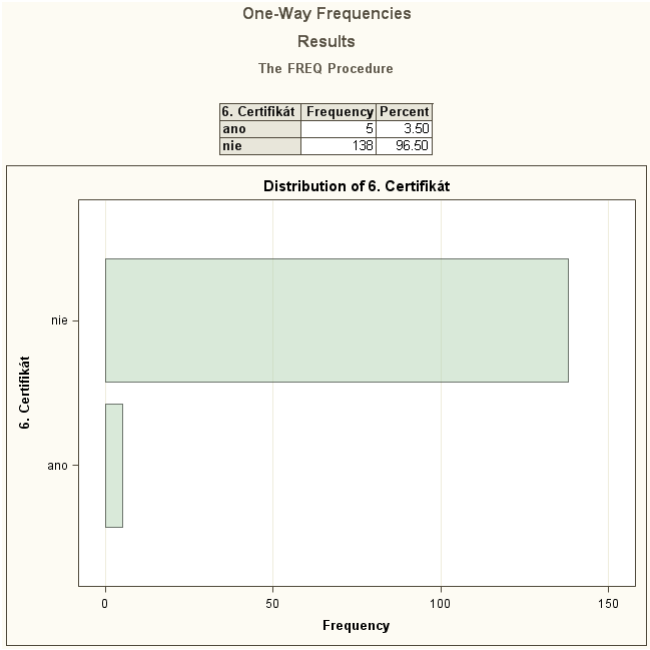
[1] Komisia Európskych spoločenstiev: Európsky indikátor jazykovej kompetencie

[2] http://ec.europa.eu/languages/languages-of-europe/eurobarometer-survey_sk.htm

generácia sa venuje hlavne štúdiu anglického jazyka a stráca sa tak trochu rozmanitosť štúdia cudzích jazykov.



Obr. 1 Početnosti študentov PHF EU podľa cudzieho jazyka počas maturitnej skúšky



Obr. 2 Početnosti študentov PHF EU vlastniacich certifikát z cudzieho jazyka

Najdlhšie sa študenti venovali štúdiu anglického jazyka, a to až 78,3 % našich respondentov. Je to možné vysvetliť aj ponukou výučby cudzích jazykov na základných školách, kde je ponuka do značnej miery okliešťená a súvisí aj s nedostatkom kvalifikovaných pedagógov, ktorí by umožnili rozšíriť výučbu o také jazyky, ako je jazyk taliansky, španielsky, francúzsky a podobne. Napriek tomu, že sa v súčasnosti začína s výučbou cudzích jazykov pomerne skoro a študenti vysokých škôl študujú aj pred prijatím na VŠ väčšinou

minimálne dva cudzie jazyky, je podiel tých respondentov, ktorí získali z ľubovoľného cudzieho jazyka medzinárodne akceptovaný certifikát mizivý. Iba 3,5 % respondentov uviedlo, že vlastní medzinárodne uznávaný certifikát z cudzieho jazyka, pritom v súčasnosti existuje možnosť náhrady maturitnej skúšky práve takýmto certifikátom istej úrovne. Pre študentov vysokých škôl by takýto certifikát mohol slúžiť aj ako doklad o jazykovej spôsobilosti, ktorá je nevyhnutná pri istých medzinárodných výmenných pobytoch, nehovoriac o možnosti ich využitia na trhu práce, či už domácom resp. trhu práce ostatných krajín EÚ.

Tab. 1 Asociácia medzi maturitou z cudzieho jazyka a prijímacími pohovormi z jazyka, ako aj najdlhšie študovaným cudzím jazykom

Table Analysis

Results

The FREQ Procedure

Table of 7. Prijímačky by 5. Maturita					
		5. Maturita			
		ANG	NEM	RUS	Total
7. Prijímačky					
ANG	Frequency	104	1	3	108
	Row Pct	96.30	0.93	2.78	
	Col Pct	96.30	4.35	27.27	
NEM	Frequency	1	22	0	23
	Row Pct	4.35	95.65	0.00	
	Col Pct	0.93	95.65	0.00	
RUS	Frequency	1	0	8	9
	Row Pct	11.11	0.00	88.89	
	Col Pct	0.93	0.00	72.73	
bez prijimac	Frequency	2	0	0	2
	Row Pct	100.00	0.00	0.00	
	Col Pct	1.85	0.00	0.00	
Total	Frequency	108	23	11	142
Frequency Missing = 1					

Statistics for Table of 7. Prijímačky by 5. Maturita

Statistic	DF	Value	Prob
Chi-Square	6	215.3607	<.0001
Likelihood Ratio Chi-Square	6	145.9069	<.0001
Mantel-Haenszel Chi-Square	1	66.5404	<.0001
Phi Coefficient		1.2315	
Contingency Coefficient		0.7763	
Cramer's V		0.8708	

WARNING: 58% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Table Probability (P)	2.250E-32
Pr <= P	1.013E-30

Effective Sample Size = 142

Table Analysis

Results

The FREQ Procedure

Table of 10. CJnajdlhšie by 5. Maturita					
		5. Maturita			
		ANG	NEM	RUS	Total
10. CJnajdlhšie					
ANG	Frequency	100	3	8	111
	Row Pct	90.09	2.70	7.21	
	Col Pct	92.59	13.04	72.73	
FRA	Frequency	2	0	1	3
	Row Pct	66.67	0.00	33.33	
	Col Pct	1.85	0.00	9.09	
NEM	Frequency	6	20	1	27
	Row Pct	22.22	74.07	3.70	
	Col Pct	5.56	86.96	9.09	
RUS	Frequency	0	0	1	1
	Row Pct	0.00	0.00	100.00	
	Col Pct	0.00	0.00	9.09	
Total	Frequency	108	23	11	142
Frequency Missing = 1					

Statistics for Table of 10. CJnajdlhšie by 5. Maturita

Statistic	DF	Value	Prob
Chi-Square	6	97.0572	<.0001
Likelihood Ratio Chi-Square	6	74.0453	<.0001
Mantel-Haenszel Chi-Square	1	30.7348	<.0001
Phi Coefficient		0.8267	
Contingency Coefficient		0.6372	
Cramer's V		0.5846	

WARNING: 67% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Table Probability (P)	7.385E-18
Pr <= P	5.759E-16

Effective Sample Size = 142

2. Rozdiely v názoroch na komunikáciu v cudzích jazykoch podľa pohlavia

Pri analýze odpovedí otázok 16 až 20 sme sa venovali ich vyhodnoteniu zvlášť pre obidve pohlavia, pretože nás zaujímalo, či existujú štatisticky významné rozdiely v odpovediach na tieto otázky. Znenie otázok bolo nasledované:

- *Otázka 16*: Štúdium cudzích jazykov považujem za dlhodobý a veľmi náročný proces.
- *Otázka 17*: Pri používaní cudzieho jazyka mám psychické bariéry, strach, že urobím chybu.
- *Otázka 18*: Obávam sa, že sa mi budú smiať, ak nebudem hovoriť v cudzom jazyku správne.

- *Otázka 19*: Je pre mňa nepríjemné komunikovať v cudzom jazyku napriek tomu, že som sa pripravoval.
- *Otázka 20*: Predtým, než začnem komunikovať v cudzom jazyku, chcem mať všetko vopred jasne premyslené.

Odpovede respondentov na hore uvedené otázky boli odstupňované od 1 – úplne súhlasím až po 7 – vôbec nesúhlasím. V tabuľke 2 sú uvedené priemery odpovedí študentov zvlášť pre mužov a ženy. Pretože táto premenná nemá normálne rozdelenie pre hodnotenie zhody stredných hodnôt sme zvolili neparametrickú analýzu rozptylu (*Nonparametric One-Way ANOVA*).

Respondenti oboch pohlaví sa pri odpovediach na 16 otázku v priemere zhodli a nebol dokázaný štatisticky významný rozdiel strednej hodnoty v ich opovediach na bežne používanej hladine významnosti 0,05. P-hodnota Kruskal-Wallis testu je P=0,7058. Pomerne nízka hodnota priemeru vyhodnotenia tejto otázky poukazuje na to, že štúdium cudzích jazykov považujú študenti PHF EU za dlhodobý a veľmi náročný proces.

Odpovede mužov a žien na otázku 17 sa štatisticky významne líšili. P-hodnota Kruskal-Wallis testu bola P=0,0004. Kým ženy dosiahli priemernú hodnotu odpovedí na túto otázku 3,1, u mužov bola priemerná odpoveď rovná až 4,25. To znamená, že muži majú pri používaní cudzieho jazyka menší strach, menšie psychické bariéry ako ženy, čo môže súvisieť s ich vyššiou suverénnosťou, priebojnosťou, čo je typickejšie pre "silnejšie" pohlavie.

Podobne si môžeme vysvetliť aj štatisticky významný rozdiel v odpovediach na otázku 18 (P-hodnota testu P=0,0225. Kým muži dosiahli priemernú úroveň odpovedí 4,63, priemer odpovedí žien bol 3,93. Ženy majú teda väčšiu obavu z toho, či budú v cudzom jazyku hovoriť správne, a či reakcie na ich nesprávnu komunikáciu nebudú prijaté úsmevne. Muži sa aj v tomto prípade obávajú výrazne menej.

Tab. 2 Priemery odpovedí respondentov na otázky 16 – 20 podľa pohlaví

Summary Tables

		Mean
1. Pohlavie(M/Z)		
M	16. ŠtúdiumCJ (1-7)	2.58
	17. Bariéry CJ(1-7)	4.25
	18. Výsmech CJ(1-7)	4.63
	19. Komunikácia CJ(1-7)	4.48
	20. Komunkáciupremyslieť(1-7)	2.75
Z	16. ŠtúdiumCJ (1-7)	2.44
	17. Bariéry CJ(1-7)	3.10
	18. Výsmech CJ(1-7)	3.93
	19. Komunikácia CJ(1-7)	3.99
	20. Komunkáciupremyslieť(1-7)	2.63

Otázky 19 a 20 boli medzi oboma pohlaviami zodpovedané porovnateľne. Na otázku 19 študenti zhodne odpovedali neutrálnym postojom voči nepríjmeným pocitom pri komunikácii v cudzom jayku, ak sa na túto komunikáciu vopred pripravovali (P-hodnota Kruskal-Wallis testu P=0,1195). Na otázku 20 zhodne obe pohlavia odpovedali, že predtým, ako začnú komunikovať v cudzom jazyku, chcú mať komunikáciu vopred premyslenú (P-hodnota Kruskal-Wallis testu P=0,7386).

Z odpovedí na otázky 19 a 20 je zrejmé, že študenti neradi improvizujú a najradšej by sa na komunikáciu v cudzom jazyku vopred pripravili, čo však v prípade konverzácie nie je vždy možné. Študenti si teda nie sú istí svojimi znalosťami a rýchlymi reakciami v cudzom jazyku.

Čo spôsobuje študentom PHF EU pri štúdiu resp. pri konverzácii v cudzom jazyku problémy, je predmetom hodnotenia ďalších otázok anonymného dotazníka, ktoré z priestorových dôvodov nie sú predmetom tohto príspevku.

3. Záver

Stratégiou lídrov Európskej únie je zabezpečiť vzdelávanie minimálne v dvoch cudzích jazykoch od útleho veku. Cieľom je, aby každý Európan získal skúsenosti a dokázal komunikovať aj v iných jazykov než vo vlastnom materinskom jazyku. Študenti PHF EU študujú dva cudzie jazyky počas svojho štúdia na fakulte. Anglický jazyk je napreferovanejším jazykom, ktorý študovali respondenti väčšinou už od základnej školy. Študenti napriek dlhodobému štúdiu cudzích jazykov neabsolvujú medzinárodne uznávané skúšky o ich znalosti, ktoré by im na základe certifikátov uľahčili medzinárodnú akceptáciu znalosti cudzieho jazyka počas štúdia resp. pri hľadaní práce v rámci jednotného pracovného trhu krajín EÚ.

Literatúra

[1] CHAJDIAK, J. – KRIŠKOVÁ, A. 2012. Usporiadanie otázok dotazníka Správanie podporujúce zravie podľa intenzity celkového hodnotenia zdravotníckych asistentov. In: *Forum Statisticum Slovacum* 2/2012. SŠDS Bratislava. 2012. ISSN 1336-7420.

[2] FIALA, T. – LANGHAMROVÁ, J. – MISKOLCZI, M. 2011. Předpokládaný vývoj úrovně vzdělání populace České republiky v letech 2000-2050. In: *Forum Statisticum Slovacum* 7/2011. SŠDS Bratislava. 2011. ISSN 1336-7420.

[3] LUHA, J. 2009. Matematicko-štatistické aspekty spracovania dotazníkových výskumov. In: *Forum Statisticum Slovacum* 3/2009. SŠDS Bratislava. 2009. ISSN 1336-7420.

[4] LUHA, J. 2006. Štatistické metódy analýzy kvalitatívnych znakov. In: *Forum Statisticum Slovacum* 2/2006. SŠDS Bratislava. 2006. ISSN 1336-7420.

[5] LÖSTER, T. – ŘEZANKOVÁ, H. – LANGHAMROVÁ, J. 2009. *Statistické metody a demografie.* 1. vydanie. VŠEM, Praha. s. 297. ISBN 978-80-86730-43-1.

[6] STANKOVIČOVÁ, I. – VOJTKOVÁ, M. 2007. *Viacrozmerné štatistické metódy s aplikáciami.* IURA EDITION, Bratislava 2007, ISBN 978-80-8078-152-1.

[7] KOMISIA EURÓPSKYCH SPLOČENSTIEV. 2005. *Oznámenie Komisie Európskemu parlamentu a Rade. Európsky indikátor jazykovej kompetencie*. Brusel, KOM (2005) 356 v konečnom znení. Dostupné z <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2005:0356:FIN:sk:PDF>.

[8] http://ec.europa.eu/languages/languages-of-europe/eu-languages_sk.htm

Adresa autora (-ov):

Silvia Megyesiová, Ing. PhD.
Podnikovohospodárska fakulta, EU
Tajovského 13, 041 30 Košice
silvia.megyesiova@euke.sk

Lucia Tóthová, PhDr.
Podnikovohospodárska fakulta, EU
Tajovského 13, 041 30 Košice
lucia.tothova@euke.sk

Silvia Kokošková, Mgr.
Podnikovohospodárska fakulta, EU
Tajovského 13, 041 30 Košice
silvia.kokoskova@euke.sk

Position of ICT Students among Other Unemployed Graduates in the Czech Republic

Postavení studentů ICT mezi ostatními nezaměstnanými absolventy v České republice

Martina Miskolczi, Jitka Langhamrova, Tomas Fiala

Abstract: The article introduces analysis of position of unemployed graduates of ICT branch among other unemployed graduates at the university level. The hypothesis is that position and status of unemployed ICT graduates is more advantageous compared to unemployed people graduated from other branches, especially in comparison with humanities. Second, it is assumed that position of ICT students among unemployed graduates during economic crisis did not worsen as much as for other branches, especially non-technical ones. Analysis of graduated students from various universities in the Czech Republic, who do not have a job, was presented as a comparison of main branches and selected sub-branches. Number of ICT unemployed graduates and economics unemployed graduates declines whereas number and proportion of unemployed graduates from humanistic branches grows. Advantageous position of ICT graduates on the labour market even in the period of economic crisis is confirmed as well.

Abstrakt: Článek představuje analýzu postavení nezaměstnaných absolventů ICT oborů mezi ostatními nezaměstnanými absolventy univerzit. Hypotéza zní, že postavení a status nezaměstnaných studentů ICT je příznivější ve srovnání s nezaměstnanými osobami – absolventy jiných oborů, zejména ve srovnání s humanitními obory. Druhá hypotéza předpokládá, že postavení studentů ICT mezi nezaměstnanými absolventy se během ekonomické krize nezhoršilo tak, jako tomu bylo u ostatních oborů, speciálně netechnických. Analýza studentů–absolventů různých univerzit v České republice, kteří nemají práci, byla prezentována jako porovnání studijních směrů a vybraných oborů. Počet nezaměstnaných absolventů ICT a ekonomie klesá, zatímco počet a podíl nezaměstnaných absolventů humanistických oborů roste. Potvrdilo se také příznivější postavení absolventů ICT na trhu práce, a to i v době ekonomické krize.

Key words: unemployment, ICT graduates, economic crisis, labour market in the Czech Republic

Klíčová slova: nezaměstnanost, absolventi ICT, ekonomická krize, trh práce v České republice

JEL classification: J62, J64

Introduction

Unemployment is a very important macroeconomic indicator with substantial impact on political, economic and social stability. In developed countries, two categories of tools are classified: active and passive policies of employment and unemployment. Active employment policy represents tools that aim to increase number of unemployed people who find a job or for whom suitable job position is created. Here, a special attention is paid to several selected groups of unemployed people such as women, pregnant women and women with babies,

young people (bellow 19 years), graduates[1] without or with short work experience only, other nationalities, socially excluded persons, people with any disability etc. This care is embedded in The Employment Act No. 435/2004 Coll.

Special care for young people and persons, who just finished their school or who worked only very shortly after their graduation, proceeds from the threat that these young people do not acquire work habits if they do not find proper job in young age. This would have devastating impact both on them personally and on society as well.

Objective of this article is to analyse position of unemployed graduates of ICT branch among other unemployed graduates, i.e. people who are unemployed and graduated from other types of schools and branches. The hypothesis is that position and status of unemployed ICT graduates is more advantageous compared to unemployed people graduated from other branches, especially in comparison with humanities. Second hypothesis concerns economic crisis and its impact on position of ICT graduates. It is assumed that position of ICT students among unemployed graduates did not worsen as much as for other branches, especially non-technical ones. Comparisons will be focused on university level.

1. Data

In the Czech Republic, statistics from Ministry of Labor and Social Affairs (MLSA) regarding young people and graduates are available. Number of unemployed young people and graduates registered at labor offices is published monthly distributed by regions and districts. Overview of graduates by their school is published half-yearly, also by regions and districts. Each school and studied branch is classified[2] according to the grade of school (for example secondary level: general, technical, vocational; university level: first stage – bachelor studies, second stage – master studies, third stage – doctoral studies etc.) and according to the branch (for example 13-01 Geography, 13-02 Cartography, 13-03 Demography etc.).

Such the statistics are available as of the end of April and the end of September each year. The difference between the two periods within a year is caused by different behavior of students during the school year. For example in September, graduated students register at labor offices after their last summer holidays, i.e. during the September. In this case, statistics from September will be analyzed.

Regional analysis of unemployed graduates with university level is complicated because people register at labor offices according to their place of permanent residence, which may and often does differ from the address of the school and such, it differs from MLSA statistics. Under this condition, analysis for the whole Czech Republic will be prepared.

2. Position of ICT Graduates among Unemployed Graduates

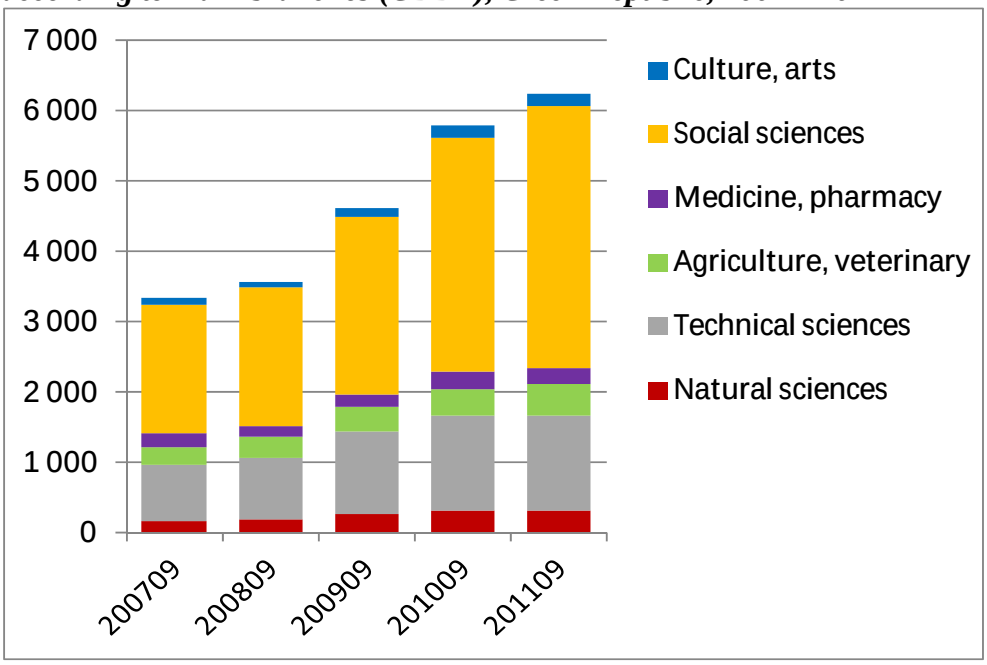
Number of graduates who are unemployed grows each year. In September 2011 the total number of unemployed graduates increased by 3,000 and almost doubled since the situation in

[1] Based on the agreement between Ministry of Education, Youth and Sports (MEYS) and Ministry of Labor and Social Affairs (MLSA) in the Czech Republic, definition of *graduate* since Jan 1st, 2004 is as follows: a jobless person that is registered at the labor office according to his/her place of permanent residence at the given date (end of April or end of September in given year) AND the time since successful graduation is shorter than 2 years.

[2] Classification of Basic Branches of Education (CBBE)

September 2007. Absolute number of unemployed graduates from technical branches, grouped in the first and second category (*natural sciences* comprise mathematics, geography, chemistry, biology physics and informatics, *technical sciences* include mining, industry, engineering, telecommunications and ICT, applied chemistry, textile industry, architecture and construction, transportation and similar) grows but in other branches the growth is even faster.

Figure 1: Number of unemployed graduates with university level (categories R, T, V) according to main branches (CBBE), Czech Republic, 2007 – 2011



Source: MLSA, own calculation

Structure of unemployed graduates presented in the following table shows that proportion of unemployed graduates from technical specialization declined between 2009 and 2011 while proportion of people who graduated at *social sciences* (philosophy, economics, tourism, sales, social care, humanities, law, services, history, pedagogy, psychology etc.) or *culture/arts* increased. What is interesting: between 2007 and 2009 proportion of technically oriented unemployed graduates increased and people graduated from social sciences and arts accounted for lower proportion compared to September 2007. This might indicate, among other reasons, changing structure of the labor demand at the labor market in the first year of the crisis.

Table 1: Structure of unemployed graduates according to main branch, Czech Republic, 2007 – 2011

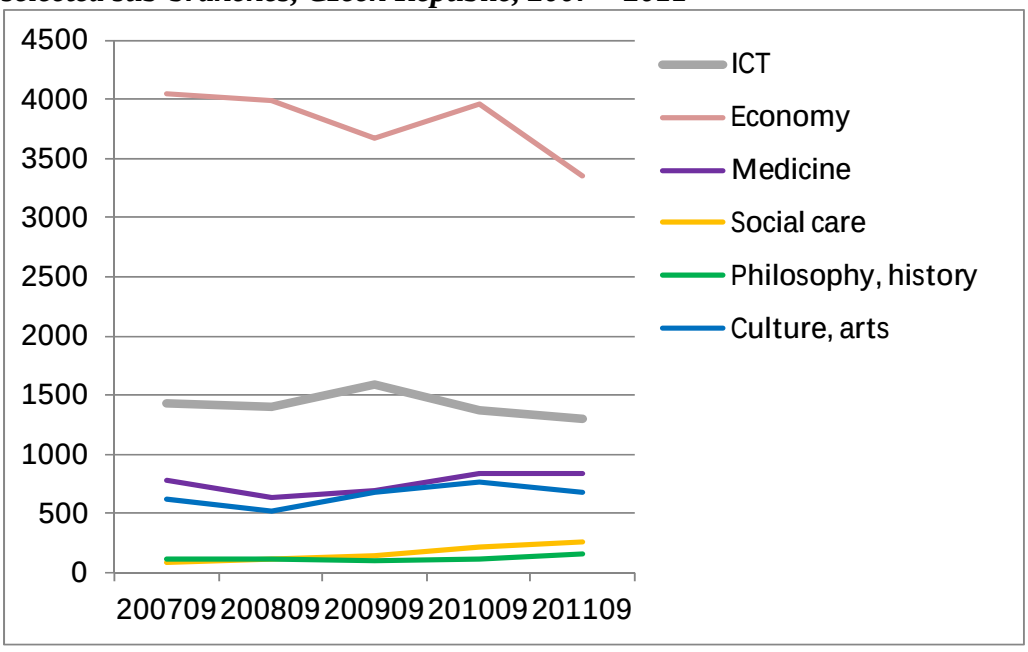
(as of Sep 30th)	**2007**	**2008**	**2009**	**2010**	**2011**
Natural sciences	5.2%	5.5%	5.8%	5.6%	5.0%
Technical sciences	23.9%	24.5%	25.5%	23.1%	21.9%
Agriculture, veterinary	7.2%	7.9%	7.3%	6.4%	6.9%
Medicine, pharmacy	5.9%	4.4%	4.2%	4.3%	3.8%
Social sciences	54.9%	55.5%	54.6%	57.7%	59.5%
Culture, arts	2.9%	2.1%	2.6%	2.9%	2.9%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Source: MLSA, own calculation

In the classification, two specific sub-branches connected with ICT could be selected: 18_Informatics and 26_Electrotechnics, telecommunications and computing. In the following figure, other sub-branches were selected to compare success of their graduates: economy, medicine, social care, philosophy & history and culture & arts.

It is clear that number of unemployed graduates from ICT and economic sub-branches decreases while number of unemployed graduates from other selected sub-branches (medicine, humanistic orientation) grows. Following figure also shows that number of unemployed graduates – economists is the highest, number of ICT graduates without job oscillates around 1,500 and other sub-branches are lower. These trends reflect opposing development of technical and humanistic branches – whereas more and more graduates from humanities cannot find suitable job position, technical branches and technical schools do not have enough candidates and their graduates are demanded on the market.

Figure 2: Number of unemployed graduates with university level (categories R, T, V) for selected sub-branches, Czech Republic, 2007 – 2011

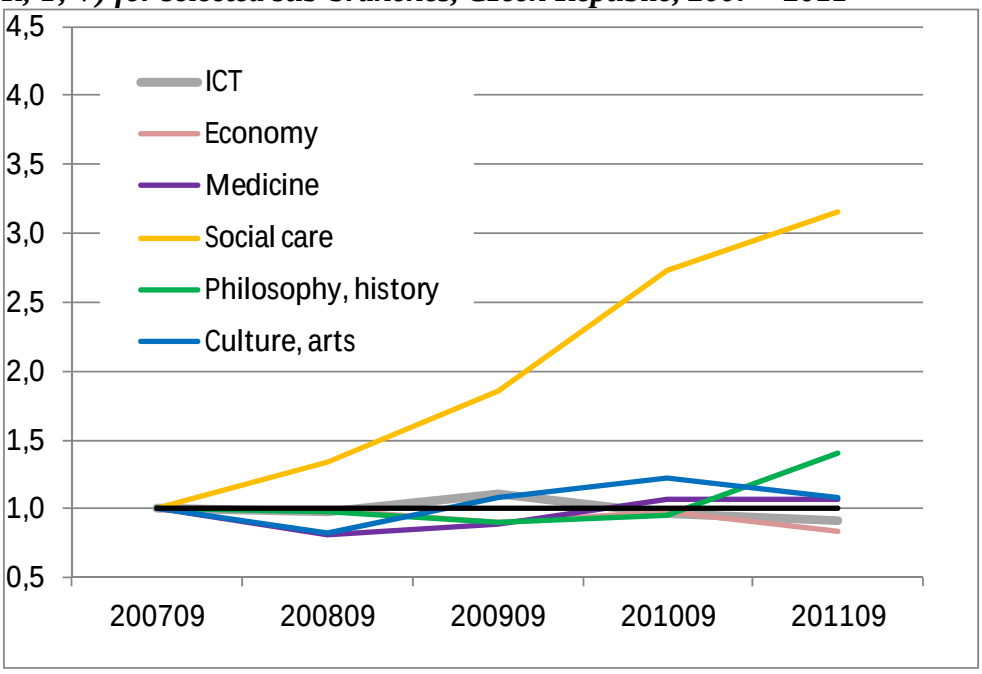


Source: MLSA, own calculation

3. Impact of Economic Crisis

Economic crisis influenced labour market for both graduates and experienced employees. In case of young people with recently finished education the impact was even harder (Miskolczi, 2010; Miskolczi, Langhamrová, 2011). Reasons for worsened position of graduates are: their lack of experience and skills for specific job position and expected fluctuation.

Figure 3: Base index of number of unemployed graduates with university level (categories R, T, V) for selected sub-branches, Czech Republic, 2007 – 2011



Source: MLSA, own calculation

In the figure above it is clear that during period of economic crisis number of ICT graduates registered at labor offices even slightly decreases in 2011. These people found their position on the labor market and were able, even in heavier conditions of economic crisis find a job, even more often. The same applies for economics graduates.

On the other hand, students graduated from humanistic branches are affected with the crisis more heavily – number of unemployed graduates from these types of school is growing, in some cases very dramatically. The growth is caused both by growing number of graduates as well as lower ability of the labor market to absorb these people.

4. Conclusion

Analysis of graduated students from various universities in the Czech Republic, who do not have a job, was presented as a comparison of main branches and selected sub-branches. Number of ICT unemployed graduates and economics unemployed graduates declines whereas number and proportion of unemployed graduates from humanistic branches grows. Advantageous position of ICT graduates on the labor market even in the period of economic crisis is confirmed as well. This conclusion has to be supported by more detailed analysis of unemployed graduates.

References

[1] The Employment Act No. 435/2004 Coll.

[2] Ministry of Labor and Social Affairs of the Czech Republic

[3] MISKOLCZI, MARTINA. Analýza nezaměstnanosti mladistvých a absolventů v krajích ČR. Praha 13.12.2010 – 14.12.2010. In: *Reprodukce lidského kapitálu – Vzájemné vazby a souvislosti* [CD-ROM]. Praha : VŠE, 2010, s. 1–8. ISBN 978-80-245-1697-4.

[4] MISKOLCZI, MARTINA, LANGHAMROVÁ, JITKA. Analýza zaměstnanosti a nezaměstnanosti vybraných skupin populace v době ekonomické krize. Praha 05.12.2011 – 06.12.2011. In: *RELIK 2011 – Reprodukce lidského kapitálu vzájemné vazby a souvislosti* [CD-ROM]. Slaný : Melandrium, 2011, s. 1–10. ISBN 978-80-86175-75-1.

Addresses

Mgr. Ing. Martina Miskolczi, MBA
martina.miskolczi@vse.cz

doc. Ing. Jitka, Langhamrová, CSc.
langhamj@vse.cz

RNDr. Tomáš Fiala, CSc.
fiala@vse.cz

Department of Demography
Faculty of Informatics and Statistics
University of Economics in Prague
nám. W. Churchilla 4
130 67 Prague 3
Czech Republic

Supported by research project IGA F4/29/2011 Analysis of population ageing and impact on labour market and economic activity

Analysis of Marriage Carrier Using Multistate Analysis and Multistate Life Tables

Analýza sňatečnostní kariéry s využitím vícestavové analýzy a vícestavových tabulek života

Martina Miskolczi, Jitka Langhamrova, Jana Langhamrova

Abstract. The article introduces application of multistate analysis on so called 'marriage career' for the case of the Czech Republic in the period of 2001-2010. Objective of the article is to clarify changes in the behaviour of women in the Czech Republic related to their marriage decision over last ten years, using method of multistate life tables. It is confirmed that women more often decide not to marry and stay unmarried. Over a studied period, women in the Czech Republic changed their decision toward marriages, namely younger women in the age 15–30 years. Tendency not to marry is stronger among young women; they stay unmarried and probably live in partnerships without official marriage. Women over 30 years changed their behaviour only little.

Abstrakt: Článek seznamuje s aplikací vícestavové analýzy na tzv. ‚sňatečnostní kariéru' v případě České republiky v období let 2001-2010. Cílem článku je vysvětlit změny v chování žen v ČR ve vztahu k jejich rozhodnutí ohledně sňatku během uplynulých deseti let, s využitím metody vícestavových tabulek života. Potvrzuje se, že ženy se častěji rozhodují neprovdat se a zůstat neprovdána. Během studovaného období změnily ženy v ČR své rozhodování vzhledem ke sňatkům, zejména mladší ženy ve věku 15-30 let. Tendence nevdávat se je silnější mezi mladými ženami, které zůstávají neprovdané a pravděpodobně žijí v partnerstvích bez oficiálního sňatku. Ženy přes 30 let změnily své chování jen nepodstatně.

Key words: mathematical demography, multistate life tables, length of stay in the status, marital status, marriage, single, married, divorced, widowed

Klíčová slova: matematická demografie, vícestavové tabulky života, délka pobytu ve stavu, rodinný stav, sňatek, svobodná, vdaná, rozvedená, vdova

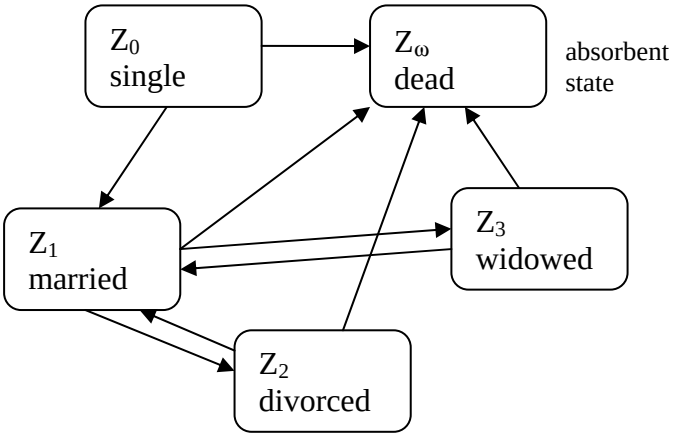
JEL classification: J12, C31

1. Introduction

Multistate demography is a part of demography that analyses states of demographic subjects and events that cause these states. For simplicity and mathematical modelling, usually only one type of demographic event is studied and the sequence of events is called 'career'. Here, 'marriage career' or 'marital status career' of women is studied in the period of 2001–2010 in the Czech Republic in order to analyze changes in the trend of nuptiality (marriages), divorcity and mortality in the Czech Republic after 2000. Method of multistate life tables enables to study occurrences of events and transfers from one state to another. (Rogers, 1975; Rogers, 1980)

Objective of this article is to verify changes in the behaviour of women in the Czech Republic related to their marriage decision over last ten years. Current trend is that people live together in cohabitation, in partnership without official marriage and that proportion of children born outside of marriage increases.

2. Definition of Terms



Events U_i: birth (U_0), marriage (U_1), divorce (U_2), death of the partner (become a widow) (U_3), death (U_ω). *States* (Z_i): single (Z_0), married (Z_1), divorced (Z_2), widowed (Z_3) and dead (Z_ω). States could be absorbent or transient. Absorbent states cannot be left, subject remains in the state. Usually, it is represented by the state 'dead'.

Randomness: Occurrence of events is considered to be random. It is assumed that an individual with certain realization of his life cycle can be found in the population with some probability.

Multistate life tables are the extension of standard (one-state) life tables. They present additional dimension(s) – marital status.

3. Intensities of Probability

For the calculation, it is necessary to estimate unknown probability distribution P(U, x, t | Z) or P(U, x | Z) for each event *U*, state *Z*, completed age *x* and *t* – time from the last transition into state Z. Probability distribution can be defined by distribution function, probability density function or intensity of probability (also called hazard rate or risk function). Using absolute frequencies (number of events, number of subjects exposed to a risk of event and length of the exposure), the intensity of probability can be estimated. (Koschin, 1992) It was proved that such an estimate is the best unbiased estimate of the intensity probability, which is constant in given interval. (Rogers, 1975)

In case of entire population, demographic data are available in annual distribution and thus, assumption about constant trends during a year has to be made. Further, each individual contributes one year to the final sum; each individual who came into the population during the year or left the population during the year contributes one half of the year. This corresponds with the assumption of uniform distribution of demographic events during the year.

Estimate of intensity of mortality: is the specific mortality rate $m_{x,t} = \frac{M_{x,t}}{\overline{S}_{x,t}.1}$, x = 0, 1, …, ω–1, where multiplication by 1 in the denominator represents length of exposure. Similarly, intensity of fertility can be estimated by $j_{x,t} = \frac{N_{x,t}}{\overline{S}_{x,t} - \frac{1}{2} N_{x,t}}$, where correction in the denominator eliminates for one half of the year those women who gave birth in the same calendar year. In the same way, specific marriage and divorce rates will be used to estimate intensity of nuptiality (marriage) and intensity of divorce. (Koschin, 1992)

For each age x = 15, 16, …59 for women in the Czech Republic, intensities of transition-probability in the form of matrix $\mathbf{H}_x$ is prepared for states single, married, divorced and widowed. Here, σ denotes nuptiality (marriage rate) with

$$H_x = \begin{pmatrix} \sigma_{Sx} + \mu_{Fx} & 0 & 0 & 0 \\ -\sigma_{Sx} & \rho_x + \mu_{Fx} + \mu_{Mx} & -\sigma_{Dx} & -\sigma_{Wx} \\ 0 & -\rho_x & \sigma_{Dx} + \mu_{Fx} & 0 \\ 0 & -\mu_{Mx} & 0 & \sigma_{Wx} + \mu_{Fx} \end{pmatrix}$$

the index according to marital status (S-single, D-divorced, W-widowed), ρ is divorce rate, μ denotes mortality of females (F) or males-husbands (M).

Then, transition-probability matrices $\mathbf{p}_x$ are calculated and, subsequently, other life tables indicators in the form of matrices, such as table number of survivors (matrices $\mathbf{l}_x$), table number of person-years (matrices $\mathbf{L}_x$), number of remaining years of life to be lived by the table generation (for entire group of individuals) in the age of *x* (matrices $\mathbf{T}_x$) and matrices of expected length of stay ($\mathbf{e}_x$; equivalent of life expectancy). (Koschin, 1992; Land & Rogers, 1982; Rogers, 1975; Raymer & Willekens, 2008)

4. Results

In the following figures, current age of the woman (15 to 59) is displayed on the X-axis and number of years spent by woman (currently being in status A) in the status B till the age of 59 is presented on the Y-axis. This indicator is called *expected length of stay*; it was calculated for years 2002–2010 for ages x = 15, 16, …, 59 years; in addition, abridged calculation of the years 2001 and 1990 are available for women in the Czech Republic. In some cases initial ages could be influenced by low number of events (marriages, divorces, …) and should be subject of further smoothing.

In the year 2010:

Typical results of life tables calculation could be interpreted as follows: Single woman in the age of 25 years may expect that she spends till the age of 59 years another 18.3 years as single, 12.5 years as married, 3.4 years as divorced and 0.3 years as widowed. Married woman in the age of 25 years may expect that she spends till the age of 59 years another 25.8 years as married, 8.2 years as divorced and 0.6 years as widowed. Divorced woman in the age of 25 years may expect that she spends till the age of 59 years another 14.1 years as married, 20.1 years as divorced and 0.4 years as widowed. Widowed woman in the age of 25 years may expect that she spends till the age of 59 years another 7.3 years as married, 1.8 years as divorced and 25.4 years as widowed.

Comparison 2010 and 2002:

Majority of woman in young ages are single, their transition into the state ‘married’ is visible between the age 25 and 30 years. It can be seen that years 2002 and 2010 differs in the number and proportion of women that remain in the state ‘single’ after 35 years of age. In 2002 the proportion of those women formed approximately 30 % and slowly decreased, whereas in 2010 the proportion is 40 % of all women from the studied group.

Another trend can be seen from results for women single or married. Their total proportion increased between years 2002 and 2010 by 5 percentage points. In the age of 40 years there were 82.9 % of single or married women, whereas in 2010 this proportion increased to 87.3 %. It shows that the state ‘single’ is in some cases preferred also by women who were divorced or widowed in 2002.

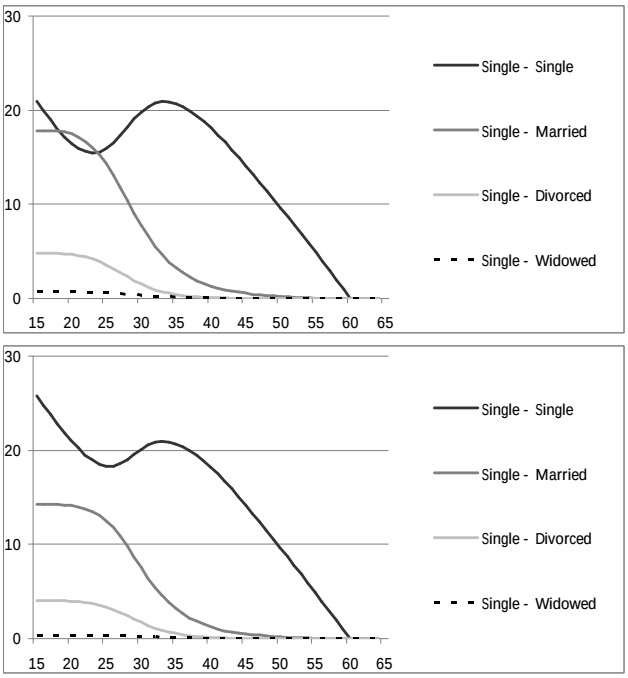


Figure 1: Expected length of stay in states for originally single women, women, 2002 and 2010, Czech Republic

Trends and changes in women's behaviour over 10 years: originally single women

It is interesting that young single women tend to marry and escape from the state 'single', but for women in ages 25 to 33 years number of expected years when they remain in the state 'single' even grows. For example in 2010, single woman in the age of 33 years has very low probability 0.044 to leave state 'single' and high probability 0.956 to remain in this state further. For older single women the probability of transition into other states (first state 'married', then possibly 'divorced' or 'widowed') is very low, expected length of stay in the state 'single' decreases proportionally with the age.

Trend in the period of 2001–2010 shows that in the age of 15 to 28 years single women tend to postpone marriage and stay single. The expected length of stay in the state 'single' prolongs and difference between 2001 and 2010 is almost five years for the age of 25 years. The common characteristic is that all lines have the same shape, i.e. decrease of expected length of stay between 15 and 25 years and then approximately between 25 and 35 years increasing chance (measured both in probability and number of expected years) that woman remains single. The largest difference is visible in first 15 years of studied part of women's lives with one exception – year 2003 differs from others in the ages of 30–45 years. This could be explained by legislative impact.

Year 1991 (Koschin, 1992) shows the remarkable change that happened over last 20 years in the Czech Republic. Young single women till their 25 years could expect to remain single for less than 10 years whereas in 2010 women till 25 years might expect to remain single another 18 to 25 years of their lives. This represents more than double number of years. On the other hand, comparable results belong to ages 32 years and more. If a woman remains single till her 32 years than there is almost no difference over last 20 years in the indicator how long she might expect to remain such.

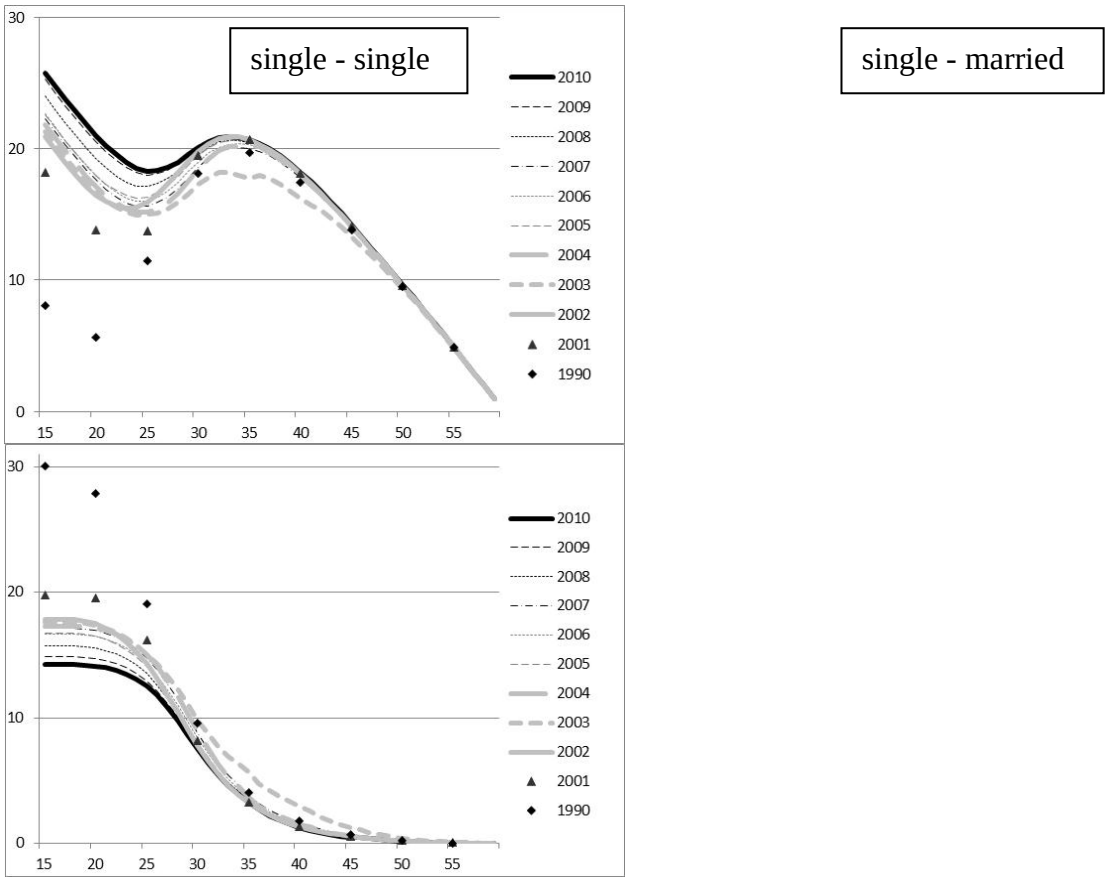


Figure 2: Expected length of stay for originally single women, 1990, 2001–2010, Czech Republic

Trends and changes in women's behaviour over 10 years: originally married women

Married women stay in the status 'married' for the same time over the last 10 years, length of marriages did not change very much. The visible difference is between the year 1991 and 2001-2010, namely for women in the age till 35 years. Transition into the state 'divorced' is more probable now compared to 1991 and length of stay as divorced woman prolonged by approximately 4 years for young women.

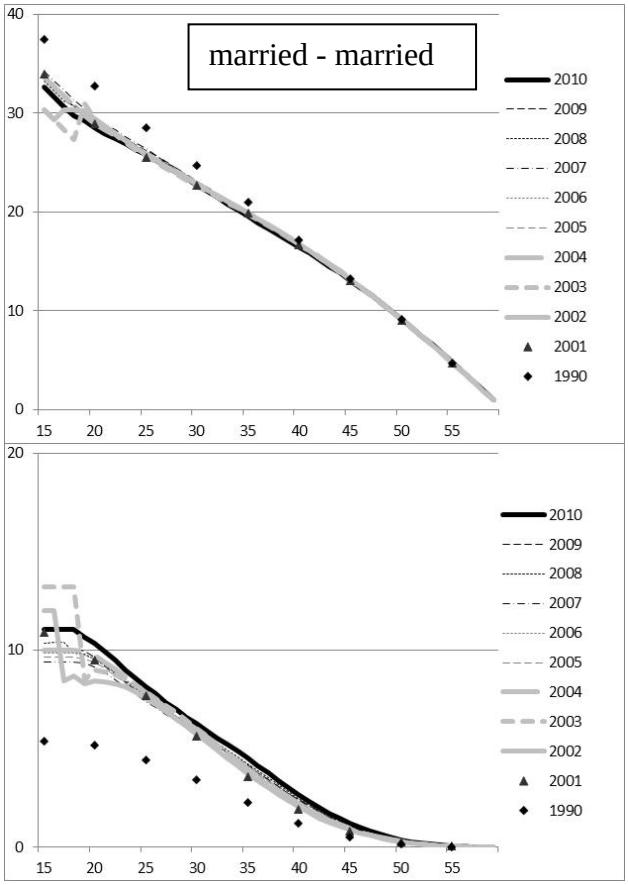


Figure 3: Expected length of stay for originally married women, 1990, 2001–2010, Czech Republic

Trends and changes in women's behaviour over 10 years: originally divorced women

Divorced young women stay in following marriages for shorter time compared to 1991 and even to 2001. For example, 30-years old divorced woman spends in following marriages 8 years till the age of 59 years. In higher ages probability of further marriage decreases quickly but it did not change very much over the last 10 years.

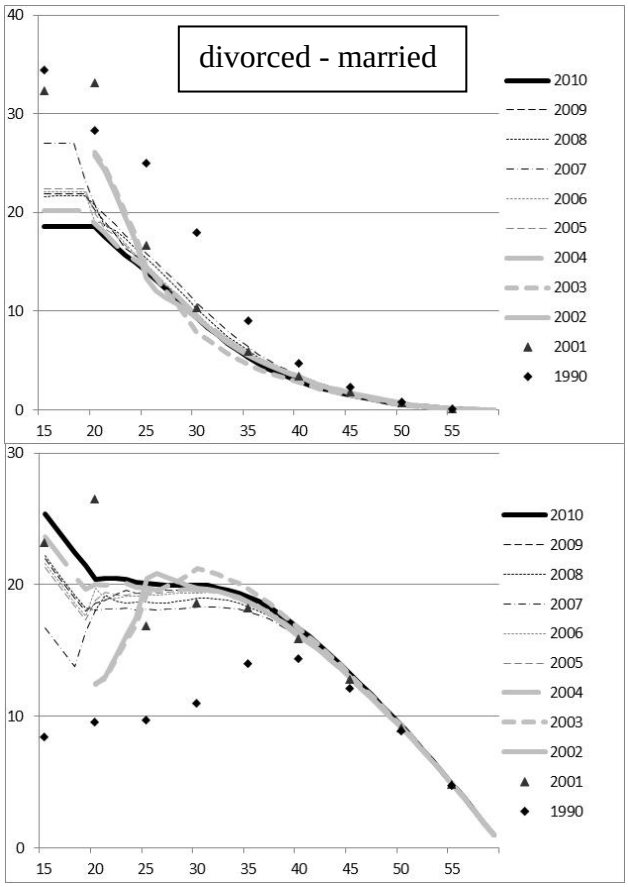


Figure 4: Expected length of stay for originally divorced women, 1990, 2001–2010, Czech Republic

Trends and changes in women's behaviour over 10 years: originally widowed women

The probability of further marriage decreases very quickly, especially after the age of 25. The length of following marriages did not change very much over the last 20 years. Behaviour in case of widowed women remained stable in studied history.

5. Conclusion

The article introduced application of multistate demographic methods onto the 'marriage career' based on real data. The analysis presented alternative approach to the analysis of marriages and divorces and additional utilization of life tables' methodology.

Calculation of multistate life table and modelling 'marriage career' of women in the Czech Republic 2001–2010 showed that women change their decision toward marriages. Objectives of the article was verified for younger women: women 15–30 years old in the Czech Republic changed their behaviour related to their marriage over last 10 20 years: tendency not to marry is stronger among young women, they stay unmarried and probably live in partnerships without official marriage. Women over 30 years changed their behaviour only little.

In this sense changes in behaviour of women in the Czech Republic related to their marriage decision over last 10 / 20 years are confirmed. The most verifiable changes happened to young women and are visible in comparison of 1991 versus 2010.

References

[1] KOSCHIN, F.: *Vícestavová demografie*. Prague, University of Economics in Prague, 1992. ISBN 80-7079-087-3.
[2] LAND, K. C., ROGERS, A. (ed.): *Multidimensional Mathematical Demography*. New York, Academic Press, 1982.
[3] RAYMER, J., WILLEKENS, F. (ed.): *International Migration in Europe. Data, Models and Estimates.* New York, John Wiley and Sons, 2008.
[4] ROGERS, A (ed.): Essays in Multistate Mathematical Demography (reprint from *Environment and Planning* A 12, 1980, s. 485–622). Laxenburg, IIASA, 1980.
[5] ROGERS, A.: *Introduction to Multiregional Mathematical Demography*. New York, John Wiley and Sons, 1975.

Addresses

Mgr. Ing. Martina Miskolczi, MBA
martina.miskolczi@vse.cz

doc. Ing. Jitka, Langhamrová, CSc.
langhamj@vse.cz

Bc. Jana Langhamrová
xlanj18@vse.cz

Department of Demography
Faculty of Informatics and Statistics
University of Economics in Prague
nám. W. Churchilla 4
130 67 Prague 3
Czech Republic

Supported by research project IGA F4/29/2011 Analysis of population ageing and impact on labour market and economic activity

Vplyv ekonomickej recesie na regionálne rozdiely nezamestnanosti v Českej republike.[1]
Impact of economic recession on regional differences in unemployment in the Czech Republic.

Tomáš Pavelka

Abstract: A Czech economy, like other states of the European Union, has undergone economic cycle in the past years. The development of gross domestic product had an impact on the development of registered unemployment rate. The Czech economy shows significant regional differences in the unemployment rate. Significant regional differences in unemployment rates indicate low mobility and flexibility of the labour market. The article deals with the impact of the economic cycle on regional differences in unemployment rates in the individual districts of the Czech Republic.

Abstrakt: Česká ekonomika si podobne ako ostatné štáty Európskej únie prešla v minulých rokoch ekonomickým cyklom. Vývoj hrubého domáceho produktu mal vplyv aj na vývoj registrované miery nezamestnanosti. Česká ekonomika vykazuje značné regionálne rozdiely v miere nezamestnanosti. Výrazné regionálne rozdiely v miere nezamestnanosti naznačujú nízku mobilitu a flexibilitu pracovného trhu. Článok sa venuje vplyvu ekonomického cyklu na regionálne rozdiely v miere nezamestnanosti v jednotlivých okresoch Českej republiky.

Key words: Unemployment rate, Gross domestic product, Economic cyklus, Flexibility of labour market.

Kľúčové slová: miera nezamestnanosti, hrubý domáci produkt, ekonomický cyklus, flexibilita trhu práce.

JEL classification: J 60; E 32.

Úvod

V české ekonomice lze v posledních letech pozorovat učebnicovou podobu hospodářského cyklu. Po vstupu České republiky do Evropské unie se růst reálného produktu rychle zvyšoval a dosahoval nejvyššího tempa v rámci členských států Evropské unie. Poté však následovala finanční krize a potažmo její přesun do reálné ekonomiky, která se projevila ve Spojených státech a v některých evropských ekonomikách v roce 2007 a 2008. Bylo pouze otázkou času, kdy se tato krize projeví i v české ekonomice. Česká republika je malou otevřenou ekonomikou, která není schopna izolovat negativní dopady světového hospodářství na vlastní ekonomiku. Jisté zpomalení růstu bylo patrné již v roce 2008, ale plnou silou dopadla na českou ekonomiku recese v roce 2009. Následné mírné oživení se ukázalo jako neudržitelné a česká ekonomika se v letošním roce opět propadá.

Na výkyvy reálného hrubého domácího produktu samozřejmě reaguje, i když s určitým zpožděním, i trh práce. Silný ekonomický růst byl spojen s rapidním poklesem míry nezaměstnanosti, propad ekonomiky naopak s velmi rychlým nárůstem počtem nezaměstnaných. Jak bude ukázáno níže, míra nezaměstnanosti v České republice vykazuje značné regionální rozdíly. Přetrvávající regionální rozdíly v míře nezaměstnanosti lze vysvětlovat existencí strukturální nezaměstnanosti a potažmo i nízkou mobilitou, či jinými slovy, nízkou flexibilitou českého trhu práce. Příspěvek si klade za cíl analyzovat dopad nedávné ekonomické recese na regionální rozdíly míry nezaměstnanosti v České republice.

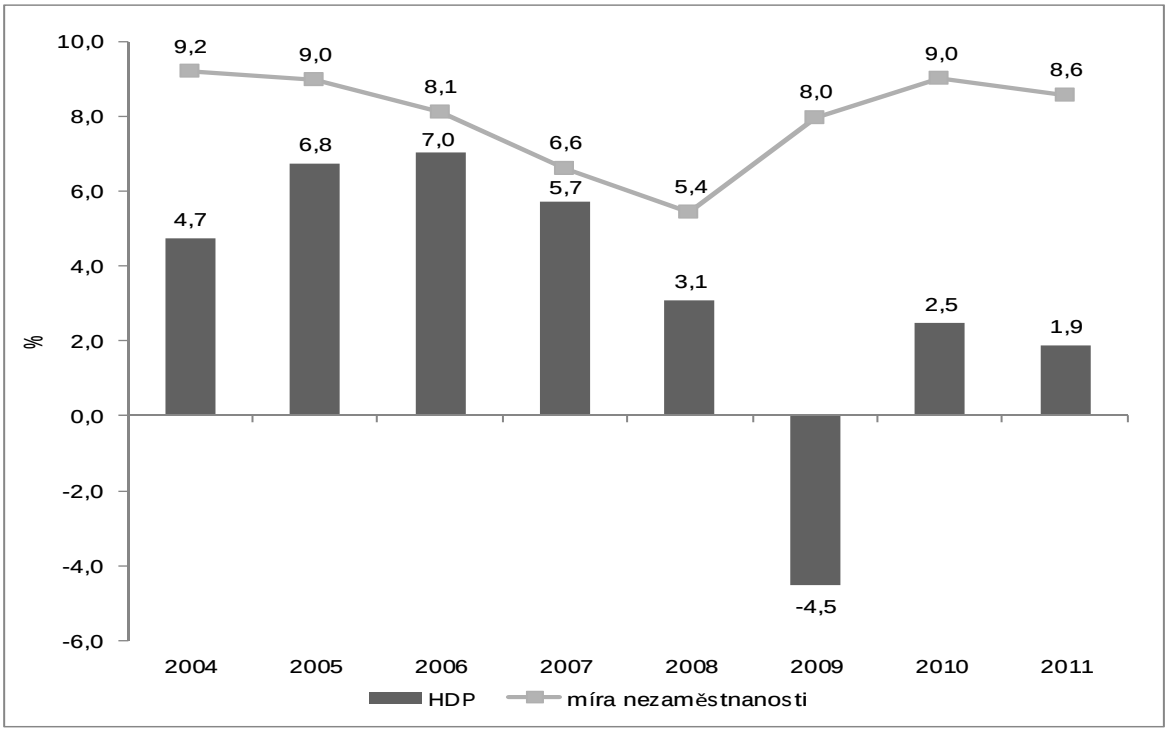
[1] Tento článek je jedním z výstupů výzkumného projektu "Flexibilita trhu práce České republiky" registrovaného Interní grantovou agenturou Vysoké školy ekonomické v Praze pod číslem MF/19/2012.

Bude řešena otázka, zda se vlivem recese regionální rozdíly zvyšují, a zda lze potvrdit názor, že míra nezaměstnanosti vlivem recese rostla v závislosti na její výši v předkrizovém období. Nezaměstnanost obecně, a zejména její cyklická složka, je spojena s řadou negativních dopadů. Ekonomické dopady cyklické nezaměstnanosti lze kvantifikovat pomocí Okunova zákona. Detailní odhad nákladů nezaměstnanosti v České republice lze nalézt ve článku „Odhad nákladů nezaměstnanosti z pohledu veřejných rozpočtů“ (Čadil a kol. 2011).

V příspěvku budou využita data o výši registrované míry nezaměstnanosti v jednotlivých okresech České republiky, tak jak je zveřejňuje Ministerstvo práce a sociálních věcí České republiky na svém internetovém portálu.

1. Ekonomický cyklus a míra nezaměstnanosti

Obrázek č. 1 zachycuje roční data o vývoji registrované míry nezaměstnanosti a o data o meziročních změnách reálného hrubého domácího produktu v České republice.



Obr. 1: Míra nezaměstnanosti a meziroční změna reálného hrubého domácího produktu.

Pramen: ČSÚ (datum citace: 18. 11. 2012)

Reálný hrubý domácí produkt České republiky v období 2004 – 2007 rostl v průměru meziročně o vysokých 6 %. Pozitivně na růst působil vstup České republiky do Evropské unie v roce 2004, díky kterému se českým výrobcům ještě více otevřely trhy ostatních členských států Evropské unie. Pozitivně se však projevil i příchod zahraničních investorů, který byl podpořen mimo jiné i investičními pobídkami v předcházejících letech. Na český vývoz, který se postupně stal jedním z hlavních zdrojů růstu hrubého domácího produktu, měla pozitivní vliv i hospodářská situace v ostatních členských státech Evropské unie. A v neposlední řadě, nelze opominout i domácí část národohospodářské poptávky. Domácnosti pod vlivem rostoucích reálných příjmů a optimistických očekávání zvyšovaly svou spotřebu a hrubá tvorba kapitálu v letech 2006 – 2007 vykazovala dvojciferná tempa růstu. V roce 2008 se však v české ekonomice začaly projevovat první příznaky ekonomické recese. V roce 2008 ještě došlo pouze ke zpomalení ekonomické růstu, ale v roce 2009 již reálný hrubý domácí

produkt propadl o 4,5 %. Tento propad byl zapříčiněn meziročním poklesem hrubé tvorby kapitálu a exportu. Spotřeba domácností stagnovala a růst vládní spotřeby nestačil na zvrat negativního vývoje. V následujících dvou letech sice reálný hrubý domácí produkt mírně rostl, ale jak naznačují předběžná data za letošní rok, česká ekonomika se opět vrací do recese.

Ekonomický cyklus měl dopad i na vývoj míry nezaměstnanosti. Míra nezaměstnanosti reaguje na vývoj produktu zpravidla s určitým zpožděním. Obrázek č. 1 však zachycuje roční průměrné hodnoty registrované míry nezaměstnanosti a meziroční změny reálného hrubého domácího produktu, díky čemuž není toto časové zpoždění patrné. Z Obrázku č. 1 je zřejmé, že registrovaná míra nezaměstnanosti klesla v období 2004 – 2008 o 3,8 p. b. K posledním dni roku 2004 evidovaly české úřady práce 541 762 nezaměstnaných a k poslednímu dni roku 2008 to bylo o 189 512 nezaměstnaných méně (celkově bylo ke konci roku 2008 nezaměstnáno 353 250 osob). Z důvodu propadu reálného hrubého domácího produktu míra nezaměstnanosti v roce 2009 meziročně vzrostla o 2,6 p. b. na 8,0 %. Růst míry nezaměstnanosti pokračoval i přes mírný růst produktu i v roce 2010. V roce 2010 činila průměrná míra nezaměstnanosti v České republice 9,0 %, ke konci tohoto roku české úřady práce evidovaly 561 551 nezaměstnaných, což představovalo oproti roku 2008 nárůst o 209 301 osob. V loňském roce registrovaná míra nezaměstnanosti v České republice klesla meziročně o 0,4 p. b. na 8,6 %.

2. Regionální rozdíly v míře nezaměstnanosti

Registrovaná míra nezaměstnanosti v České republice vykazuje značné regionální rozdíly. Regionální rozdíly v míře nezaměstnanosti lze analyzovat podle jednotlivých krajů či podle jednotlivých okresů České republiky. Tento příspěvek analyzuje regionální rozdíly míry nezaměstnanosti v rámci 77 okresů České republiky. Regionálními rozdíly podle krajů se zabývají dva články Löstera (2011a, 2011b).

Tab. 1: Regionální míra nezaměstnanosti

	2004	2005	2006	2007	2008	2009	2010	2011
Praha-východ	3,5	3,1	2,3	2,0	1,7	2,7	3,8	3,6
Praha-západ	2,8	2,7	2,4	1,8	1,7	3,1	4,2	4,0
Praha	3,6	3,4	3,0	2,5	2,1	3,0	3,9	4,0
Ml. Boleslav	4,6	3,9	3,4	2,6	2,2	4,0	4,8	4,8
Benešov	4,3	4,5	4,1	3,4	2,8	4,3	5,4	5,4
Hodonín	13,7	14,3	13,8	11,3	9,8	13,4	14,9	14,1
Děčín	14,2	15,0	14,1	11,7	10,2	13,4	15,1	14,3
Bruntál	16,2	16,0	14,1	11,4	9,7	13,7	15,6	15,6
Jeseník	15,0	16,5	14,9	11,6	9,6	13,2	15,6	15,8
Most	22,8	22,0	20,5	17,6	13,1	15,5	16,2	15,9

Pramen: MPSV, http://portal.mpsv.cz/sz/stat/nz, (datum citace: 18. 11. 2012)

Tabulka č. 1 zachycuje pět okresů v České republice s nejnižší registrovanou mírou nezaměstnanosti a pět okresů s nejvyšší registrovanou mírou nezaměstnanosti v období 2004 – 2011. Uvedené okresy jsou setříděné podle situace v roce 2011. Nejnižší míru registrované nezaměstnanosti vykazují dva okresy sousedící s hlavním městem, Praha - východ a Praha -

západ, a také samotné hlavní město Praha. Velká část ekonomicky aktivních osob ze sousedních okresů hlavního města dojíždí za zaměstnáním právě do Prahy. Stejné vysvětlení platí také pro Benešov a částečně i Mladou Boleslav, kde však klíčovou roli hraje automobilka Škoda. Pouze v prvních dvou letech se do pětice okresů s nejnižší mírou nezaměstnanosti dostaly České Budějovice a v roce 2007 Pelhřimov. Naopak nejvyšší míru registrované nezaměstnanosti ve všech sledovaných letech vykazoval okres Most. S velkou pravděpodobností se jedná o strukturální nezaměstnanost, která je spojena s nevhodnou strukturou hospodářství pocházející ještě z dob před rokem 1989. Mezi okresy s nejvyšší výší registrované nezaměstnanosti se zařazuje dlouhodobě i další Severočeský okres, a to Děčín a dále pak okresy na severu Moravy - Bruntál, Jičín a často i Karviná. Vedle nevhodné struktury hospodářství svou roli hraje i nedostatečná dopravní infrastruktura a zejména nedostatečné dopravní spojení těchto okresů se zbytkem České republiky. Dlouhodobě vysokou registrovanou míru nezaměstnanosti vykazují i dva zemědělské okresy z jihu Moravy, a to Hodonín a Znojmo.

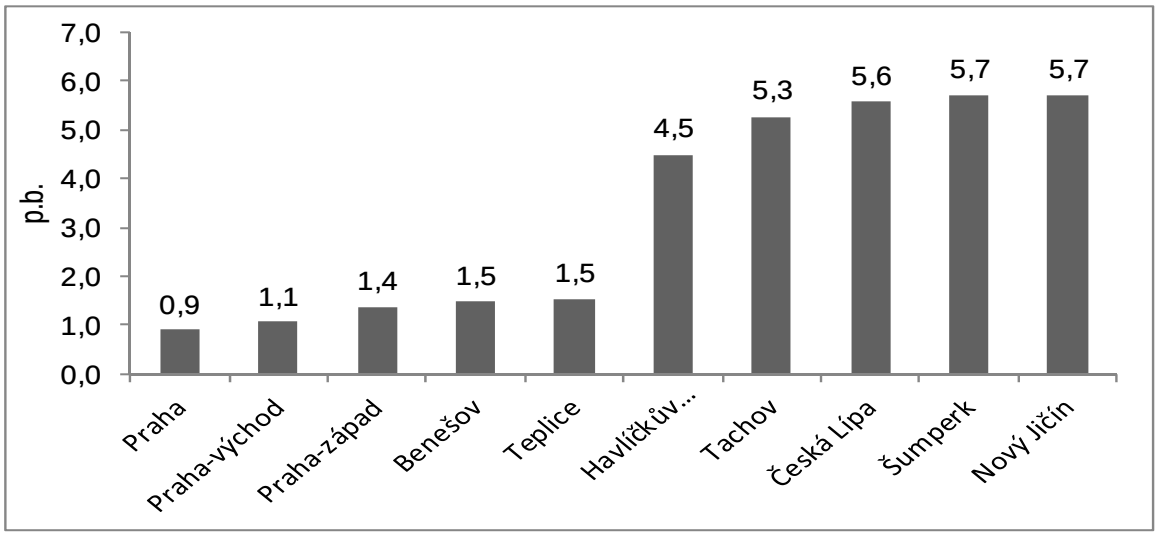
Z tabulky č. 2 je zřejmé, že pozitivní ekonomický vývoj v průběhu let 2004 – 2008 byl spojen nejen s poklesem celorepublikové registrované míry nezaměstnanosti (pokles o 3,8 p.b.), ale zároveň klesala i maximální a minimální míra nezaměstnanosti podle okresů. V letech 2004 – 2008 se snížila míra nezaměstnanosti v okrese s nejnižší mírou nezaměstnanosti o 1,1 p.b. (o 39 %) a míra nezaměstnanosti v okrese s nejvyšší mírou nezaměstnanosti o 9,7 p.b. (o 42,5 %). Procentně tak došlo v podstatě k obdobnému poklesu registrované míry nezaměstnanosti v obou okresech. Rozdíl mezi mírou nezaměstnanosti v okrese s nejvyšší mírou nezaměstnanosti a mírou nezaměstnanosti v okrese s nejnižší mírou nezaměstnanosti se z 20 p.b. v roce 2004 snížil na 11,4 p.b. v roce 2008. Jak je patrné také z tabulky č. 2, mezi roky 2004 – 2008 postupně klesala i směrodatná odchylka z 3,86 na 2,38. Odchylky míry nezaměstnanosti v jednotlivých okresech České republiky od její průměrné výše se vlivem ekonomického růstu postupně zmenšovaly.

Tab. 2: Míra nezaměstnanosti v okresech ČR – variabilita dat

	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Min. míra nezaměstnanosti	2,8	2,7	2,3	1,8	1,7	2,7	3,8	3,6
Max. míra nezaměstnanosti	22,8	22	20,5	17,6	13,1	15,5	16,2	15,9
Variační rozpětí	20	19,3	18,2	15,8	11,4	12,8	12,4	12,3
Směrodatná odchylka	3,86	3,75	3,59	3,04	2,38	2,74	2,77	2,75

Pramen: MPSV, http://portal.mpsv.cz/sz/stat/nz, vlastní výpočty, (datum citace: 18. 11. 2012)

V roce 2009 nastal z důvodu ekonomické recese zlom i ve vývoji registrované míry nezaměstnanosti. V roce 2009 došlo k meziročnímu nárůstu registrované míry nezaměstnanosti ve všech 77 okresech České republiky. Pět okresů s nejnižším a pět okresů s nejvyšším meziročním přírůstkem registrované míry nezaměstnanosti v roce 2009 zachycuje obrázek č. 2. V roce 2010 došlo k dalšímu prohloubení registrované míry nezaměstnanosti v 74 okresech České republiky. V okrese Česká lípa registrovaná míra nezaměstnanosti v roce 2010 meziročně klesla o 0,1 p.b. a v okresech Šumperk a Rokycany se registrovaná míra nezaměstnanosti meziročně nezměnila. V loňském roce se registrovaná míra nezaměstnanosti meziročně zvýšila v 9 okresech, v 5 okresech se nezměnila a ve zbývajících 63 okresech pokračovala v růstu.



Obr. 2: Meziroční změny registrované míry nezaměstnanosti v roce 2009

Pramen: vlastní výpočty

Z tabulky č. 2 je zřejmé, že od roku 2009 se zvýšila registrovaná míra nezaměstnanosti jak v okrese s nejnižší mírou nezaměstnanosti, tak v okrese s nejvyšší mírou nezaměstnanosti. V okrese s nejnižší mírou nezaměstnanosti se registrovaná míra nezaměstnanosti mezi roky 2008 a 2011 zvýšila o 1,9 p. b. (o 111,8 %) a v okrese s nejvyšší mírou nezaměstnanosti o 2,8 p.b. (pouze o 21,4 %). Z uvedeného se zdá, že ekonomická recese se projevila v růstu míry nezaměstnanosti výrazněji v okrese (či okresech), který vykazoval dlouhodobě nízkou míru nezaměstnanosti. Variační rozpětí se zvýšilo z 11,4 v roce 2008 na 12,8 v roce 2009. V následujících dvou letech se však rozdíl míry nezaměstnanosti mezi okresem s nejvyšší a nejnižší mírou nezaměstnanosti snižoval. Také hodnoty směrodatné odchylky naznačují mírnější zvýšení odchylek míry nezaměstnanosti od její průměrné míry v porovnání s počátkem sledovaného období.

3. Záver

Ekonomický cyklus, kterému byla v posledních letech vystavena česká ekonomika, měl dopad i na regionální rozdíly míry nezaměstnanosti. Rychlý růst reálného hrubého domácího produktu vedl k poklesu míry nezaměstnanosti, k poklesu rozdílu mezi nejvyšší a nejnižší mírou nezaměstnanosti podle okresů a snížila se také odchylka míry nezaměstnanosti v jednotlivých okresech od její průměrné výše.

Ekonomický propad v roce 2009 způsobil růst registrované míry nezaměstnanosti ve všech 77 okresech České republiky. Míra nezaměstnanosti výrazněji vzrostla v okresech s nejnižší mírou nezaměstnanosti v porovnání s růstem míry nezaměstnanosti v okresech s nejvyšší mírou nezaměstnanosti.

Zjednodušeně lze uvést, že nejnižší míru nezaměstnanosti vykazuje dlouhodobě hlavní město Praha a okresy v jeho blízkosti. Naopak nejvyšší míru nezaměstnanosti dlouhodobě vykazují některé okresy na severu Čech a na Severu Moravy.

Literatura

[1]ČADIL, J., PAVELKA, T., KAŇKOVÁ, E., VORLÍČEK, J. 2011.Odhad nákladů nezaměstnanosti z pohledu veřejných rozpočtů. *Politická ekonomie*, , roč. 59, č. 5, s. 618–637. ISSN 0032-3233

[2] LÖSTER, T., LANGHAMROVÁ, J. 2011a. Analysis of differences in unemployment rate between regions of the Czech Republic. In: *2nd International Sientific Conference Whither our Economies – 2012. Conference Proceedings.* Vilnius: Mykolas Romeris University, ISSN 2029-8501.

[3]LÖSTER, T., LANGHAMROVÁ, J. 2011b. Disparities between regions of the Czech Republic for non business aspects of labour market. In: PAVELKA, Tomáš, LÖSTER, Tomáš (ed.). *International Days of Statistics and Economics.* Slaný: Melandrium, ISBN 978-80-86175-79-9.

Adresa autora:

Tomáš Pavelka, Ing. Ph.D.
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
pavelkat@vse.cz

Popis tvarovej variability synaptonemálneho komplexu s použitím algoritmu neurónového plynu

Description of shape variability of the synaptonemal complex using the neural gas algorithm

Lukáš Pastorek, Hana Řezanková

Abstract: This paper is concerned with the analysis of image obtained from the fluorescence microscope, and our aim is to describe the shape of the synaptonemal complex with the modified neural gas algorithm so that algorithm is applicable on similar images with various shape of the synaptonemal complex.

Abstrakt: V práci sa budeme zaoberať analýzou obrazu získaného z flourescenčného mikroskopu, pričom našou snahou bude popísať tvarovú variabilitu synaptonemálneho komplexu upraveným algoritmom neurónového plynu, tak aby bol algoritmus aplikovateľný pri analýze podobných obrázkov s rôznym tvarom synaptonemálneho komplexu.

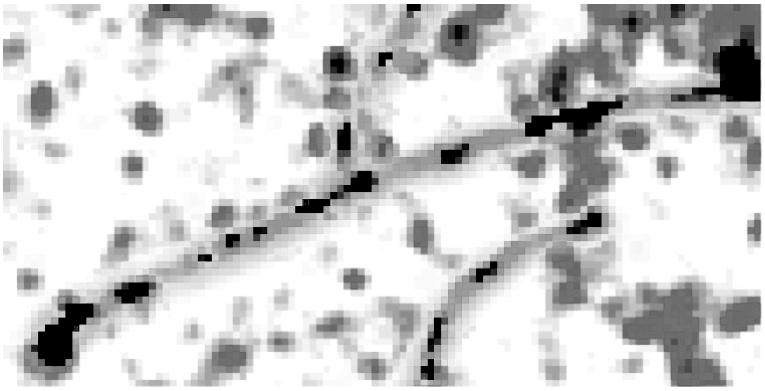
Key words: Neural gas, shape variability, synaptonemal complex

Kľúčové slová: Neurónový plyn, tvarová variabilita, synaptonemálny komplex

JEL classification: Z19

Úvod

Tento príspevok nadväzuje na obsiahlejšiu prácu, ktorej primárnym cieľom je objaviť hlbšie súvislosti proteínu (pre potreby tohto článku označovaného ako NEMO1) a tzv. synaptonemálneho komplexu. Primárnym cieľom je zistiť, či lokalizácie proteínu NEMO1 v okolí synaptonemálneho komplexu vykazujú náhodný alebo systematický charakter (viď obrázok 1). Bude nutné vytvoriť metódu, ktorá dokáže zamietnuť alebo nezamietnuť hypotézu o náhodnom rozložení lokalizácií proteínu v blízkosti synaptonemálneho komplexu. Problém, ktorý vyvstal pri riešení tejto úlohy, spočíva v rôznorodosti tvarovej variability komplexu. Bolo teda potrebné objaviť spôsob, akým popíšeme jeho tvar, aby sme dokázali následne popísať vzťah medzi lokalizáciami a komplexom. Túto úlohu sme sa rozhodli riešiť s využitím biologicky inšpirovaného algoritmu neurónového plynu.



Obr. 1: Ukážka obrazu z mikroskopu zobrazujúci lokalizácie proteínu NEMO1 a synaptonemálneho komplexu (pre potreby článku upravený do čiernobielych odtieňov).

Pozn.: Výrazná čiara smerujúcu z ľavého dolného rohu k pravému hornému rohu predstavuje celý synaptonemálny komplex, zatiaľ čo malé útvary v jeho okolí sú lokalizácie proteínu NEMO1.

1. Popis dátového súboru

Dátový súbor je bežný obrázok v štandardizovanom formáte JPEG s príponou *jpg*. Tento farebný obrázok je tvorený preložením troch vrstiev hodnôt RGB, teda červeného, zeleného a modrého svetla. Celý obrázok je tvorený z malých stavebných prvkov – pixelov, pričom každý pixel je charakterizovaný tromi hodnotami spomínanej červenej, zelenej a modrej. Prekryv hodnôt (ich kombinácia) vyvolá výslednú farbu pixelu. Hodnoty jednotlivých zložiek RGB sa pohybujú medzi 0 a 255. Každý pixel má presne zadefinovanú pozíciu na osi x a y. Preto sme úpravou v prostredí Matlab dokázali pretransformovať celý súbor obrázku na klasický štatistický súbor (v našom prípade 5-rozmerný) obsahujúci v prvých dvoch stĺpcoch hodnoty priestorových súradníc pixelu na osy x a y a zvyšné tri stĺpce obsahujú hodnoty červenej, zelenej a modrej.

2. Popis metódy

Algoritmus *neurónového plynu* (angl. *Neural Gas;* popísaný v [1]), spadá do oblasti metód fyzikálne inšpirovaného strojového učenia s učením bez učiteľa, kedy sa model učí len na základe predložených dátových vstupov, bez možnosti upravovať svoje váhy s ohľadom na kvalitu výstupov.

Počas *sekvenčného* tréningu vyberáme v každom iteračnom kroku náhodne jeden zo vstupných vektorov $\mathbf{x}_i$ z dátového súboru $X = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{R}^d, i = \{1, \mathbf{K}, n\}\}$, kde n je dĺžka trénovacieho súboru a $\mathbf{R}^d$ je d-rozmerný vektorový priestor. Tento vstupný vektor následne predložíme populácií váhových vektorov modelu $C = \{\mathbf{c}_j \mid \mathbf{c}_j \in \mathbf{R}^d, j = \{1, \mathbf{K}, l\}\}$, kde c_j je váhový vektor modelu, ktorý označujeme ako *neurón j-teho zhluku* a k je zvolený *počet* zhlukov, ku ktorým chceme dáta priradiť. *Neuróny* súťažia o najbližšiu pozíciu k predloženému vstupnému vektoru z hľadiska euklidovskej vzdialenosti

$$\|\mathbf{x} - \mathbf{c}_v\| = \arg\min_j \|\mathbf{x}_i - \mathbf{c}_j\|, \tag{1}$$

kde $\mathbf{c}_v$ je *víťazný neurón* (referenčný váhový vektor), ktorý reprezentuje zhluk, do ktorého bude vstupný vektor prvotne patriť. Učiaci algoritmus *neurónového plynu* však obsahuje navyše zoznam tzv. *„poradie susedov*“ (angl. ”neighborhood ranking”), kedy sú všetky váhové vektory modelu radené podľa euklidovskej vzdialenosti od vstupného vektora $\mathbf{x}_i$, $(\mathbf{c}_{j_0}, \mathbf{c}_{j_1}, \mathbf{K}, \mathbf{c}_{j_{l-1}})$, kde $\mathbf{c}_{j_0}$ je najbližší váhový vektor k vstupnému vektoru $\mathbf{x}_i$, $\mathbf{c}_{j_1}$ je druhý najbližší váhový vektor a $\mathbf{c}_{j_m}$, $m = 0, \mathbf{K}, l-1$ je váhový vektor modelu, pre ktorý existuje m vektorov $\mathbf{c}_j$ takých, že

$$\|\mathbf{x} - \mathbf{c}_j\| < \|\mathbf{x} - \mathbf{c}_{j_m}\|. \tag{2}$$

Index m, ktorý je spojený s každým vektorom $\mathbf{c}_j$, je určený funkciou $m(\mathbf{x}_i, \mathbf{c}_j)$ závislej na $\mathbf{x}_i$ a $C = \{\mathbf{c}_j \mid \mathbf{c}_j \in \mathbf{R}^d, j = \{1, \mathbf{K}, l\}\}$. Adaptačný krok je vyjadrený vzorcom

$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t) + a(t) h_l\left(m(\mathbf{x}_i, \mathbf{c}_j)\right)\left[\mathbf{x}_i(t) - \mathbf{c}_j(t)\right], \tag{3}$$

kde $a(t)$ je rýchlosť učenia a $h_l\left(m(\mathbf{x}_i, \mathbf{c}_j)\right) = e^{-m/l(t)}$ je exponenciálna forma funkcie okolia. Parameter $l(t)$ určuje počet váhových vektorov, ktoré upravujú svoje váhy (polohu) a je definovaný diskrétnou klesajúcou funkciou: $l(t) = \left(l_f / l_i\right)^{t/T}$.

3. Úprava dátového súboru a metódy

Metódu algoritmu *neurónového plynu* sme sa rozhodli upraviť s ohľadom na špecifickosť dát a zámer, s ktorým danú metódu uplatňujeme. Našim primárnym cieľom nie je rozdeliť dátový súbor do zhlukov, ale skôr popísanie priestoru synaptonemálneho komplexu a nájdenie prirodzených pozícii referenčných modelových vektorov na skúmanom obrázku, ktoré vystihujú približne krivku pohybu a zakrivenie komplexu.

Keďže komplex je na originálnom obrázku charakterizovaný len zelenou farbou, zatiaľ čo lokalizácie proteínu NEMO1 len červenou, môžeme pre potreby metódy neu*rónového plynu* redukovať priestor dátového súboru len na trojrozmerný súbor $\mathbf{R}^3$ a s dimenziu hovoriacou o lokalizáciách proteínu zatiaľ vôbec nepracovať (ako následok nám zmiznú proteínové lokalizácie z obrázku a na obraze zostane len synaptonemálny komplex; viď obrázok 2). Modrá farba sa v našom obrázku nevyskytuje, preto vylúčime i túto dimenziu. Zároveň sme z obrázku odfiltrovali malé časti iných synaptonemálnych komplexov, ktoré do skúmaného obrázku zasahovali.

Obr. 2: Ukážka „vyčisteného“ obrázku zobrazujúceho synaptonemálny komplex

Metódu *neurónového plynu* sme sa rozhodli použiť i z dôvodu, že tvar synaptonemálneho komplexu málo kedy vykazuje lineárny priebeh (ako je na obrázku 2), preto ho väčšinou nie je možné popísať bežne používanými regresnými modelmi.

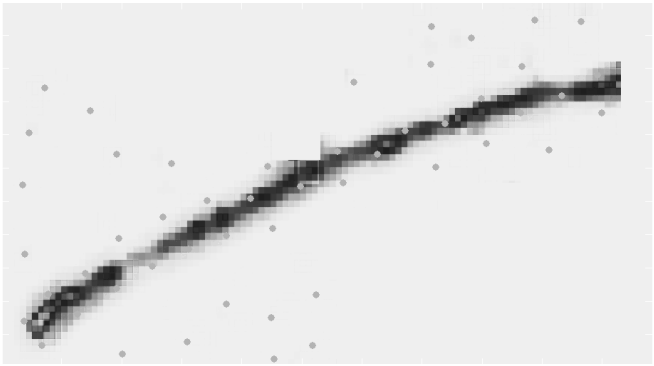
Metódu sme sa rozhodli upraviť s ohľadom na výpočtový čas a fakt, že obrovská časť pixelov má nulové hodnoty zelenej (šedý priestor na obrázku 2). Tieto budú z metódy vylúčené. Teda $X = \left\{ \mathbf{x}_{ik} \mid \mathbf{x}_{ik} \in \mathbf{R}^3, \mathbf{x}_{i3} \neq 0 \right\}$, kde $\mathbf{x}_{ik}$ je hodnota *k-tej zložky i-teho vektora* trénovacieho súboru X, pričom $i = \{1, \mathbf{K}, n\}$ a po vynechaní červenej a modrej zložky RGB sa $k = \{1,2,3\}$, kde 1 = index dimenzie súradníc osy x, 2 = index dimenzie súradníc osy, 3 = index dimenzie hodnôt zelenej.

Ponecháme teda v súbore len nenulové hodnoty (hypotéza o ponechaní len tých pixelov, na ktorých leží synaptonemálny komplex). Následne bola uplatnená metóda *neurónového plynu,* s päťdesiatimi neurónmi.

Po zobrazení výsledkov (viď obrázok 3), sa nám ukázala naša hypotéza o súbore obsahujúcom len tie pixely, na ktorých leží synaptonemálny komplex, ako mylná.

Rozptýlenie sa neurónov po celom priestore obrázku napovedalo o existencií pixelov s malými hodnotami zeleného svetla, ktoré neboli viditeľné na obrázku a ktoré spôsobili odklon neurónov od plochy synaptonemálneho komplexu. Tento problém sme sa rozhodli neriešiť prostredníctvom vylúčenia hodnôt menších ako pevne stanovená hranica ale rozhodli sme sa

skôr upraviť samotný algoritmus neurónového plynu tak, aby bol aplikovateľný i pri analýze ďalších obrázkov.



Obr. 3: Konečná poloha neurónov po prvotnom spustení algoritmu neurónového plynu

Rozhodli sme sa upraviť vzorec (3) do tvaru:

$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t) + a(t) * g(\mathbf{x}_{i3}(t)) * h_l\left(m(\mathbf{x}_i, \mathbf{c}_j)\right) * \left[\mathbf{x}_i(t) - \mathbf{c}_j(t)\right],$$

kde $g(\mathbf{x}_{i3}(t))$ je funkciou hodnôt *tretej zložky i-teho vektora* (teda hodnôt zeleného svetla). Môže nadobúdať váhy v intervale $\langle 0,1 \rangle$. Funkcia je daná predpisom :

$$g(\mathbf{x}_{i3}(t)) = \begin{cases} \dfrac{|V|}{|X|} & \mathbf{x}_{i3}(t) > \mathbf{x}_{i3_q} \\ 0 & \textit{inak} \end{cases}, \qquad (4)$$

kde $\mathbf{x}_{i3_q}$ je hodnota *q*- percentného kvantilu nenulových hodnôt dimenzie zeleného svetla, $|X|$ je počet vektorov (pixelov) s nenulovou hodnotou zeleného svetla väčších ako hodnota *q*-percentného kvantilu všetkých nenulových hodnôt a $|V|$ je počet vektorov (pixelov) s nenulovou hodnotou zeleného svetla väčšou ako hodnota *q*- percentného kvantilu všetkých nenulových hodnôt, a zároveň menšou hodnotou ako je hodnota zeleného svetla vstupného vektoru v čase *t* . Jedná sa o mohutnosť množín:

$$X = \left\{\mathbf{x_{i3}} \mid \mathbf{x_{i3}} \neq 0 \wedge \mathbf{x}_{i3} > \mathbf{x}_{i3_q}\right\}$$

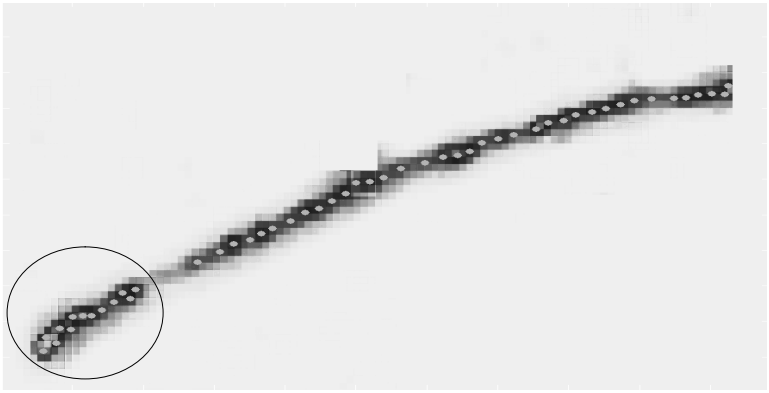
a

$$V = \left\{\mathbf{x_{i3}} \mid \mathbf{x_{i3}} \neq 0 \wedge \mathbf{x}_{i3_q} < \mathbf{x_{i3}} < \mathbf{x}_{i3}(t)\right\}.$$

V prípade, že je hodnota tretej zložky daného vektora väčšia ako je hodnota kvantilu, váha , ktorá vstupuje do výpočtu v adaptačnom vzorci (4) sa vypočíta ako *podiel_počtu* nenulových hodnôt zeleného svetla *menších ako hodnota* zeleného svetla *vstupného vekt*oru ale zároveň *väčších* ako *hodnota q- percentného kvantilu* všetkých nenulových hodnôt k *celkovému počtu* nenulových hodnôt zeleného svetla *väčších* ako *hodnota q- percentného kvantilu* všetkých nenulových hodnôt.

Týmto spôsobom minimalizuje vplyv hodnôt, ktoré sú príliš malé a zároveň hodnotám väčším ako je daný *q*- percentný kvantil sa váha odvíja od kumulatívnej distribučnej funkcie všetkých hodnôt väčších ako je hodnota daného *q*- percentného kvantilu. Vychádzajúc z histogramu hodnôt sme určili hodnotu *q* = 0,8; teda 80 – percentného kvantilu všetkých nenulových hodnôt zeleného svetla.

Výsledná poloha neurónov viditeľná na obrázku 4 odhalila ďalšiu prekážku pri popisovaní tvaru synaptonemálneho komplexu. Napriek tomu, že neuróny správne ukotvili svoju polohu na pixeloch komplexu, v dolnej časti komplexu sa objavil dôsledok nastavenia príliš početnej skupiny neurónov. Nastavenie príliš veľkého počtu spôsobí ich rozptýlenie na pixely bližšie pri krajoch komplexu a teda lepšieho pokrytia priestoru. To, čo je žiadaný efekt v prípade zhlukovania, je v našom prípade prekážkou. Tento problém je však jednoduché odstrániť zvolením menšieho počtu neurónov, tak ako je vidieť na obrázku 5.



Obr. 4: Poloha neurónov po spustení upraveného algoritmu neurónového plynu (50 neurónov) spolu so zvýrazneným úsekom, kde došlo k prílišnému rozptýleniu sa neurónov na synaptonemálnom komplexe

Obr. 5: Poloha neurónov po spustení upraveného algoritmu neurónového plynu (25 neurónov)

Záver

V príspevku sme sa snažili poukázať na spôsob popísania tvaru synaptonemálneho komplexu s použitím fyzikálne inšpirovaného výpočtového modelu s učením bez učiteľa. Pri jeho použití sme však museli uskutočniť zmeny, ktoré viedli k jeho optimalizácií vzhľadom na špecifickosť dát. Do výpočtu sme zahrnuli len vektory s nenulovou hodnotou zeleného svetla a do adaptačného vzorca neurónového plynu sme vložili funkciu, ktorá udeľovala váhy jednotlivým predkladaným vstupným vektorom v závislosti od ich hodnôt zeleného svetla.

Nakoniec sme museli znížiť počet neurónov, aby sme sa vyhli prílišnému rozptýleniu neurónov po povrchu synaptonemálneho komplexu. Otázkou a námetom na ďalšiu prácu zostáva, ako určiť počet neurónov, aby sme dostatočne pokryli povrch komplexu, vystihli jeho tvar ale vyhli sa prílišnému rozptýleniu.

Poďakovanie: *Príspevok bol vytvorený s podporou vedeckovýskumného projektu IGA VSE F4/6/2012*

4. Literatúra

[1] MARTINETZ, T. M., SCHULTEN. K. J. 1991. A „neural-gas" network learns topologies. In: Kohonen, T., Makisara K., Simula K., editori. *Artificial Neural Networks*, s. 397–434. North Holland, Amsterdam.

[2] PASTOREK, L. 2012. Porovnanie sekvenčného a dávkového učenia pri metódach samo-organizujúcich sa máp. In: *Sborník prací účastníků vědeckého semináře doktorského studia FIS VŠE*. s. 227–232. Oeconomica, Praha.

Adresa autora (-ov):

Lukáš Pastorek, Mgr.
Katedra statistiky a pravdepodobnosti
Fakulta informatiky a statistiky VŠE v Praze
nám. W. Churchilla 4
130 67 Praha 3
lukas.pastorek@vse.cz

Hana Řezanková, prof. Ing. CSc.
Katedra statistiky a pravdepodobnosti
Fakulta informatiky a statistiky VŠE v Praze
nám. W. Churchilla 4
130 67 Praha 3
hana.rezankova@vse.cz

Spojenie medzi rovnomernou a seriálnou korelačnou štruktúrou v modeli rastových kriviek

Connection between uniform and serial correlation structure in a growth curve model

Rastislav Rusnačko

Abstract: In this text, we show special correlation structure in the growth curve model which we can be viewed as transition between the serial and the uniform correlation structure in this model. We show estimators of unknown variance parameters for these cases.

Key words: Growth curve model, uniform correlation structure, serial correlation structure, variance parameters, maximum likelihood estimators, Toeplitz matrix

Kľúčové slová: Model rastových kriviek, rovnomerná korelačná štruktúra, seriálna korelačná štruktúra, variančné parametre, maximálne vierohodné odhady, Toeplitzova matica

JEL classification: C13, C29, C39

1. Úvod

Model rastových kriviek predstavili Potthoff a Roy v roku 1964, keď sa snažili odpovedať na otázku, či vzdialenosť medzi hypofýzou a pterygomaxilárnou brázdou u chlapcov a dievčat je rovnaká a či rýchlosť jej rastu je rovnaká. S týmto modelom sa stretneme aj pod názvom zovšeobecnený model viacrozmernej analýzy variancie alebo Potthoffov a Royov model. Časom sa objavili rôzne špeciálne prípady tohto modelu, napríklad v závislosti od uvažovanej korelačnej štruktúry. Medzi často uvažované štruktúry patria rovnomerná a seriálna korelačná štruktúra. V tomto článku ukážeme špeciálnu štruktúru, ktorá môže byť považovaná za spojenie medzi týmito dvomi štruktúrami.

2. Model rastových kriviek

Štandardný model rastových kriviek je tvaru

$$Y = XBZ + \varepsilon, \quad Vec(\varepsilon) \sim N\big(0,\ \Sigma \otimes I\big), \tag{1}$$

kde $Y_{n\times p}$ je matica pozorovaní, $X_{m\times n}$ je ANOVA matica, $Z_{r\times p}$ je matica regresných konštánt, $B_{m\times r}$ je matica neznámych parametrov, $\varepsilon_{n\times p}$ je matica náhodných chýb s normálnym rozdelením, I je jednotková matica, $\Sigma_{p\times p}$ je variančná matica riadkov matice Y a vec operátor skladá stĺpce matice pod seba a vytvára tak z matice stĺpcový vektor. Pri takomto označení indexov predstavuje p počet meraní, n počet objektov a m počet skupín. To znamená, že jednotlivé merania ukladáme do riadkov matice Y. V prípade, že variančná matica Σ je úplne neznáma, tak jej rovnomerne najlepším nestranným odhadom v prípade normality je matica

$$S = \frac{1}{n-r(X)} Y' M_X Y, \tag{2}$$

pričom $\mathrm{M_X} = \mathrm{I} - \mathrm{P_X} = \mathrm{I} - \mathrm{X(X'X)^{-}X'}$.

3. Rovnomerná korelačná štruktúra

Pri tejto štruktúre predpokladáme rovnaké diagonálne a rovnaké mimodiagonálne prvky variančnej matice. Teda je tvaru

$$\Sigma_{p\times p} = \sigma^2[(1-\rho)I + \rho\mathbf{1}\mathbf{1}'] = \sigma^2\begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix}, \tag{3}$$

kde $\sigma > 0$ a $\rho \in \left(-\frac{1}{p-1}, 1\right)$ sú neznáme parametre. Keďže matica S definovaná vzťahom (2) je nevychýleným odhadom variančnej matice, tak $\mathrm{E}\big(\mathrm{Tr}(S)\big) = p\sigma^2$ a $\mathrm{E}(\mathbf{1}'S\mathbf{1}) = p\sigma^2[1 + (p-1)\rho]$. Na základe nevychýlených odhadovacích rovníc sú odhady neznámych parametrov v tvare

$$\hat{\sigma}^2{}_U = \frac{\mathrm{Tr}(S)}{p} \quad \text{a} \quad \hat{\rho}_U = \frac{1}{p-1}\left(\frac{\mathbf{1}'S\mathbf{1}}{\mathrm{Tr}(S)} - 1\right). \tag{4}$$

Tieto odhady odvodil Žežula v [2]. Samozrejme odhady neznámych parametrov závisia od odhadu variančnej matice s ktorým pracujeme. Ďalšie možné odhady variančnej matice sú:

1.) maximálne vierohodný odhad variančnej matice

$$S^M = \frac{1}{n}(Y'M_XY + M_{Z'}Y'P_XYM_{Z'}), \tag{5}$$

2.) odhad pomocou vonkajšieho súčinu

$$S^* = \frac{1}{n-r(X)}Y'M_XY + \frac{1}{n}M_{Z'}Y'YM_{Z'} - \frac{1}{n-r(X)}M_{Z'}Y'M_XYM_{Z'}. \tag{6}$$

Maximálne vierohodné odhady ${}_M\hat{\sigma}_U^2$ a ${}_M\rho_U$ sú rovnakého tvaru ako odhady (4), ale matica S je nahradená maticou S^M.

4. Seriálna korelačná štruktúra

$$\Sigma_{p\times p} = \sigma^2\sum_{i=2}^{p} r^{\,i-1}W_i + s^{\,2}I\,, \tag{7}$$

kde $\sigma > 0$ a $\rho \in (-1, 1)$ sú neznáme parametre, $W_k = (w_{ij}(k))$ je matica typu $p \times p$ s prvkami $\big(w_{ij}(k)\big) = 1$ alebo 0 podľa toho, či $|i-j| = k-1$ alebo $|i-j| \neq k-1$, $k = 2, \dots, p \geq 3$. Táto štruktúra je teda symetrická Toeplitzova matica s prvkami $(1, \rho, \rho^2, \dots, \rho^{p-1})$. Keďže $\mathrm{E}\big(\mathrm{Tr}(S)\big) = p\sigma^2$ a $\mathrm{E}(\mathbf{1}'S\mathbf{1}) = \sigma^2[p + 2\rho(p-1) + 2\rho^2(p-2) + \cdots + 2\rho^{p-1}]$, tak na základe nevychýlených odhadovacích rovníc sú odhady neznámych parametrov v tvare

$$\hat{\sigma}^2{}_S = \frac{\mathrm{Tr}(S)}{p} \quad \text{a} \quad \hat{\rho}_S(p-1) + \hat{\rho}_S{}^2(p-2) + \cdots + \hat{\rho}_S{}^{p-1} = \frac{p}{2}\left(\frac{\mathbf{1}'S\mathbf{1}}{\mathrm{Tr}(S)} - 1\right). \tag{8}$$

Maximálne vierohodné odhady neznámych parametrov v prípade seriálnej korelačnej štruktúry odvodil Žežula a Klein v [2]. Nech $e_{i:p}$ je vektor dĺžky p obsahujúci 1 na i –tom mieste a všade inde nuly. Ďalej nech C_p je matica $I_p - e_{1:p}e'_{1:p} - e_{p:p}e'_{p:p}$ a G_p je symetrická Toeplitzova matica typu $p \times p$ s prvkami $(0, 1, 0, \dots, 0)$. Označme

$$s_1(Y) = \mathrm{Tr}[Y'Y - Y'P_X Y P_{Z'}],$$
$$s_2(Y) = \mathrm{Tr}\big[(Y - P_X Y P_{Z'})'(Y - P_X Y P_{Z'})C_p\big],$$
$$s_3(Y) = \mathrm{Tr}\big[(Y - P_X Y P_{Z'})'(Y - P_X Y P_{Z'})G_p\big].$$

Potom maximálne vierohodné odhady neznámych parametrov pre seriálnu korelačnú štruktúru sú v tvare

$$_M\hat{\sigma}^2{}_S = \frac{1}{np(1-\hat{\rho}^2)}\big(s_1(Y) + \hat{\rho}^2 s_2(Y) - \hat{\rho}s_3(Y)\big), \tag{9}$$

pričom $_M\,\hat{\rho}_S$ je z intervalu $(-1, 1)$ a je riešením rovnice

$$\frac{p-1}{p}s_2(Y)_M\,\hat{\rho}_S{}^3 + \frac{2-p}{2p}s_3(Y)_M\,\hat{\rho}_S{}^2 - \left(\frac{s_1(Y)}{p} + s_2(Y)\right)_M\,\hat{\rho}_S + \frac{s_3(Y)}{2} = 0. \tag{10}$$

5. Spojenie medzi rovnomernou a seriálnou korelačnou štruktúrou

Uvažujme korelačnú štruktúru v tvare

$$\Sigma_{p\times p} = \sigma^2[(1-\rho)I + \rho A]. \tag{11}$$

Rozdiel medzi touto a rovnomernou štruktúrou je v matici A, o ktorej predpokladáme, že je známa symetrická Toeplitzova matica s prvkami napríklad $\left(1, 1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{p-1}\right)$ alebo $\left(1, 1, \frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^{p-1}}\right)$. Pri takýchto uvažovaných tvaroch matice A dostávame korelačné štruktúry v tvare Toeplitzových symetrických matíc s prvkami $\left(1, \rho, \frac{\rho}{2}, \frac{\rho}{3} \ldots, \frac{\rho}{p-1}\right)$, respektíve s prvkami $\left(1, \rho, \frac{\rho}{2}, \frac{\rho}{4}, \ldots, \frac{\rho}{2^{p-1}}\right)$. Našim cieľom je odhadnúť neznáme parametre s^2 a r pomocou metódy maximálnej vierohodnosti. Pre zjednodušenie zápisu označme

$$N = \left(I + \frac{\rho}{1-\rho}A\right)^{-1}. \tag{12}$$

Potom pre determinant a inverziu matice (11) platí

$$\Sigma^{-1} = s^{-2}(1-\rho)^{-1}\left[I - \frac{\rho}{1-\rho}AN\right] \quad \text{a} \quad |\Sigma| = s^{2p}(1-\rho)^p|N^{-1}|. \tag{13}$$

Ak označíme $y = vec(Y)$, $W = Z' \otimes X$, $b = vec(B)$ a $e = vec(\varepsilon)$, tak za predpokladu normality pozorovaní možme model rastových kriviek prepísať do tvaru jednorozmerného modelu

$$y = W\,b + e, \qquad e \sim N_{np}\big(0, \Sigma \otimes I\big).$$

Vierohodnostná funkcia potom je

$$l\big(b, s^2, \rho, y\big) = -\frac{np}{2}\ln(2p) - \frac{np}{2}\ln\big[s^2(1-\rho)\big] - \frac{n}{2}\ln|N^{-1}| -$$
$$-\frac{1}{2s^2(1-\rho)}\left[\big(y - W\,b\big)'\left[\left(I - \frac{\rho}{1-\rho}AN\right)\otimes I\right]\big(y - W\,b\big)\right]. \tag{14}$$

Deriváciou tejto funkcie podľa parametrov b, s^2 a ρ dostávame rovnice

$$W'\left[\left(I - \frac{\rho}{1-\rho}AN\right)\otimes I\right]W\,b = W'\left[\left(I - \frac{\rho}{1-\rho}AN\right)\otimes I\right]y, \tag{15}$$

$$np\,s^{\,2}(1-\rho) = \big(y - W\,b\,\big)'\left[\left(I - \frac{\rho}{1-\rho}AN\right)\otimes I\right]\big(y - W\,b\,\big), \tag{16}$$

$$n\,s^{\,2}[p(1-\rho) - \mathrm{Tr}(NA)] =$$

$$= \big(y - W\,b\,\big)'\left(\left[(3-2\rho)I - \frac{1+\rho}{1-\rho}AN + \frac{\rho}{(1-\rho)^2}ANAN\right]\otimes I\right)\big(y - W\,b\,\big). \tag{17}$$

Z rovnice (15) hneď vidieť, že maximálne vierohodný odhad parametra b je v tvare

$$\widehat{b} = (W'W)^{-1}W'y. \tag{18}$$

Využitím vzťahu $vec(ABC) = \big(C'\otimes A\big)vec(B)$ môžeme jednoduchým spôsobom odvodiť maximálne vierohodný odhad matice regresných konštánt v tvare

$$\hat{B} = (X'X)^{-1}X'YZ'(ZZ')^{-1}. \tag{19}$$

Na základe tohto odhadu matice neznámych parametrov dostávame, že platí $W\widehat{b} = P_W y$ a $y - W\widehat{b} = M_W y = vec(Y - P_X Y P_{Z'})$. Pre jednoduchosť označme teraz

$$d_1(Y) = \mathrm{Tr}[(Y'Y - Y'P_X Y P_{Z'})],$$

$$d_2(Y) = \mathrm{Tr}[(Y - P_X Y P_{Z'})'(Y - P_X Y P_{Z'})AN],$$

$$d_3(Y) = \mathrm{Tr}\left[(Y - P_X Y P_{Z'})'(Y - P_X Y P_{Z'})A\frac{\partial N}{\partial \rho}\right].$$

Použitím tohto označenia a známeho vzťahu $\mathrm{Tr}(A'BCD') = (vec\,A)'\big(D\otimes B\big)vec\,C$ môžeme ľahko vidieť, že vierohodnostná funkcia (14) je nanajvýš rovná funkcii

$$l\left(\widehat{b}\,,\,s^{\,2},\rho,y\right) = -\frac{np}{2}\ln\big[s^{\,2}(1-\rho)\big] - \frac{n}{2}\ln|N^{-1}| - \frac{1}{2\,s^{\,2}(1-\rho)}\left[d_1(Y) - \frac{\rho}{1-\rho}d_2(Y)\right].$$

Deriváciou tejto funkcie podľa parametrov $s^{\,2}$ a ρ dostávame maximálne vierohodné odhady neznámych parametrov v tvare

$$\widehat{{}_A s^2} = \frac{1}{np\left(1 - \widehat{{}_A r}\right)}\left[d_1(Y) - \frac{\widehat{{}_A r}}{1 - \widehat{{}_A r}}d_2(Y)\right], \tag{20}$$

pričom $\widehat{{}_A r}$ je riešením rovnice

$$1 + \frac{\widehat{{}_A r} - 1}{p}\frac{\partial \ln|N^{-1}|}{\partial\, r}\bigg|_{r\,=\,\widehat{{}_A r}} - \frac{1 - \widehat{{}_A r}}{\left(1 - \widehat{{}_A r}\right)d_1(Y) - \widehat{{}_A r}\,d_2(Y)} \times$$

$$\times\left[d_1(Y) - \frac{1 + \widehat{{}_A r}}{1 - \widehat{{}_A r}}d_2(Y) - \widehat{{}_A r}\,d_3(Y)\right] = 0. \tag{21}$$

6. Simulácia na Potthoffových a Royových dentálnych dátach

Pri zavedení modelu rastových kriviek Potthoff a Roy pracovali s dentálnymi dátami, ktoré znázorňovali nameranú vzdialenosť medzi hypofýzou a pretygomaxilárnou brázdou u jedenástich dievčatách a šestnástich chlapcoch vo veku 8, 10, 12 a 14 rokov. Ak by sme jednotlivé merania usporiadali do matice, dostaneme maticu pozorovaní $Y_{27\times 4}$ so stĺpcami C_1, C_2, C_3 a C_4, pričom

$C_1 = [21, 21, 20.5, 23.5, 21.5, 20, 21.5, 23, 20, 16.5, 24.5, 26, 21.5, 23, 20, 25.5, 24.5, 22,$

$$24, 23, 27.5, 23, 21.5, 17, 22.5, 23, 22]',$$
$$\mathrm{C}_2 = [20, 21.5, 24, 24.5, 23, 21, 22.5, 23, 21, 19, 25, 25, 22.5, 22.5, 23.5, 27.5, 25.5, 22,$$
$$21.5, 20.5, 28, 23, 23.5, 24.5, 25.5, 24.5, 21.5]',$$
$$\mathrm{C}_3 = [21.5, 24, 24.5, 25, 22.5, 21, 23, 23.5, 22, 19, 28, 29, 23, 24, 22.5, 26.5, 27, 24.5,$$
$$24.5, 31, 31, 23.5, 24, 26, 25.5, 26, 23.5]',$$
$$\mathrm{C}_4 = [23, 25.5, 26, 26.5, 23.5, 22.5, 25, 24, 21.5, 19.5, 28, 31, 26.5, 27.5, 26, 27, 28.5, 26.5,$$
$$25.5, 26, 31.5, 25, 28, 29.5, 26, 30, 25]'.$$

Matica analýzy variancie je v tomto prípade tvaru

$$\mathrm{X}_{27\times 2} = \begin{pmatrix} \mathbf{1}'_{11} & \mathbf{0}'_{16} \\ \mathbf{0}'_{11} & \mathbf{1}'_{16} \end{pmatrix}'$$

a matica regresných konštánt

$$\mathrm{Z}_{2\times 4} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 10 & 12 & 14 \end{pmatrix}.$$

Na výpočet odhadov neznámych parametrov budeme potrebovať maticu S definovanú vzťahom (2). Po výpočte dostávame, že je tvaru

$$\mathrm{S} = \begin{pmatrix} 5{,}41545 & 2{,}71682 & 3{,}91023 & 2{,}71023 \\ 2{,}71682 & 4{,}18477 & 2{,}92716 & 3{,}31716 \\ 3{,}91023 & 2{,}92716 & 6{,}45574 & 4{,}13074 \\ 2{,}71023 & 3{,}31716 & 4{,}13074 & 4{,}98574 \end{pmatrix}.$$

Pri predpoklade rovnomernej korelačnej štruktúry (3) sú odhady neznámych parametrov

$$\hat{\sigma}^2{}_U = 5{,}26043 \quad \text{a} \quad \hat{\rho}_U = 0{,}62455.$$

Na vyčíslenie príslušných maximálne vierohodných odhadov potrebujeme odhad S^M, ktorý je pre dentálne dáta v tvare

$$S^M = \begin{pmatrix} 5{,}05448 & 2{,}45776 & 3{,}61570 & 2{,}53199 \\ 2{,}45776 & 3{,}95816 & 2{,}71703 & 3{,}03919 \\ 3{,}61570 & 2{,}71703 & 5{,}97877 & 3{,}82170 \\ 2{,}53199 & 3{,}03919 & 3{,}82170 & 4{,}62922 \end{pmatrix}.$$

Teda maximálne vierohodné odhady sú

$${}_M\hat{\sigma}^2_U = 4{,}90516 \quad \text{a} \quad {}_M\rho_U = 0{,}61783.$$

Pri predpoklade seriálnej korelačnej štruktúry (7) dostávame odhady

$$\hat{\sigma}^2{}_S = \hat{\sigma}^2{}_U = 5{,}26043 \quad \text{a} \quad \hat{\rho}_S = 0{,}74353.$$

Maximálne vierohodné odhady sú pri tejto štruktúre definované vzťahmi (9) a (10) a po ich vyčíslení dostávame odhady

$${}_M\hat{\sigma}^2{}_S = 4{,}89087 \quad \text{a} \quad {}_M\hat{\rho}_S = 0{,}60673.$$

Predpokladajme teraz korelačnú štruktúru (11) s Toeplitzovou maticou A_1 s prvkami $\left(1, 1, \frac{1}{2}, \frac{1}{3}\right)$. Odvodený tvar maximálne vierohodných odhadov máme v (20) a (21). Po vyčíslení dostávame nasledujúce odhady neznámych parametrov

$${}_{A_1}\widehat{s^2} = 4{,}29310 \quad \text{a} \quad {}_{A_1}\widehat{r} = 0{,}48032.$$

Ak by sme predpokladali danú korelačnú štruktúru s Toeplitzovou maticou A_2 s prvkami $\left(1, 1, \frac{1}{2}, \frac{1}{4}\right)$, tak odhady budú v tvare

$$\widehat{{}_{A_2}s^2} = 4{,}32044 \quad \text{a} \quad \widehat{{}_{A_2}r} = 0{,}48878.$$

7. Záver

V tomto texte sme ukázali odhady neznámych parametrov v prípade rovnomernej a seriálnej korelačnej štruktúry v modeli rastových kriviek. Odvodili sme aj maximálne vierohodné odhady neznámych parametrov pre špeciálnu korelačnú štruktúru, ktorú môžeme považovať za ich prepojenie. V poslednej kapitole sme ukázali ich použitie na Potthoffových a Royových dentálnych dátach.

Literatúra

[1] ŽEŽULA, I. – KLEIN, D. 2009. The maximum likelihood estimators in the growth curve model with serial covariance structure. In: *Journal of Statistical Planning and Inference*, č. 139, 2010, 3270– 3276.

[2] ŽEŽULA, I. 2006. Special variance structure in the growth curve model. In: *Journal of Multivariate Analysis*, č. 97, 2006, 606 – 618.

Adresa autora:

Rastislav Rusnačko, RNDr.
Prírodovedecká fakulta UPJŠ
Jesenná 5
040 01 Košice
rastislav.rusnacko@student.upjs.sk

Tento článok vznikol za podpory grantu VEGA 1/0410/11 a VVGS-PF-2012-45.

Vlastnosti odhadů J-divergence credit scoringových modelů při Beta rozloženém score

Properties of J-divergence estimators for credit scoring models with Beta distributed scores

Martin Řezáč, Iveta Stankovičová

Abstract: J-divergence is widely used to assess discriminatory power of credit scoring models, i.e. models that try to predict a probability of client's default. However, empirical estimate using deciles of scores, which is the common way how to compute it, may lead to strongly biased results. The main aim of this paper is to describe properties of alternative estimators of J-divergence for credit scoring models with Beta distributed scores. Indeed, better estimator leads to better assessment of models, what may lead to better credit scoring model.

Abstrakt: J-divergence je široce používána k posouzení diskriminační síly credit scoringových modelů, tedy modelů, které se snaží předpovědět pravděpodobnost selhání klienta. Nicméně, empirický odhad pomocí decilů skóre, což je klasický způsob jak ji spočítat, může vést k výrazně vychýleným výsledkům. Hlavním cílem tohoto článku je popsat vlastnosti alternativních odhadů J-divergence pro credit scoringové modely s Beta rozloženým score. Je zřejmé, že lepší odhad vede k lepšímu hodnocení modelů, což může vést k lepšímu credit scoringovému modelu.

Key words: J-divergence, Information Value, Credit Scoring, Beta Distribution.

Klíčová slova: J-divergence, Informační hodnota, Credit scoring, Beta distribuce.

JEL classification: E51, C14, C63

1. Úvod

J-divergence patří mezi často používané způsoby popisu rozdílu mezi dvěma rozděleními pravděpodobnosti. Známá je také pod názvem *informační hodnota (IV),* a to v případě jejího využití pro účely scoringových modelů, tj. např. credit scoring modelů, které jsou používány k určení pravděpodobnosti selhání klienta (tj. situace, kdy klient nedostojí svým úvěrovým závazkům). Credit scoringové modely se v praxi využívají u většiny rozhodnutí, které se týkají poskytování úvěrů, a jsou tak neodmyslytelnou součástí procesů (schvalovacích, vymáhacích, obchodních,...) ve finančním sektoru. Metodologii vývoje credit scoringových modelů a metody posuzování jejich kvality lze nalézt v článcích jako jsou Hand a Henley [3] nebo Thomas [12] a knihách, jako je Anderson [1], Siddiqi [11] nebo Thomas [13].

Článek se primárně zabývá J-divergencí, která je jedním z široce používaných indexů (vedle Giniho indexu a K-S statistiky, viz Wilkie [14] nebo Řezáč a Řezáč [10]) pro posouzení kvality credit scoringových modelů. Většinou se počítá pomocí diskretizace score do intervalů pomocí decilů s požadavkem na nenulový počet pozorování ve všech intervalech. To ale může vést k značně zkreslenému odhadu J-divergence. Mezi alternativní algoritmy odhadu patří empirické odhady založené na ESIS, viz Řezáč [6], nebo přístup založený na teorii jádrových odhadů hustoty, viz Řezáč [7].

Hlavním cílem této práce je popsat vlastnosti vybraných odhadů J-divergence credit scoringových modelů při Beta rozloženém score. Ve druhé a třetí kapitole je uvedena metodologie těchto odhadů včetně algoritmů nebo odkazů na přislušnou literaturu. Čtvrtá kapitola je následně věnována zdůvodnění vhodnosti volby Beta rozložení a odhadům J-

divergence na reálných datech. Dále jsou zde pomocí simulační studie diskutovány vlastnosti jednotlivých odhadů.

2. J-divergence pro Beta rozložené score

Jeffreyho divergence (J-divergence) dvou náhodných veličin X_0 a X_1 s hustotami $f_0(x)$ a $f_1(x)$ je definovaná jako symetrizovaná Kullback-Leiblerova divergence, tj.

$$D_J(X_0, X_1) = D_{KL}(X_0 : X_1) + D_{KL}(X_1 : X_0) = \int_{-\infty}^{\infty} \big(f_0(x) - f_1(x)\big) \ln\left(\frac{f_0(x)}{f_1(x)}\right) dx, \qquad (1)$$

kde Kullback-Leiblerova divergence $D_{KL}(X_0 : X_1)$ je dána vztahem

$$D_{KL}(X_0 : X_1) = \int_{-\infty}^{\infty} \big(f_0(x)\big) \ln\left(\frac{f_0(x)}{f_1(x)}\right) dx. \qquad (2)$$

Uvažujme tedy dvě náhodné veličiny X_0 a X_1 představující vhodně transformované výstupy daného credit scoringového modelu pro špatné (klienti v selhání) a dobré klienty. Nechť se tyto náhodné veličiny řídí Beta rozdělením s hustotami $f_0(x)$ a $f_1(x)$ definované vztahy:

$$f_0(x) = \begin{cases} \dfrac{1}{B(a_0, b_0) \cdot s_0^{\,a_0+b_0-1}} (x - J_0)^{a_0-1} \cdot (s_0 + J_0 - x)^{b_0-1} & \text{pro } J_0 < x < J_0 + s_0 \\ 0 & \text{pro } x \le J_0 \text{ nebo } x \ge J_0 + s_0 \end{cases} \qquad (3)$$

$$f_1(x) = \begin{cases} \dfrac{1}{B(a_1, b_1) \cdot s_1^{\,a_1+b_1-1}} (x - J_1)^{a_1-1} \cdot (s_1 + J_1 - x)^{b_1-1} & \text{pro } J_1 < x < J_1 + s_1 \\ 0 & \text{pro } x \le J_1 \text{ nebo } x \ge J_1 + s_1. \end{cases} \qquad (4)$$

Snadno se ukáže, že transformace $h_i(x) = \dfrac{x - J_i}{s_i}$, i = 0,1, převedou náhodné veličiny X_0 a X_1 na náhodné veličiny Y_0 a Y_1 s hustotami

$$g_0(x) = \begin{cases} \dfrac{1}{B(a_0, b_0)} x^{a_0-1} \cdot (1-x)^{b_0-1} & \text{pro } 0 < x < 1 \\ 0 & \text{jinak} \end{cases} \qquad (5)$$

$$g_1(x) = \begin{cases} \dfrac{1}{B(a_1, b_1)} x^{a_1-1} \cdot (1-x)^{b_1-1} & \text{pro } 0 < x < 1 \\ 0 & \text{jinak.} \end{cases} \qquad (6)$$

Pro takto rozložené náhodné veličiny lze nalézt analytické vyjádření Kullback-Leiblerovi divergence, a tedy i J-divergence. Dostáváme tedy

$$\begin{aligned} D_J(Y_0, Y_1) = {} & (a_1 - a_0) \cdot \big(y(a_1) - y(a_0)\big) + (b_1 - b_0) \cdot \big(y(b_1) - y(b_0)\big) + \\ & + (a_1 - a_0 + b_1 - b_0) \cdot \big(y(a_0 + b_0) - y(a_1 + b_1)\big), \end{aligned} \qquad (7)$$

kde $y(t)$ je digamma funkce (detaily o digamma funkci viz Gradshtein a Ryzhik [2] nebo Medina a Moll [5]). Pro výpočet lze použít také aproximativní vzorec využívají vztahu $y(t) \approx \ln(t - 0{,}5)$ (viz Johnson, Kotz a Balakrishnan [4]). Pak platí

$$D_J(Y_0, Y_1) \approx \ln\left[\left(\frac{a_1 - 0{,}5}{a_0 - 0{,}5}\right)^{a_1-a_0} \cdot \left(\frac{b_1 - 0{,}5}{b_0 - 0{,}5}\right)^{b_1-b_0} \cdot \left(\frac{a_0 + b_0 - 0{,}5}{a_1 + b_1 - 0{,}5}\right)^{a_1-a_0+b_1-b_0}\right]. \qquad (8)$$

Pro praktický odhad J-divergence je zapotřebí ještě odhadnout parametry a_0, a_1, b_0 a b_1. Typicky se tak děje pomocí MLE odhadů. Ty nabízí např. procedura Univariate systému SAS. Výpočetní schéma pak lze nalézt v Johnson, Kotz a Balakrishnan [4].

3. Neparametrické odhady J-divergence

V praxi mezi nejčastěji používané neparametrické estimátory J-divergence patří tzv. empirické odhady. Ty jsou založeny na myšlence nahrazení neznámých hustot pomocí empirických odhadů těchto hustot, de facto pomocí vhodných relativních četností. Uvažujme n_0 hodnot score $s_{0_i}, i = 1, \mathbf{K}, n_0$ pro špatné klienty a n_1 hodnot score $s_{1_i}, i = 1, \mathbf{K}, n_1$ pro dobré klienty a označme L (resp. H) minimum (resp. maximum) všech těchto hodnot. Rozdělme interval $[L,H]$ na r podintervalů $[q_0, q_1], (q_1, q_2], \ldots, (q_{r-1}, q_r]$, kde $q_0 = L - 1, q_r = H + 1$, $q_i, i = 1, \mathbf{K}, r-1$ a $q_i, i = 1, \mathbf{K}, r-1$ jsou vhodné hraniční body, např. Příslušné kvantyly score všech klientů. Označme

$$\begin{aligned} n_{0_j} &= \sum_{i=1}^{n_0} I(s_{0_i} \in (q_{j-1}, q_j]) \\ n_{1_j} &= \sum_{i=1}^{n_1} I(s_{1_i} \in (q_{j-1}, q_j]) \quad j = 1, \mathbf{K}, r \end{aligned} \tag{9}$$

pozorované počty špatných, resp. dobrých, klientů v každém intervalu. Dále označme $\hat{f}_{IV}(j)$ příspěvek k J-divergenci na j-tém intervalu, definovaný jako

$$\hat{f}_{IV}(j) = \left(\frac{n_{1_j}}{n_1} - \frac{n_{0_j}}{n_0} \right) \ln \left(\frac{n_{1_j} n_0}{n_{0_j} n_1} \right), \; j = 1, \mathbf{K}, r\,. \tag{10}$$

Empirický odhad J-divergence je poté dán vztahem

$$\hat{D}_J = \sum_{j=1}^{r} \hat{f}_{IV}(j). \tag{11}$$

Speciálním případem je tzv. decilový odhad využívající pro určení hranic intervalů q_i score všech klientů a r =10. Dalšími jsou pak např. algoritmy ESIS (viz Řezáč [6]), ESIS1 (viz Řezáč a Koláček [9]) nebo ESIS2 (viz Řezáč [8]).

Dalším možným přístupem k odhadu J-divergence je využití teorie jádrových odhadů hustoty. Pro M+1 ekvidistantních bodů score $L = x_0, x_1, \mathbf{K}, x_M = H$ máme

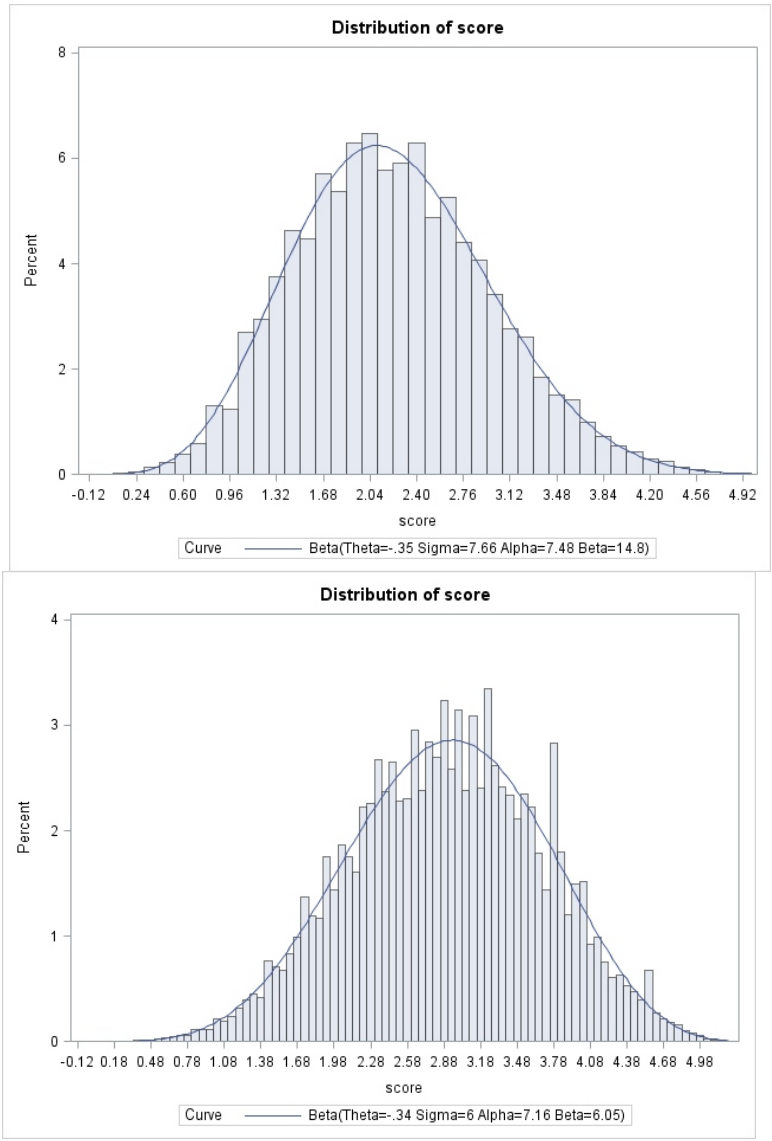
$$\hat{D}_J = \frac{H - L}{2M} \left(\tilde{f}_{IV}(L) + 2 \sum_{i=1}^{M-1} \tilde{f}_{IV}(x_i) + \tilde{f}_{IV}(H) \right), \tag{12}$$

kde $\tilde{f}_{IV}(L)$ jsou odhadnuté příspěvky k J-divergenci dané vhodnými jádrovými odhady neznámých hustot score špatných a dobrých klientů. Více viz Řezáč [7].

4. Výsledky

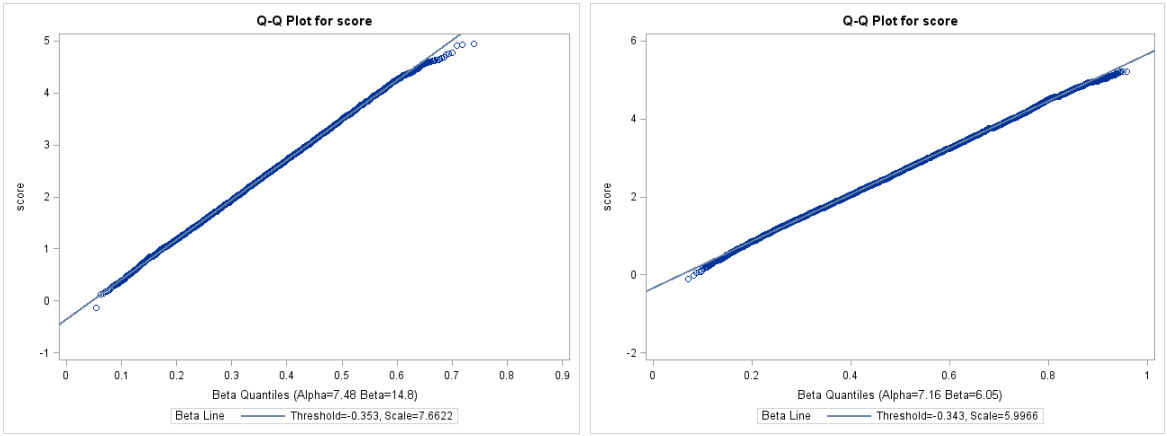
Logickou otázkou je proč uvažovat právě Beta rozložení. Odpověď lze nalézt v následujících obrázcích 1 a 2 a tabulkách 1 a 2, získané v systému SAS pomocí procedury Univariate. K dispozici byla reálná data poskytnutá jistou finanční institucí obsahující výstup credit scoringové funkce (inverzní logit z 1 mínus pravděpodobnost defaultu) a ukazatel

dobrého klienta. Celkový rozsah dat byl 176 878 pozorování (bližší popis viz Řezáč a Řezáč [10]).



Obrázek 1: Nafitované Beta rozdělení score pro špatné (vlevo) a dobré (vpravo) klienty.

Z obrázku 1 (nafitované hustoty Beta rozložení a histogramy) a především z obrázku 2 (Q-Q grafy) je zřejmé, že volba Beta rozložení byla správná.



Obrázek 2: Q-Q grafy pro Beta rozdělení score pro špatné (vlevo) a dobré (vpravo) klienty.

Následující tabulka 1 obsahuje výsledky testů shody zkoumaných dat s Beta rozložením. V případě score špatných klientů všechny uvažované testy nezamítají hypotézu o Beta rozložení. Na druhou stranu pro score dobrých klientů došlo k přibližně desetinásobnému růstu testových statistik pro Cramer-von Misesův i Anderson-Darlingův test a zachování testové statistiky Kolmogorovova-Smirnovova testu na přibližné úrovni. Celkově v tomto případě došlo k zamítnutí hypotézy o Beta rozloženém score u všech třech zmíněných testů. Problém spočívá ve značném rozsahu (cca 160 000 pozorování) dat pro score dobrých klientů. Při provedení stejných testů pro náhodný výběr score dobrých klientů čítající 10% původního počtu pozorování dostaneme přibližně stejné výsledky jako pro score špatných klientů (dokonce s vyššími p-hodnotami u všech tří testů). Celkově tedy považujeme za vhodné pokládat score špatných i dobrých klientů za Beta rozložené.

Tabulka 1: Testy pro Beta rozložení pro score špatných (vlevo) a dobrých (vpravo) klientů.

Goodness-of-Fit Tests for Beta Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.00863	Pr > D	0.117
Cramer-von Mises	W-Sq	0.09981	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.73667	Pr > A-Sq	>0.250

Goodness-of-Fit Tests for Beta Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.00905	Pr > D	<0.001
Cramer-von Mises	W-Sq	1.05559	Pr > W-Sq	0.002
Anderson-Darling	A-Sq	8.24061	Pr > A-Sq	<0.001

Tabulka 2 dále obsahuje parametry nafitovaných Beta rozložení. Nejvýraznější rozdíl mezi score špatných a dobrých klientů je dán parametrem beta (14.8 vs. 6.05) a sigma (7.66 vs. 5.99).

Tabulka 2: Parametry Beta rozdělení score špatných (vlevo) a dobrých (vpravo) klientů.

Parameters for Beta Distribution		
Parameter	Symbol	Estimate
Threshold	Theta	-0.35327
Scale	Sigma	7.662249
Shape	Alpha	7.479128
Shape	Beta	14.81027
Mean		2.217774
Std Dev		0.749697

Parameters for Beta Distribution		
Parameter	Symbol	Estimate
Threshold	Theta	-0.34301
Scale	Sigma	5.996635
Shape	Alpha	7.15525
Shape	Beta	6.05231
Mean		2.905692
Std Dev		0.792681

Následující tabulka 3 udává hodnoty D_J odhadnuté pomocí výše uvedených algoritmů. Poslední řádek obsahuje parametrický odhad daný (7) s MLE odhady parametrů. Jako nejvhodnější neparametrický odhad se jeví hodnota 3,117 daná algoritmem ESIS1.

Tabulka 3: Odhady D_J.

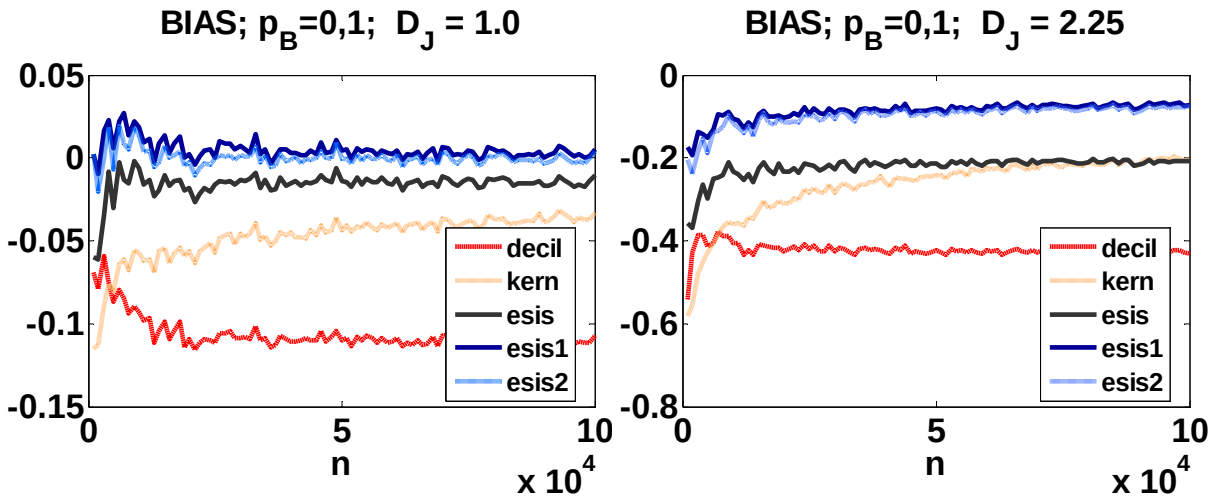
	D_J
decil	2,508551
kern	2,797372
esis	2,945658
esis1	3,117013
esis2	2,967163
param	3,403594

Otázkou je ovšem obecné chování zmíněných algoritmů. Velmi častým způsobem posouzení kvality/vlastností odhadů nějakého parametru či statistiky jsou vychýlení (Bias) a střední kvadratická chyba (MSE) definované jako

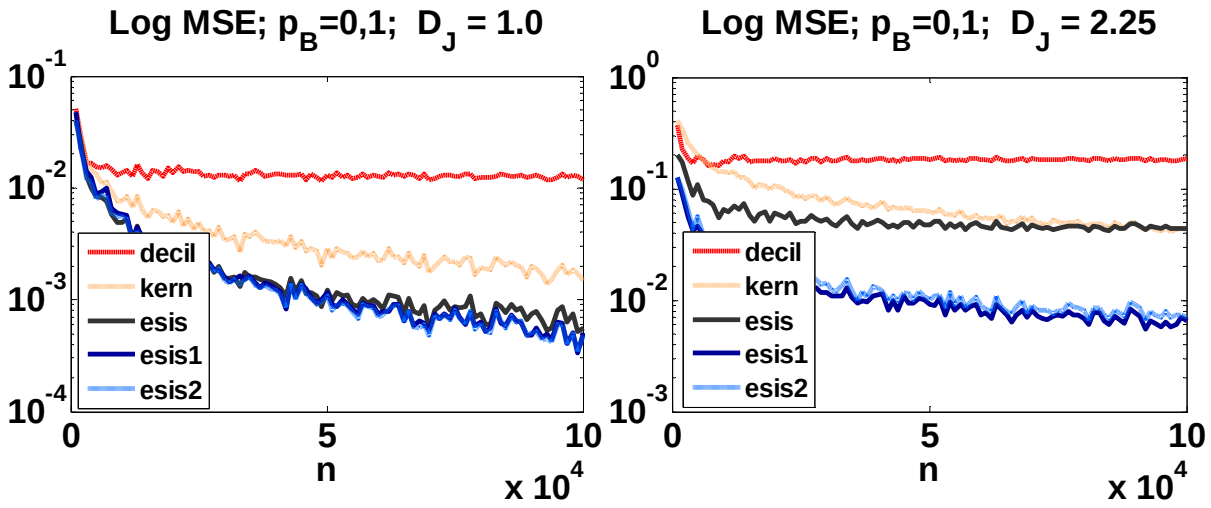
$$bias = E\left(\hat{D}_J\right) - D_J, \tag{13}$$

$$MSE = E\left(\left(\hat{D}_J - D\right)^2\right) \tag{14}$$

Následující obrázky 3 a 4 zobrazují vlastnosti zmíněných algoritmů právě z tohoto pohledu. Simulační studie vedoucí k těmto výsledkům byla provedena následovně. Uvažujeme n klientů, $n \cdot p_B$ špatných a $n \cdot (1 - p_B)$ dobrých (p_B je relativní četnost špatných klientů), v našem případě volíme p_B = 0,1, což nejvíce odpovídá výše zmíněným reálným datům. Dále uvažujeme parametry Beta rozložení vedoucí k hodnotě D_J = 1,00 a 2,25. Velikost datového vzorku volíme n = *1000* až n = *100 000*. Nejprve vygenerujeme score špatných a dobrých klientů v závislosti na zvolených parametrech. Následně spočteme všechny výše zmíněné odhady $\hat{D}_J$. Tento postup zopakujeme tisíckrát. Střední hodnoty pro vychýlení a MSE jsou pak spočteny jako příslušné aritmetické průměry.



Obrázek 3: Vychýlení odhadů $\hat{D}_J$ pro Beta rozdělené score.



Obrázek 4:Logaritmus MSE odhadů $\hat{D}_J$ pro Beta rozdělené score.

Z obrázků 3 a 4 je patrné, že decilový odhad je značně vychýlen, přesněji řečeno podhodnocen. Hodnota log MSE se poměrně záhy stabilizuje a s rostoucím počtem pozorování neklesá. Celkově jde tedy o odhad ne příliš vhodný. Naproti tomu algoritmy ESIS1 a ESIS2 vedou v případě slabšího modelu ($D_J = 1{,}00$) k téměř nevychýlenému odhadu. Pro silnější model ($D_J = 2{,}25$) jsou jejich vlastnosti horší, nicméně nejlepší ze všech uvažovaných metod odhadu D_J.

5. Závěr

Cílem této práce bylo popsat vlastnosti vybraných odhadů J-divergence (též Informační hodnota) credit scoringových modelů při Beta rozloženém score. Byl uveden vzorec pro teoretickou hodnotu J-divergence za předpokladu tohto typu rozložení. Jeho znalost umožnila jednak počítat parametrický odhad, ale také posoudit kvalitu neparametrických odhadů. Na reálných datech byl představen výpočet odhadů J-divergence. Mimo to, na simulovaných datech z Beta rozložení byly demenstrovány vlastnosti uvedených odhadů, konkrétně vychýlení a MSE pro rozsahy dat od 1000 po 100 000. Zcela zřejmě se ukázaly slabiny klasického decilového empirického odhadu. Naproti tomu se zdá, že vhodným odhadem J-divergence při Beta rozloženém score jsou algoritmy ESIS1 a ESIS2.

Literatura

[1] ANDERSON, R. 2007. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press, Oxford.

[2] GRADSTEIN, I. S. AND RYZHIK, I. M. 1965. *Tables of integrals, sums, series and products*. Academic Press, New York and London.

[3] HAND, D. J. AND HENLEY, W. E. 1997. Statistical Classification Methods in Consumer Credit Scoring: a review. In: *Journal. of the Royal Statistical Society*, Series A, 160,No.3, s. 523-541.

[4] JOHNSON, N. L., KOTZ, S., BALAKRISHNAN, N. 1995. *Continuous Univariate Distributions, volume 2*, 2nd edition, Wiley, New York.

[5] MEDINA, L. A. AND MOLL, V. H. 2009. The integrals in Gradshteyn and Ryzhik. Part 10: The digamma function. In: *Scientia, Series A: Mathemaical Sciences* 17, s. 45-66.

[6] ŘEZÁČ, M. 2011. Estimating Information Value for Credit Scoring Models. In: *Aplimat - Journal of Applied Mathematics*, s. 1619-1628.

[7] ŘEZÁČ, M. 2011. Advanced empirical estimate of information value for credit scoring models. In: *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis* LIX (2), s. 267-273.

[8] ŘEZÁČ, M. 2012. Information Value Estimator for Credit Scoring Models. In: *Proceedings of ECDM 2012*, Lisboa, s. 188-192.

[9] ŘEZÁČ, M. AND KOLÁČEK, J. 2011. Computation of Information Value for Credit Scoring Models. In: Workshop of the Jaroslav Hájek Center and Financial Mathematics in Practice I, Book of short papers, s. 75-84.

[10] ŘEZÁČ, M. AND ŘEZÁČ, F. 2001. How to Measure the Quality of Credit Scoring Models. In: *Finance a úvěr - Czech Journal of Economics and Finance* 61 (5), s. 486-507.

[11] SIDDIQI, N. 2006. *Credit Risk Scorecards: developing and implementing intelligent credit scoring*, Wiley, New Jersey.

[12] THOMAS, L. C. 2000. A survey of credit and behavioural scoring: forecasting _nancial risk of lending to consumers. In: *International Journal of Forecasting* 16 (2), s. 149-172.

[13] THOMAS, L. C. 2009. *Consumer Credit Models: Pricing, Profit, and Portfolio*. Oxford University Press, Oxford.

[14] WILKIE, A. D. 2004. Measures for comparing scoring systems. In: Thomas, L. C., Edelman, D. B., Crook, J. N. (Eds.), *Readings in Credit Scoring*. Oxford University Press, Oxford, s. 51-62.

Adresa autorů:

Martin Řezáč, Mgr. Ph.D.
Ústav matematiky a statistiky PřF MU
Kotlářská 2
611 37 Brno
mrezac@math.muni.cz

Iveta Stankovičová, doc. Ing. PhD.
Katedra informačných systémov FM UK
Odbojárov 10
820 05, Bratislava
iveta.stankovicova@fm.uniba.sk

Inspecting Correlations of World Stock Market Indices using Self-Organizing Maps

Vyšetrovanie korelácií medzi svetovými burzovými indexami so samoorganizujúcimi sa mapami

Miroslav Sabo

Abstract: We analyze dependencies among world stock market indices using self-organizing maps. Mutual dependencies between pairs of indices are first analyzed and many positive correlations are revealed. Indices are then clustered with Kohonen maps and two clusters of similar indices are found. Before and after crisis behaviour of selected indices is finally analyzed.

Abstrakt: V tomto článku analyzujeme závislosti medzi svetovými burzovými indexami pomocou samoorganizujúcich sa máp. Najskôr skúmame vzájomné vzťahy medzi vybranými indexami, ktoré neskôr vstupujú do zhlukovej analýzy za účelom nájdenia skupín indexov s podobným správaním. Nakoniec porovnávame správanie sa indexov pred a po kríze.

Key words: crisis, stock market index, neural networks, dependency, self-organizing maps.

Kľúčové slová: kríza, burzový index, neurónová sieť, závislosť, samoorganizujúca sa mapa

JEL classification: C45, F37, C3

1. Introduction

Stock market indices are considered as primary indicators of country´s economic strength. They are also used in other areas, i.e. to measure performance of portfolios (mutual funds). National indices (for example Dow Jones Industrial Average) are mostly composed of the stocks of large companies of the country, but there are also composite indices (for example Euro Stoxx 50) that are cumputed from several international companies to measure economic strength of larger regions. Therefore it is important to investigate their time-varying behaviour and mutual dependencies. One possible application of this knowledge may be to understand inner relations among them and then to predict future values of one index with the knowledge of the others. Past articles concerning stock market indices may be divided into two groups: articles studying future prediction methods (mostly for one index) and articles discussing mutual relations between more than one index.

From all branches of science, prediction analysis in finance has special importance since it can be very helpful in many areas. Another rarity is the fact that financial time series are not dependent only on past events, but also on hidden factors that cannot be measured. Especially stock market indices and world exchange rates are dependent not only on other world indices or past events, but mainly on psychological and political factors [7]. Therefore, many people are interested in modelling financial time series.

There are many philosophical views on predicting economical times series. Some people say they can be predictable, but there are also opinions that any prediction in finance is impossible, because financial time series are strictly stochastic (with strong nonlinear dependencies that are not able to be revealed) [4].

From the second group of applications, there are some multivariate parametric approaches for modelling multidimensional dependencies, widely used are copulas. In [2] international stock market is analyzed via copula methods.

2. Self-organizing maps (SOM)

SOM (proposed by [3]) is a kind of artificial neural network that is used as nonlinear dimension reduction method. Their graphical output, called U-matrix (see Fig. 4), is interpreted as similarity map and near objects have also similar features. Color intensities are used to depict distances between adjacent parts of map. In other words, U-matrix can be used as visualization tool where clusters of similar objects can be identified easily. Moreover, if we also want to describe features of objects from any part of U-matrix, we can look at component planes (see Fig.3) where color intensities depict values of corresponding feature. See [3] for more details about SOM.

3. Preprocessing data

The data we are using is available on Eurostat website. In section Monetary and other financial statistics, there is Share prices indices section. For further analysis we downloaded two datasets: share prices indices - annual data and share prices indices - monthly data (all available indices from database were used).

Moreover, we selected time period from January 2000 to July 2011 for the first dataset and period from 2001 to 2010 for the second one. Data was rebased according to values in year 2005 (to eliminate too different values, if for example, data was rebased according to year 1995). Before analysis, we first removed all indices with at least one missing value.

In further analyses, we used only statistical open-source software R and Matlab with SOM toolbox [1].

4. Dependencies among price indices (monthly data)

The task for this section is to analyse relations among indices, both linear and nonlinear. It is well known that most indices are strongly dependent on the other. The situation is easily seen especially in crisis period, when decrease of one index causes decrease of many others. In Fig. 1, there is R visualization of Pearson linear correlation matrix. Size of balls represents the strength of dependence between corresponding indices. Almost all dependencies are positive. We also investigated nonlinear dependencies using Kendall and Spearman correlation coefficients, but only small differences were revealed. Therefore in further we will assume only Pearson correlations.

Correlation matrix describes only pairwise correlations and from that it is very difficult to see how indices dependent on others simultaneously. To see this, we used self-organizing maps mentioned above, since it is effectively able to project original data to 2D space with preserved toology, i.e. indices that are closely related will be also projected to close neurons. In Fig. 2, there is Matlab visualization of U-matrix. From this, we can conclude two informations:

- When looking at color of neurons, one can immediately see that there are two significant clusters - first in the bottom part of U-matrix and all other objects create second (and bigger) one.
- Moreover, bottom cluster is divided into two subclusters - left one consisting of Slovakia, Czech Republic, Austria, Estonia, Latvia and Lithuania and right one consisting of Poland, Turkey and Norway. Hungary is in the middle of these two subclusters. Comparing all cluster-one countries, we can conclude that it is created by Central Europe and Baltic region countries.

The second cluster is more homogenous and consists of all assumed summary indices that were projected in the upper left part of U-matrix (white neurons in this part mean very

strong dependencies). In the lower left part of this cluster, there are Irish and Belgian indices that are separated from other cluster-two objects.

Special case is Malta which is in the middle position of two biggest clusters, far from other neighboring indices (indicated by black neurons around it).

Fig. 1: R visualization of Pearson correlation matrix among selected stock market indices
Source: Data from http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ and code from http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=152.

Dependencies among price indices (annual data)

Fig. 2 displays U-matrix of countries based on annual data (not all countries were used according to missing values, therefore there are not the same countries as before). When comparing these two, no significant differences can be seen (since annual data is only average value of monthly data). When looking at component planes (see Fig. 3), one can conclude the following:

- There are strong correlations between years 2001-2004 and 2007-2010. It is not a surprise, since these years are consecutive. Note, that component planes can be used also for correlation analysis. If two or more are similar, than there exist strong (linear and nonlinear) dependencies between variables, that these planes represent.
- Matrix for year 2005 was omitted, since all indices have the same value in this time equal to 100, since all ones were rebased according to this time. Moreover, there are two similar groups of component planes - periods 2001-2004 and 2007-2010, what indices before and after crisis times.
- Looking at period 2001-2004, we can see, that upper cluster countries had high values of their indices during this period and lower cluster ones low. On the contrary, during crisis, there were some countries from lower cluster (Poland, Turkey and Norway, projected to the bottom right neuron) that had high values of indices despite the fact, that all other indices decreased.

5. Inspecting before-after crisis behaviour simultaneously (annual data)

In this section we will show how can self-organizing maps be helpful in inspecting before-after crisis behaviour of indices (we have to remark that after crisis period does not mean period after the crisis ended, but period after the crisis started). We first divided all indices to two periods-2001-2004 and 2007-2010 (according to previous findings). Therefore, number of assumed objects has now doubled (each country was projected to SOM map two times - first according to its before crisis values and the second time according to crisis values).

In Fig. 4, there is U-matrix for this situation. We had to slightly move some labels to avoid overlapping, but there is only small mismatch with true situation. Results are following:

- Upper half of U-matrix consists of almost all before crisis indices, whereas to the lower half part, there were projected almost all after crisis indices.
- There can be seen some special behaviour examples, see for instance Amsterdam Exchange Index before crisis and the same one after crisis (lower right part of matrix, denoted by arrow). Both were projected to close neurons, what indicates similar index values before and after crisis.
- There are three clusters of objects-upper cluster, bottom left one and rest of matrix creates third (and the biggest cluster). First one consists of before crisis indices with too low values of price indices. This cluster is created by almost Central Europe and Baltic region countries with Romania, Bulgaria, Turkey and Norway.
- The second cluster (bottom left) contains Poland, Turkey, Norway and Germany after the crisis started. As mentioned before, these countries had higher indices values after crisis then before, therefore they create special cluster. Moreover, countries that are close to these have also similar behaviour (for example Finland).
- We can also see particular time movement of each country, for example see Wrasawzki Indeks Giedovy that moved the most (from upper right to bottom left part, in Fig. 4 represented by arrow), what indicates significant increase of its value after the crisis started.

Similarly as mentioned before, Amsterdam Exchange index before crisis value was projected close to its after crisis value, what indicates no significant change.

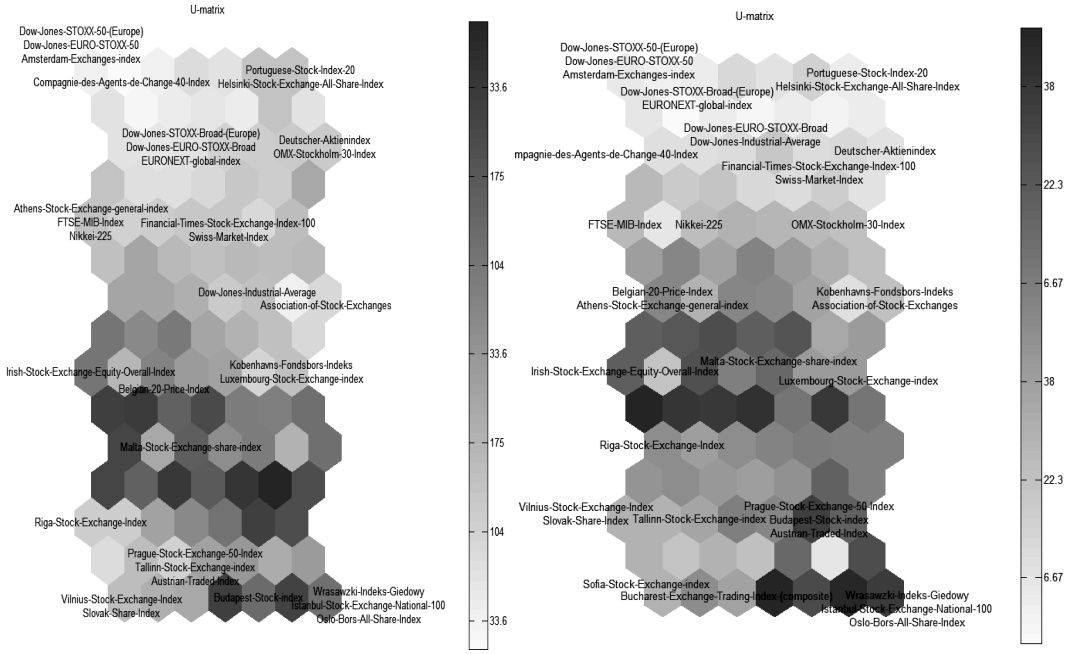


Fig. 2: U-matrix for monthly data (left) and for annual data (right)
Source: Own construction in MATLAB.

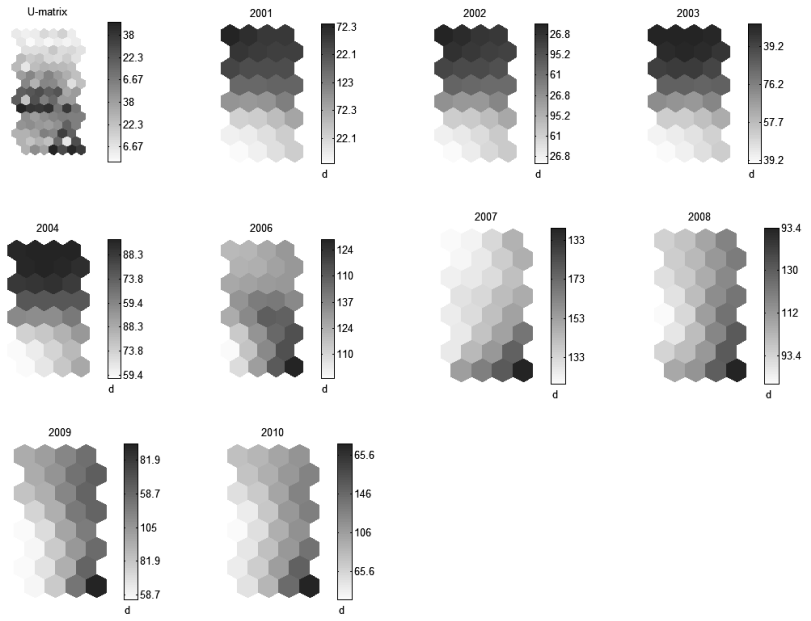


Fig. 3: U-matrix (upper left) and component planes for all assumed years (annual data)
Source: Own construction in MATLAB.

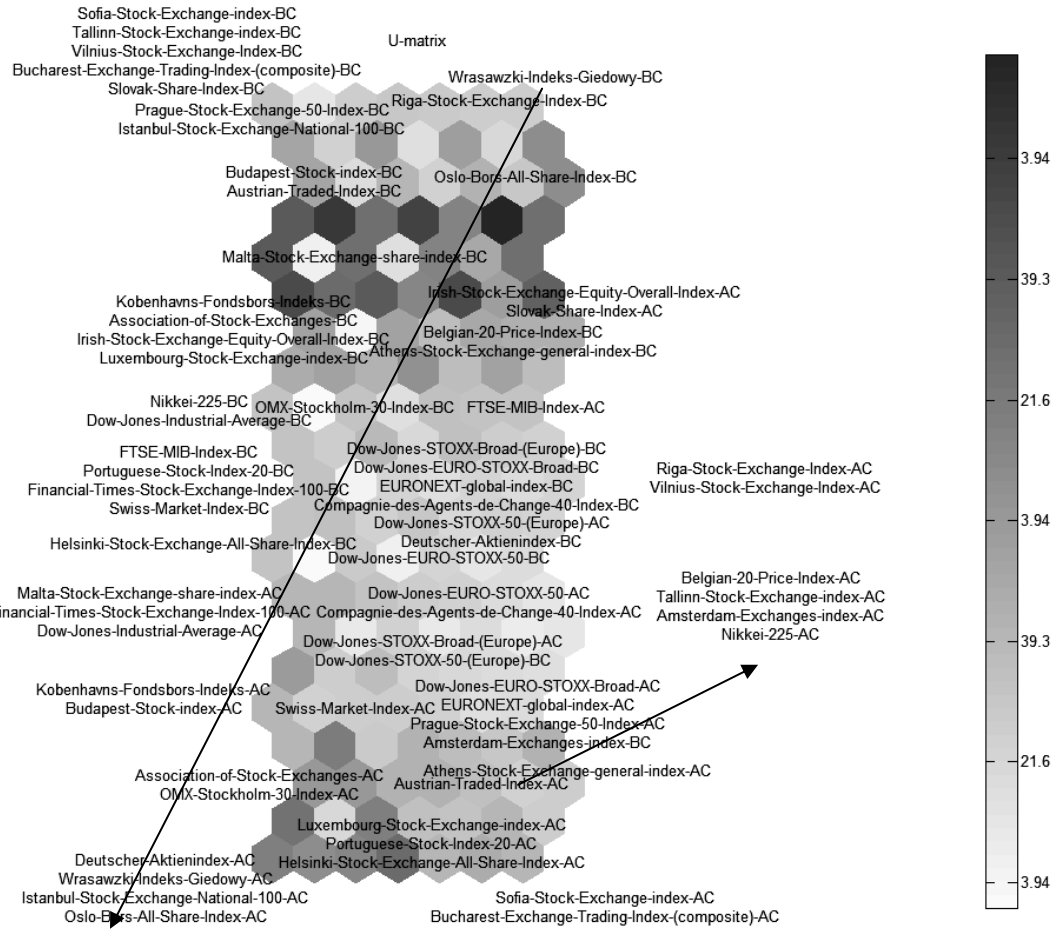


Fig. 4: U-matrix for before-after crisis comparison (annual data)
Source: Own construction in MATLAB.

6. Conclusion

This article had two aims - to inspect overall dependencies among selected world national and composite stock market indices and to analyse behaviour of these before and after crisis using Kohonen maps.

In the first part, we revealed two significant clusters of similar indices - cluster one created by Central Europe and Baltic region countries together with Turkey and Norway and the second one consisting from all composite indices, rest European countries indices and other assumed world indices. These clusters represent similar behaviour of objects that were projected to them, i.e. linear and nonlinear dependencies.

It has also been showed that there are only positive dependencies among all assumed world indices.

In the second part, indices were studied according to their before and after crisis movements. Indices, that belong to the same clusters according to the first part of article showed also similar behavior after the crisis started. Moreover, there are also some national indices, whose values increased significantly after the crisis started (Poland, Turkey, Norway).

Since self-organizing maps allow simple visual analysis of multivariate datasets, we recommend them as effective tool when monitoring more than two indicators simultaneously is needed (when they cannot be visualized with simple scatterplots).

Acknowledgments: This work was supported by VEGA 1/0143/11.

References

[1] ALHONIEMI, E. – HIMBERG, J. – PARHANKANGAS, J. – VESANTO, J. 2011. Som toolbox, version 2.0 beta http://www.cis.hut.fi/projects/somtoolbox/

[2] JONDEAU, E. – ROCKINGER, M. 2006. The Copula-GARCH model of conditional dependencies: An international stock market application. In: *Journal of International Money and Finance*, Volume 25 (5), pp. 827 - 853.

[3] KOHONEN, T. 1997. Self-Organizing Maps. *Series in Information Sciences*, Vol. 30. Springer, Heidelberg. Second ed. 1997.

[4] MARKEI, B.G. 1999. *A random walk down wall street*, W.W. Norton and Company, New York, London.

[5] MATLAB version 7.6. 2010. Natick, Massachusetts: The MathWorks Inc.

[6] R DEVELOPMENT CORE TEAM 2011. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[7] ZHANG, Y. Q. – WAN, X. 2007. Statistical fuzzy interval neural networks for currency exchange rate time series prediction. In: *Applied Soft Computing*, Volume 7 (4), pp. 1149 - 1156.

[8] [cit. 6.8.2011] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

[9] [cit. 6.8.2011] http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=152

Adresa autora:

Miroslav Sabo, Mgr.
Stavebná fakulta STU v Bratislave
Radlinského 11, 813 68 Bratislava
sabo@math.sk

Probability Modelling by Generalized Kappa Distribution

Pravdepodobnostné modelovanie zovšeobecneným kappa rozdelením

Ľubica Sipková, Juraj Sipko

Abstract: The aim of this article is to present the useful tool for fitting a probability model of observed or generated continuous data with extremely skewed empirical distribution. Generalized forms of theoretical probability distributions, defined by inverses of their distribution functions, are extremely flexible in this regard, but also are difficult to fit. The Generalized Kappa Distribution (GKD), originally proposed by Hosking, is also defined by its quantile function, that has proved useful in a number of different applications. Because it can assume a wide variety of shapes, the GKD could be fitted as a model of extreme events such as floods, wind storms or heavy rains, and is essential in the design of water-related structures in hydrology. Furthermore, the GKD offers risk managers great flexibility in modelling a broad range of financial data, and in social studies it is manageable tool for modelling of income data as the outcome of a stochastic process.

Key words: Kappa model, Generalized probability model, Inverse distribution function, Quantile function, Income inequality, Quantiles.

Kľúčové slová: kappa model, zovšeobecnený pravdepodobnostný model, inverzná distribučná funkcia, kvantilová funkcia, príjmová nerovnosť, kvantily.

JEL classification: J31, C14, C46.

Príspevok je súčasťou riešenia VEGA 01/0127/11: *Priestorová distribúcia chudoby v EÚ*.

1. Úvod

Definovanie modelov v tvare štvor-parametrických a viac-parametrických zovšeobecnených kvantilových funkcií (inverzných distribučných funkcií – inverzií *cdf*s; z angl. *Inverse Cumulative Distribution Functions*, často označovaných aj ako percentilových funkcií; z angl. *Percentile Functions*) umožňuje transformovať komponenty modelu namiesto matematickej transformácie údajov, ako aj robustnú estimáciu parametrov v prípade nedodržania predpokladov klasických parametrických metód. Je využiteľné na pravdepodobnostné modelovanie netypických tvarov rozdelení v prípade, keď nebol nájdený vhodný známy tvar funkcie hustoty (*pdf*) teoretických asymetrických spojitých rozdelení (obsiahnutých v knižniciach štatistických programových balíkov, ako napr. lognormálneho, logistického, gamma, Weibullovo, loglogistického, extrémnych-hodnôt,...) po overení zhody rozdelení viacerými známymi testami dobrej zhody, alebo v prípade veľkých výberov radšej koeficientom korelácie.

Zovšeobecnené kvantilové modely sú spravidla dostatočne elastické na modelovanie veľmi asymetrických nepravidelných tvarov rozdelení a ich problematických dlhých koncov. Ich komplexné tvary najčastejšie obsahujú parameter polohy, parameter rozloženia/stupnice a parametre vlastného (základného) tvaru, z ktorých dva sú často v exponentoch (určujú tvar pravého a ľavého konca). V štrukturálnych a regionálnych štúdiách je výhodné definovanie rovnakého funkčného tvaru pre celú populáciu, ale aj pre jej rôzne volené podsúbory, čo komplexné – zovšeobecnené tvary kvantilových modelov zväčša uspokojivo umožňujú vďaka svojej extrémnej elasticite.

2. Modelovanie inverznými distribučnými funkciami

V teórii kvantilového modelovania sa rozlišuje dvojaký prístup, a to vytváraním nového tvaru pre danú aplikáciu použitím tzv. pravidiel modifikácie kvantilových funkcií (Gilchrist,

2000; Sipková - Sodomová, 2007) alebo presným odhadom parametrov niektorého z už známych, ale vysoko elastických viac-parametrických komplexných tvarov, tzv. zovšeobecnených kvantilových funkcií (z angl. *Generalized Quantile Fnctions*).

Aplikáciou metód Gilchristovho kvantilového modelovania (Gilchrist 1997, Gilchrist 2000, Sipková – Sodomová, 2007) a teórie poriadkových štatistík (z angl. *Order Statistics Theory*, pozri napr. David – Nagaraja, 2003) možno konštruovať ďalšie nové analytické kvantilové tvary príjmových modelov kombinovaním a matematickými transformáciami už známych inverzií *cdf*s. Napríklad pre rozdelenie príjmov domácností SR na základe údajov EU SILC bol konštruovaný Weibullov-Paretov a Gamma-Paretov kvantilový model (Sipková, 2007).

Mnohé novo-konštruované kvantilové tvary však často vzájomne analyticky súvisia a môžu byť špeciálnymi tvarmi zovšeobecnených tvarov (najčastejšie s hodnotami 0, alebo 1 niektorého z ich viacerých parametrov).

Zovšeobecnené tvary kvantilových modelov, ktoré boli už v minulosti použité sú napr. zovšeobecnené lamda rozdelenie – GLD (Ramberg and Schmeiser 1974; Sipková, 2004, 2005, Pacáková – Sipková, 2007), zovšeobecnené Weibullovo rozdelenie (Mudholkar a Kolia, 1994), tiež zovšeobecnené kappa rozdelenie (Hosking, 1994, Tarsitano 2011), alebo zovšeobecnené Paretovo rozdelenie (Stopa, 1990) pre extrémne vysoké hodnoty.

Nie je však jednoduché ich správne „napasovať“, t. j. vhodne odhadnúť ich parametre. Je to hlavne z dôvodu ich extrémnej elasticity. Skúmané sú preto ich možné tvary, oddelene v rôzne volených regiónoch hodnôt ich parametrov v exponentoch. Rozpracované sú metódy odhadu parametrov vysoko elastických štvor-parametrických a päť-parametrických zovšeobecnených tvarov na rôznych základoch, napr. optimalizáciou mier „dobrej zhody“, maximalizáciou koeficienta korelácie empirických a teoretických hodnôt, minimalizáciou štvorca/absulútnej veľkosti distribučných rezíduí, na základe tzv. L-momentov, atď. Metódy estimácie parametrov komplexných a vysoko elastických pravdepodobnostných modelov, ktoré sú definované inverznou distribučnou funkciou s vlastnými parametrami tvaru v exponentoch sú predmetom mnohých vedeckých prác (pre GLD napr. Dudewicz – Karian, 1999; Karian – Dudewicz, 1999, 2000).

Tak, ako estimácia výrazne asymetrických kvantilových tvarov s hrubým/dlhým koncom, tak aj verifikácia ich dobrej zhody v prípade veľkých rozsahov súborov je často problematická. Je známe, že chí-kvadrát test dobrej zhody pri veľmi veľkých výberoch nie je aplikovateľný, preto že takmer vždy pomocou neho dôjde k zamietnutiu modelu. K posúdeniu kvality modelu často slúži koeficient korelácie, pripadne koeficient determinácie.

Linearita usporiadania bodov v *Q-Q* grafoch (*X-Y* graf kvantilov empirického rozdelenia oproti simulovanému rankitu teoretického tvaru, prípadne oproti rankitu mediánov teoretického rozdelenia počítaného cez inverznú beta funkciu) je častou vizualizáciou „dobrej zhody“ pri kvantilovom modelovaní. Vhodná kvantifikácia tejto zhody empirického a modelovaného rozdelenia je pomocou najznámejšej miery linearity, t. j. koeficienta korelácie.

V nasledujúcej časti príspevku priblížime tvar zovšeobecneného kappa kvantilového rozdelenia (GKD). Výsledky pravdepodobnostného modelovania pomocou normovaného tvaru GKD celkového rozdelenia miezd zamestnancov v Slovenskej republike, ako aj v podsúboroch podľa jej ôsmich krajov, názorne prezentujeme priamo v *Q-Q* grafických zobrazeniach v poslednej časti príspevku. Dve miery, miera kvality zhody zovšeobecneného kappa modelu – koeficient korelácie, ako Tarsitanova miera nerovnosti *V-nerovnosť*, smernica v prípade normovaného tvaru GKD, sú uvádzané priamo v *Q-Q* grafoch.

3. Zovšeobecnené kappa rozdelenie

Zovšeobecnený tvar štvor-parametrického kappa rozdelenia (GKD) definoval Hosking [5]. Je lineárno-exponenciálnou transformáciou Gumbelovho-exponenciálneho tvaru a definovaný je kvantilovou funkciou takto:

$$Q(p, l) = l_1 + \frac{l_2}{l_3}\left[1 - \left(\frac{1 - p^{l_4}}{l_4}\right)^{l_3}\right]; \quad 0 < p \le 1,\ l_2 > 0 \qquad (1)$$

Viacero známych rozdelení extrémnych hodnôt je špeciálnym prípadom zovšeobecneného kappa rozdelenia. Napríklad obmenami kappa funkcie podľa (1) získame zovšeobecnené Paretovo rozdelenie, keď $l_4 = 1$; rovnomerné rozdelenie ($l_3 = 1,\ l_4 = 1$); zovšeobecnené logistické rozdelenie ($l_4 = -1$); exponenciálne rozdelenie ($l_3 = 1,\ l_4 = 0$); reflexne-exponenciálne ($l_3 = 0,\ l_4 = 1$) [6].Tvar GKD súvisí aj s Burr-3 ($l_3 < 0,\ l_4 < 0$) a Burr-12 rozdeleniami ($l_3 < 0,\ l_4 > 0$). Zovšeobecnený tvar Weibullovho rozdelenia je obráteným tvarom GKD a opačne [12].

Pearsonove momenty GKD (*r*-tý moment) možno definovať podľa vzťahu:

$$E\left[1 - l_3\left(\frac{X - l_1}{l_2}\right)^r\right] = \begin{cases} \left(\frac{1}{l_4}\right)^{1+rl_3} B(1 + r l_3,\ l_4^{-1}) & \text{pre } l_4 > 0 \\ \left(-\frac{1}{l_4}\right)^{1+rl_3} B(1 + r l_3, -r l_3 - r l_4^{-1}) & \text{pre } l_4 < 0 \end{cases}$$

Všetky štyri momenty sú konečné čísla v prípade keď $l_3 > 0,\ l_4 > 0$. Keď $l_4 < 0$, potom *r*-tý moment je konečné číslo keď $l_3 > -1$ [12].

Preto že GKD obsahuje v sebe viacero často využívaných rozdelení extrémnych hodnôt a je veľmi elastické, jeho parametre musia byt odhadované precízne, s veľkou presnosťou. Najviac rozpracovaná metóda odhadu parametrov GKD je metóda L-momentov. Mudolkar and Hutson (1998) rozšírili metódu L-momentov na nový typ „akoby" momentových odhadov s názvom LQ-momenty. Tvrdia, že LQ-momenty vždy existujú a je ich jednoduchšie určiť, pričom majú podobné vlastnosti ako L-momenty. Ani and Jemain v roku 2006 rozvinuli a vylepšili metódu LQ-momentov pre rozdelenie Extremnych hodnôt typu 1. Nová metóda LQ-momentov pre GKD bola rozpracovaná podľa Ani Sharbi and Abdul Aziz Jemain v roku 2010. Nové prístupy k estimácii štyroch parametrov tohto extrémne elastického rozdelenia sú neustále rozpracovávané.

4. Kappa model rozdelenia miezd – aplikácia

Podľa teórie Amartya Sena (1997) model rozdelenia príjmov má byť definovaný takou analytickou funkciou, ktorá umožňuje nadväzujúce analýzy relatívnej nerovnosti. Najznámejšie miery nerovnosti sú už definované pomocou inverzií *cdf*s, t. j. pomocou tzv. rankitu – stredných hodnôt usporiadaných štatistík. Z viacerých dôvodov tvar modelu rozdelenia príjmov definovaný kvantilovou funkciou, alebo jej novo-konštruovaným tvarom zo známych inverzií *cdf*s pomocou primeraného matematického aparátu, alebo niektorým známym zovšeobecneným tvarom, je vhodným východiskom k aplikáciám v regionálnych štúdiách nerovností príjmov. Naviac totiž umožňuje štúdium extrémov v nadväznosti na Monte Carlo simulácie z odhadnutých kvantilových tvarov.

Konkrétna aplikácia modelovania príjmového rozdelenia pomocou štvor-parametrického zovšeobecneného kappa rozdelenia a regionálne porovnanie pomocou *V-nerovnosti* je na základe 1%-ného náhodného výberu priemerných mesačných miezd

oficiálneho výberového zisťovania hrubých miezd (premenná: hmes_mzda) zamestnancov SR v roku 2010 organizáciou Trexima, s.r.o. Vychádzame z individuálnych údajov zisťovania.

Aplikujeme GKD v normovanom tvare (v základnom tvare s umiestnením stredu do 0 a s jednotkovou variabilitou). Pri štúdiu nerovnosti je kappa rozdelenie bez parametra polohy l_1 a parametra stupnice l_2 (rozloženia, variability) a má len dva parametre l_3, l_4, oba v exponentoch a určujú spoločne šikmosť a špicatosť, a teda aj hrubosť/predĺženosť koncov.

Výpočet stredných hodnôt usporiadaných štatistík tohto tvaru, funkčný tvar tzv. rankitu, predstavuje Dagumov jednomodálny tvar prvého typu, ktorý je konkrétnym tvarom kappa rozdelenia keď $l_3 < 0$, $l_4 < 0$ a $l_2 = l_3(l_4)\hat{}\, l_3$; $l_1 = l_2/l_3)$ [12, vzťah (18)]:

$$E(X_{i:n}) = w_{i:n} = \left[p_i^{-l_4} - 1\right]^{-l_3}, i = 1, 2 \ldots n \qquad (2)$$

kde p_i je proporcionálne postavenie i-tej hodnoty v grafickom zobrazení, tzv. pozícia.

V príspevku aplikujeme pozíciu zobrazovania v *Q-Q* grafoch, ktorú pre asymetrické príjmové rozdelenie odporúča Landwehr (s proporciami 0,35 a 0,65), t. j. pozícia je počítaná podľa vzťahu $p_i = (i - 0{,}35)/n$.

Výsledné hodnoty dvoch parametrov tvaru odhadneme pomocou downhill simplexovej maximalizácie koeficienta korelácie (Tarsitano, str. 19):

$$r(l_3, l_4) = \frac{\sum_{i=1}^{n} \left\{\left[p_i^{-l_4} - 1\right]^{-l_3} - w_n(l_3, l_4)\right\} X_{i:n}}{\sqrt{S_x b_n(l_3, l_4)}} \qquad (3)$$

kde $X_{i:n}; i = 1, 2 \ldots n$ sú poriadkové štatistiky (*Order Statistics*), ktorých náhodnou realizáciou je usporiadaná *n*-tica zistených hodnôt miezd zamestnancov. Do vzťahu (3) ďalej dosadzujeme priemernú hodnotu rankitu

$$w_n = n^{-1} \sum_{i=1}^{n} w_{i:n}, \; i = 1, 2 \ldots n \qquad (4)$$

Súčet štvorcov vzdialeností zistených hodnôt od ich priemeru je $S_x = \sum (X_{i:n} - m_n)^2$.

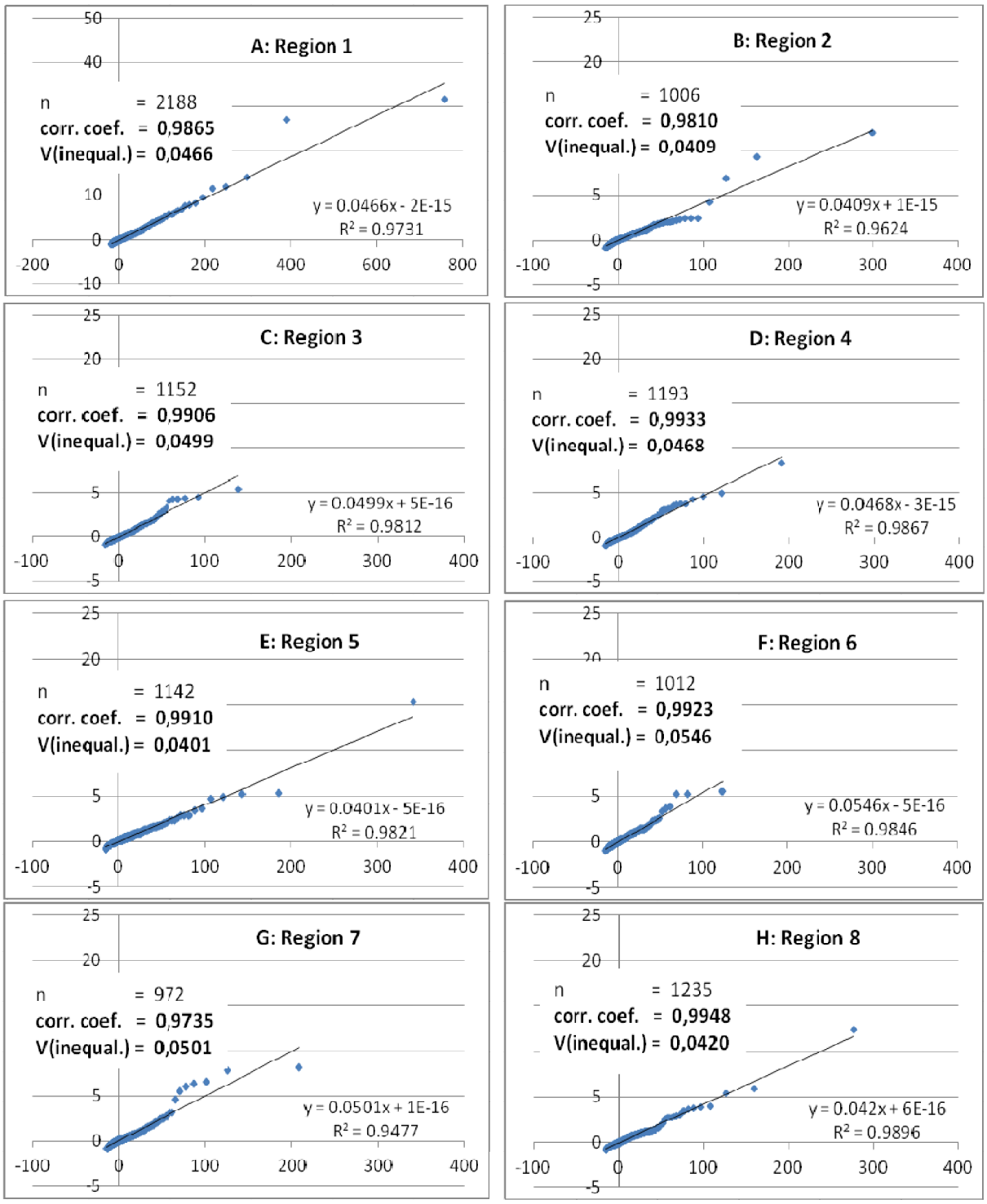
Súčet druhých mocnín vzdialeností hodnôt rankitu podľa vzťahu (2) od ich strednej hodnoty je $b_n(l_3, l_4) = \sum_{i=1}^{n} (w_{i:n} - w_n)^2$.

Odvodenie miery nerovnosti (pozri Tarsitano, 2006, str. 61 až 64, alebo verzia publikovaná na internete str. 8 až str. 12) s označením *V-nerovnosť*:

$$V_n(l_3, l_4) = \sum_{i=1}^{n} \left[\frac{\left[p_i^{-l_4} - 1\right]^{-l_3} - w_n(l_3, l_4)}{\left[p_n^{-l_4} - 1\right]^{-l_3} - w_n(l_3, l_4)} \right] \frac{X_{i:n}}{n m_n} \qquad (5)$$

kde w_n zistíme v súlade so vzťahom (4).

Zámerom Tarsitana bolo zvoliť takú monotónnu transformáciu stupníc v *Q-Q* grafickom zobrazení, aby miera nerovnosti (*V-nerovnosť*) mohla byť získaná ako odhad smernice priamky vhodne preloženej bodmi grafu. V prispevku je aplikovaný normovaný tvar zovšeobecneného kappa modelu. Miera relatívnej príjmovej nerovnosti nie je konštruovaná oddelene od overenia vhodnosti aplikovaného odhadnutého tvaru príjmového modelu. Je získaná odhadom najmenších štvorcov smernice v Tarsitánom doporučovanej lineárnej transformácii vstupujúcich hodnôt do *Q-Q* grafického zobrazenia (pozri Tarsitano, 2006).



Obrázok 1: Q-Q grafy pre kappa modely priemerných hrubých miezd zamestnancov :

A: Bratislavský; B: Trnavský; C: Trenčiansky; D: Nitriansky; E: Žilinský; F: Banskobystrický; G: Prešovský; H: Košický

Zdroj: Vlastné zobrazenie, údaje Trexima, s.r.o.(2010)

Podľa veľkosti Tarsitanom navrhnutej miery nerovnosti je poradie od najväčšej hodnoty po najnižšiu takéto: 6. Banskobystrický, 7 Prešovský, 3 Trenčiansky, 4 Nitriansky, 1 Bratislavský, 8 Košický, 2 Trnavský a 5 Žilinský. Interpretujeme v súlade s teóriou podľa Sena aj ako poradie *„možného vnímania nerovnosti v existujúcej mzdovej distribúcii*“ samotnými zamestnancami v regiónoch SR od najsilnejšieho po najslabšie.

5. Záver

Zovšeobecnený kappa kvatilový tvar funkcie sa ukázal primerane elastický pre regionálnu analýzu mzdových rozdelení v SR (posledné empirické hodnoty v *Q-Q* grafoch ležia pod priamkou). Napriek tomu v regióne 7 koeficient korelácie 0,9735 je nízky a miera *V*-nerovnosti môže byť tým do určitej miery ovplyvnená. Je však otázne, či veľkú menlivosť empirických hodnôt v pravých koncoch mzdových rozdelení je možné lepšie modelovať inou súvislou a relatívne „hladkou" funkciou pre celé rozdelenie. Zvlnený pravý koniec empirického rozdelenia je výsledkom vzájomnej korelovanosti usporiadaných hodnôt výberu.

Literatúra

[1] DAVID, H.A. – NAGARAJA, H.N. 2003. *Order Statistics.* 3rd Edn., John Wiley and Sons, USA. ISBN 0-471-38926-9.

[2] DUDEWICZ, E. J. – KARIAN, Z. A. 1999. Fitting the Generalized Lambda Distribution system by a method of percentiles, *American Journal of Mathematical and Management Sciences*, 19, (1).

[3] FREIMER, M., MUDHOLKAR, G.S., KOLLIA, G., LIN, C.T. 1988. A study of the generalized Tukey Lambda family, In: *Communications in Statistics, Theory and Methods*, 17,(10), 3547-3567.

[4] GILCHRIST, W.G. 2000. *Statistical modelling with quantile functions*, Chapman & Hall. ISBN 1-5848-8174-7.

[5] HOSKING J. R. M. 1994. The four-parameter kappa distribution. *IBM Journal of Research, Development*, 38, 251-258.

[6] PARK, J.S. & PARK, B.J. 2002. Maximum likelihood estimation of the four-parameter Kappa distribution using the penalty method. *Computers & Geosciences* 28: 65-68.

[7] SIPKOVÁ, Ľ. Nový prístup - nové možnosti štatistického modelovania, alebo ako "ušiť" pravdepodobnostný model na mieru, In *Forum statisticum Slovacum*, Bratislava : SSDS, Roč. 3, č. 6, 147-151, 2007. ISSN 1336-7420.

[8] SIPKOVÁ, Ľ. Zovšeobecnené lambda rozdelenie a odhad jeho parametrov, In: *Ekonomika a informatika*, 1/2004, Bratislava, SR, 2004. ISSN 1336-3514.

[9] SIPKOVÁ, Ľ., SIPKO, J. 2012. Analysis of income inequality of employees in the Slovak Republic. In *International days of statistics and economics* : conference proceedings, September 13-15, 2012, Prague, CR. University of Economics Prague. ISBN 978-80-86175-79-9, s. 11.

[10] SIPKOVÁ, Ľ., SIPKO, J. 2012. Regionálne porovnanie nerovnosti miezd zamestnancov SR aplikáciou kappa kvantilových modelov, In: Nerovnosť a chudoba v Európskej únii a na Slovensku: Zborník statí. Košice : Ekonomická fakulta, TU Košice (v tlači).

[11] SIPKOVÁ, Ľ., SODOMOVÁ, E. 2007. *Modelovanie Kvantilovými funkciami*, Vydavateľstvo Ekonóm, Bratislava, SR. ISBN 978-80-225-2346-2.

[12] TARSITANO, A. 2006. A new Q-Q plot and its application to income data, In: *Statistica & Applicazioni*, Vol.IV, special issue n. 1. V upravenej verzii dostupné na internete: http://www3.unisi.it/eventi/GiniLorenz05/ABSTRACT_PAPER_24%20May/PAPER_Tarsitano.pdf

Adresy autorov:

Ľubica Sipková, Ing. PhD.
Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky
Dolnozemská 1, Bratislava
lubica.sipkova@euba.sk

Juraj Sipko, doc. Ing. MBA PhD.
Paneurópska vysoká škola,
Fakulta ekonómie a podnikania
Tematínska 10, 851 05 Bratislava
jsipko@gmail.com

Možnosti testování sezonních jednotkových kořenů demografických časových řad v systému GRETL

The possibilities in testing of seasonal unit roots in demographic time series with GRETL system

Ondřej Šimpach, Petra Dotlačilová, Jitka Langhamrová

Abstract: The aim of this study is to present the options for testing seasonal unit roots in quarterly published demographic time series, in which the presence of seasonality is generally expected. The testing using sophisticated HEGY test will be demonstrated in an econometric package GRETL and with selected demographic time series with quarterly frequency the absence of stationarity will be proved.

Abstrakt: Předkládaná studie představí možnosti testování sezonních jednotkových kořenů u čtvrtletně publikovaných demografických časových řad, u nichž je obecně přítomnost sezonnosti očekávána. Testování s využitím sofistikovaného HEGY testu bude demonstrováno v prostředí ekonometrického balíčku GRETL a na vybraných demografických časových řadách s čtvrtletní frekvencí bude nepřítomnost stacionarity prokázána.

Key words: seasonal unit roots, HEGY test, demographic time series

Klíčová slova: sezonní jednotkové kořeny, HEGY test, demografické časové řady

JEL classification: C22, C52

Úvod

Pro potřeby demografické analýzy je zapotřebí, aby demografická časová řada měla mimo jiné i předpoklad nestacionarity, neboli aby se v příslušné časové řadě vyskytoval trend. Právě přítomnost trendu je jednou z mnoha podmínek, že analyzovaná demografická časová řada může být využita pro modelování a případné predikce. Předkládaný článek pojedná o možnostech exaktního určení, zda příslušná demografická časová řada, publikovaná s čtvrtletní frekvencí, je nebo není stacionární. Obecně neplatí, že každá čtvrtletně publikovaná časová řada je sezonní (viz Arlt, Arltová [1]). Pravděpodobnost, že čtvrtletní demografická časová řada sezonní je, je ovšem vysoká, proto bude představen přístup hledání sezonních jednotkových kořenů v čtvrtletně publikovaných demografických časových řadách, poprvé představen Hyllebergem et al. [6]. Ověření stacionarity je jednou z podmínek, aby s časovou řadou mohlo být následně nakládáno např. dle přístupu autorů Boxe a Jenkinse [3] pro modelování časových řad. Ve zvolených demografických časových řadách figurují *počty sňatků, počty rozvodů, počty živě narozených osob, počty potratů celkem, počty zemřelých celkem, počty přistěhovalých a počty vystěhovalých* a byly publikovány Českým statistickým úřadem (ČSÚ).

1. Metodika

Při pohledu na průběh ekonomické či demografické časové řady je možné vyslovit hypotézu o stacionaritě časové řady. V případě, že trend roste, respektive kolísá a má cykly, je možno tvrdit, že konkrétní časová řada je nestacionární. Pak také autokorelační funkce (ACF) má svou první hodnotu velmi vysokou, blížící se k jedné a ostatní zbylé hodnoty klesají jen velmi pomalu. Grafické zobrazení a definici ACF podává např. Arlt et. al v [2]. V případě, že je časová řada nesezónní a je třeba stacionaritu ověřit exaktně, je možno použít tzv. testů jednotkových kořenů, sestavených Dickeyem a Fullerem v [4]. K následujícím modelům:

a) $X_t = \Phi X_{t-1} + a_t$

b) $X_t = c + \Phi X_{t-1} + a_t$

c) $X_t = c + \Upsilon_t + \Phi X_{t-1} + a_t$

kde X_t je konkrétní časová řada, c je konstanta a a_t je reziduum, se vyslovuje hypotéza:

H_0: $\Phi = 1$, tj. časová řada je typu I(1),

H_1: $\Phi < 1$, je typu I(0).

Použité testové kritérium

$$T = \frac{\widehat{\Phi} - 1}{S_{\widehat{\Phi}}} \tag{1}$$

má t-rozdělení. Však vzhledem k tomu, že testovaná hypotéza je „nestacionarita“, nemá statistika t standardní t rozdělení, ale rozdělení, které bylo odvozené Dickeyem a Fullerem např. v [4]. Ve prospěch alternativní hypotézy svědčí nízké hodnoty testového kritéria t, respektive hodnoty nižší než α procentní kvantil rozdělení D-F. V případě, že v modelech a) b) a c) je zbytková složka autokorelovaná, potom se konstruuje tzv. rozšířený D-F test, který se liší pouze tím, že modely rozšíří o složku $\sum_{i=1}^{p-1} \Upsilon\, \Delta X_{t-i}$, takže

a) $\Delta X_t = \Phi X_{t-1} + \sum_{i=1}^{p-1} \Upsilon\, \Delta X_{t-i} + a_t$

b) $\Delta X_t = c + \Phi X_{t-1} + \sum_{i=1}^{p-1} \Upsilon\, \Delta X_{t-i} + a_t$

c) $\Delta X_t = c + \Upsilon_t + \Phi X_{t-1} + \sum_{i=1}^{p-1} \Upsilon\, \Delta X_{t-i} + a_t$

a $\Delta X_{t-i} = X_{t-1} - X_{t-i-1}$ je rozdíl sousedních hodnot.

V případě, že uvažujeme sezónní časovou řadu, nelze použít přístup testu jednotkového kořene autorů Dickey-Fuller. V současné době je možno použít přístupu, sestaveného autory Hylleberg et. al v [6], označeném jako HEGY, který byl speciálně sestaven pro testování přítomnosti sezónních jednotkových kořenů ve čtvrtletně pozorovaných časových řadách Y_t. Vše spočívá v testování statistické významnosti parametru π_i, kde $i = 1, \dots, 4$ v regresní rovnici, která dle Harveyho a Dijka [7] může mít následnou podobu:

$$\Delta_4 y_t = \mu_t + \pi_1 y_{1,t-1} + \pi_2 y_{2,t-1} + \pi_3 y_{3,t-1} + \pi_4 y_{3,t-1} + \sum_{j=1}^{k} \Phi_j \Delta_4 y_{t-j} + \varepsilon_t \tag{2}$$

kde $t = 1, \dots, T$. Zároveň mějme Δ_k představující tzv. filtr, který může být definován jako

$$\Delta_k y_t \equiv (1 - B^k) y_t \equiv y_t - y_{t-k} \qquad \forall\, k = 1,2, \dots \tag{3}$$

kde B je operátor zpoždění. V regresní rovnici představuje μ_t deterministický trend. Nyní nechť

$$y_{1,t} = (1 + B + B^2 + B^3) y_t \tag{4}$$

$$y_{2,t} = -(1 + B + B^2 + B^3) y_t \tag{5}$$

$$y_{3,t} = -(1 - B^2) y_t \tag{6}$$

a vzhledem k tomu, že $(1 - B^4) = (1 - B) \cdot (1 + B) \cdot (1 + B^2)$, pak y_t může obsahovat sezónní jednotkové kořeny. Uvedené filtry, vedoucí k $y_{1,t}$, $y_{2,t}$ a $y_{3,t}$ odstraňují všechny jednotkové kořeny, kromě jednoho, který vyplývá ze skutečnosti, že roční filtr $(1 - B^4)$ může být rozložen jako

$(1 - B^4) = (1 + B + B^2 + B^3) \cdot (1 - B)$ nebo

$(1 - B^4) = -(1 + B + B^2 + B^3) \cdot (1 + B)$ nebo

$(1 - B^4) = -(1 - B^2) \cdot (1 + B^2)$.

Když se v regresní rovnici (2) parametr

- $\pi_1 = 0$, pak rovnice y_t obsahuje nesezónní jednotkový kořen,
- $\pi_2 = 0$, pak rovnice y_t obsahuje sezónní jednotkový kořen v pololetních četnostech, zmenšených o jednotku,
- $\pi_3 = \pi_4 = 0$, pak rovnice y_t obsahuje sezónní jednotkové kořeny v ročních četnostech $\pm i$, kde $i = 1, \dots, 4$.

Autoři Hylleberg et. al v [6] doporučují použití jednostranných klasických t-testů k určení statistické významnosti parametrů π_1 a π_2 a dále F-testu pro společnou statistickou významnost parametrů π_3 a π_4. V ekonometrickém systému Gretl je možno využít doplňku HEGY pro testování jednotkových kořenů v sezónních řadách. Testovaná hypotéza

H_0: časová řada je stacionární

H_1: non H_0

je zde vyjádřena jako

H_0: parametr z_1 ; z_2 ; z_3 ; $z_4 \neq 0$

H_1: non H_0

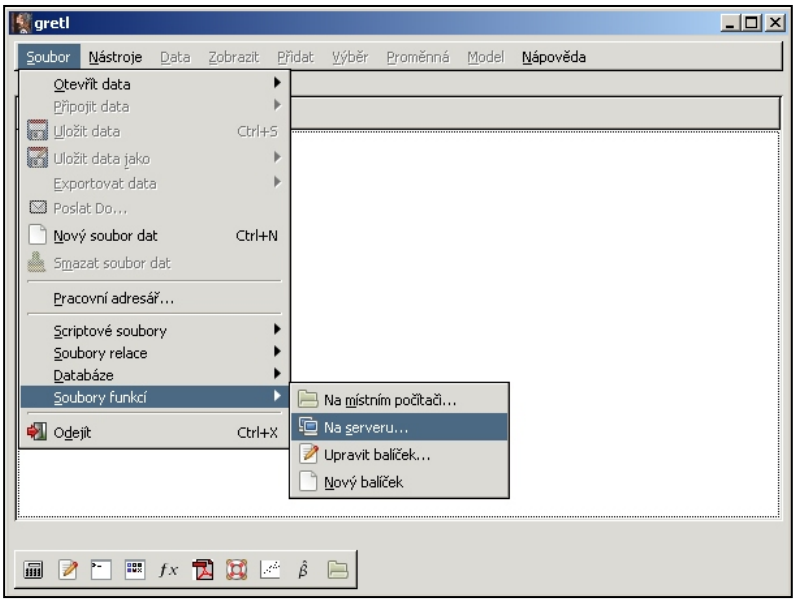
Oproti přístupu autorů Dickeyho a Fullera, kde nulová hypotéza znamená nestacionaritu, je v tomto případě testovaná hypotéza stacionarita. Systém Gretl má parametry π_1 ; π_2 ; π_3 a π_4 označeny jako z_1 ; z_2 ; z_3 a z_4 .

2. Instalace doplňku HEGY do systému GRETL

Bude výhodné ukázat postup, jak doplněk HEGY do systému získat, neboť ověřování stacionarity přístupem autorů Hylleberg et al. se zatím ve standardních balíčcích statistických a ekonometrických systémů příliš nepoužívá.

Krok 1)

Po spuštění systému Gretl vybereme v hlavní nabídce tlačítko *„Soubor“* (v anglické verzi *„File“*), z nabídky zvolíme *„Soubory funkcí“* (v a.v. *„Function files“*), a klikneme na tlačítko *„Na serveru…“* (v a.v. *„On Server…“*), jak ukazuje obrázek 1.

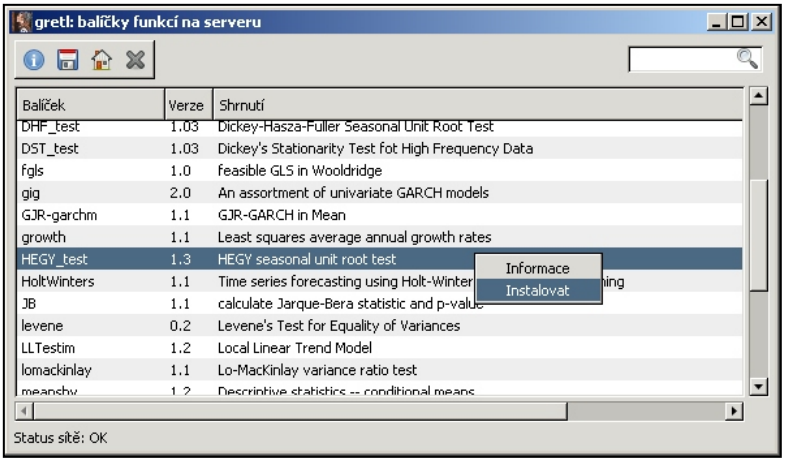


Obr. 1: Hlavní obrazovka systému Gretl

Krok 2)

Zobrazí se nové okno, které je zobrazeno na obrázku 2, ve kterém přibližně v polovině seznamu nalezneme položku pojmenovanou HEGY_test. Klikneme na ni pravým tlačítkem myši a vybereme nabídnutou možnost *„Instalovat“* (v a.v. *„Install“*). Po kliknutí na toto

tlačítko se již nic viditelného nestane. Okno můžeme zavřít a vrátíme se zpět k hlavní obrazovce systému Gretl.



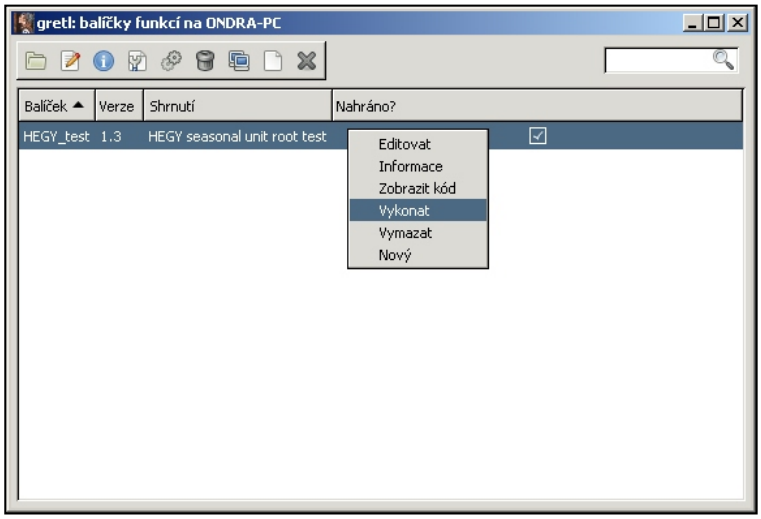
Obr. 2: Balíčky funkcí na serveru

Krok 3)

Po návratu do hlavní obrazovky Gretlu klikneme opět na tlačítko *„Soubor“* (v a.v. *„File“*), z nabídky zvolíme *„Soubory funkcí“* (v a.v. *„Function files“*), a klikneme na tlačítko *„Na místním počítači…“* (v a.v. *„On local machine…“*). Tlačítko je možno vidět i na obrázku 1.

Krok 4)

Zobrazí se nové okno, ve kterém bude seznam nainstalovaných doplňků na lokálním PC. Doplněk HEGY_test se spouští tak, že se na něj klikne pravým tlačítkem myši a vybere se možnost *„Vykonat“* (v a.v. *„Execute“*). Doplněk se spustí pouze v případě, že v základní obrazovce má uživatel otevřená data. Jinak zahlásí chybovou hlášku.



Obr. 3: Spuštění HEGY testu

3. Dosažené výsledky

Ze zvolených demografických časových řad, které byly ČSÚ publikovány s čtvrtletní frekvencí (počty sňatků, počty rozvodů, počty živě narozených osob, počty potratů celkem, počty zemřelých celkem, počty přistěhovalých a počty vystěhovalých), byla ve všech případech zamítnuta nulová hypotéza stacionarity na 5% hladině významnosti. Minimálně jeden z parametrů z_1, z_2, z_3 nebo z_4 byly na 5% hladině významnosti prokázány jako nulové. Výsledky shrnuje tabulka 1.

Tab. 1: Výsledky HEGY testu pro vybrané čtvrtletní demografické časové řady

SNATKY	**koeficient**	**směr. chyba**	**t-podíl**	**p-hodnota**
const	6345,35000	1841,35000	3,446	0,0010
time	-13,8541000	7,14161000	-1,940	0,0565
z1	-0,11011200	0,02968890	-3,709	0,0004
z2	-0,02753800	0,03003090	-0,917	0,3624
z3	-0,00762274	0,00969591	-0,786	0,4345
z4	-0,01323320	0,00961451	-1,376	0,1732
d4y_1	0,42566300	0,09966930	4,271	0,0001
ROZVODY				
const	5053,850	1419,7700	3,560	0,0007
time	-2,044170	4,2410800	-0,482	0,6314
z1	-0,160991	0,0455561	-3,534	0,0007
z2	-0,205903	0,0885868	-2,324	0,0231
z3	0,220502	0,0959667	2,298	0,0247
z4	0,325187	0,0923045	3,523	0,0008
d4y_1	0,190322	0,1228250	1,550	0,1259
ZIVE_NAROZENI				
const	2283,4700	952,52400	2,397	0,0193
time	10,920100	6,0440800	1,807	0,0752
z1	-0,0276901	0,0101182	-2,737	0,0079
z2	-0,1776990	0,0855798	-2,076	0,0416
z3	0,0333594	0,0391601	0,852	0,3973
z4	0,0416267	0,0388963	1,070	0,2883
d4y_1	0,6023090	0,0938876	6,415	0,0000
POTRATY_CELKEM				
const	2951,8100	710,68900	4,153	0,0001
time	-14,220900	4,6117300	-3,084	0,0030
z1	-0,0518291	0,0116865	-4,435	0,0000
z2	-0,1387010	0,0590383	-2,349	0,0217
z3	0,0190186	0,0888773	0,214	0,8312
z4	0,2785010	0,0824795	3,377	0,0012
d4y_1	0,3841340	0,0915528	4,196	0,0001
ZEMRELI_CELKEM				
const	18675,0000	7336,9800	2,545	0,0132
time	-22,428500	12,854700	-1,745	0,0855
z1	-0,1623660	0,0622092	-2,610	0,0111
z2	-0,2576140	0,0776817	-3,316	0,0015
z3	0,0212165	0,0507676	0,418	0,6773
z4	0,1215390	0,0496249	2,449	0,0169
d4y_1	-0,0316371	0,1187440	-0,266	0,7907
PRISTEHOVALI				
const	302,05700	825,39200	0,366	0,7155
time	17,259800	24,308600	0,710	0,4801
z1	-0,0302791	0,0187532	-1,615	0,1110
z2	-0,5104460	0,1236140	-4,129	0,0001
z3	0,2699530	0,1143450	2,361	0,0211
z4	0,4748740	0,1032190	4,601	0,0000
d4y_1	0,1024440	0,1251530	0,819	0,4159
VYSTEHOVALI				
const	188,09000	422,94100	0,445	0,6579
time	3,2107400	10,355200	0,310	0,7575
z1	-0,0257016	0,0176582	-1,456	0,1501
z2	-0,1948100	0,0829772	-2,348	0,0218
z3	0,6054780	0,1136490	5,328	0,0000
z4	0,3641730	0,1274180	2,858	0,0057
d4y_1	0,0978633	0,1227800	0,797	0,4282

Zdroj: vlastní konstrukce

4. Závěr

Výstupy ze softwaru jsou velmi přehledné a poskytují sofistikovaný důkaz o tom, že v demografické časové řadě s čtvrtletní frekvencí (která je navíc předpokládána s přítomnou sezonností) se vyskytuje trend. Po stažení doplňkové funkce HEGY se i po zavření software a ukončení všech dalších úkonů daný doplněk nevymaže. (Vymazání doplňku můžeme provést tak, že v nabídce *„Soubory funkcí“* – *„Na místním počítači“* (v a.v. *„Function files“* – *„On local machine“*), klikne na zvolený doplněk pravým tlačítkem myši a vybereme možnost *„Vymazat“* (v a.v. *„Delete“*), viz obrázek 3). Doplněk je však možné využít pouze pro řady s frekvencí čtvrtletní, nikoliv měsíční. To ovšem není zas až tak velký nedostatek, neboť s měsíční frekvencí se ve statistice příliš mnoho údajů nepublikuje.

Příspěvek byl zpracován v rámci projektu VŠE IGA 29/2011 „Analýza stárnutí obyvatelstva a dopad na trh práce a ekonomickou aktivitu“.

5. Literatura

[1] ARLT, J., ARLTOVÁ, M.: *„Ekonomické časové řady“*, Grada Publishing, 2007.

[2] Arlt, J., Arltová, M., Rublíková, E.: *„Analýza ekonomických časových řad s příklady“*, 2. vyd. Skripta VŠE Praha, 2004.

[3] BOX, G.E.P., JENKINS, G.: „*Time series analysis: Forecasting and control“*, San Francisco, Holden-Day, 1970.

[4] DICKEY, D.A., FULLER, W.A.: „Distribution of the Estimators for Autoregressive Time Series with a Unit Root“, *Journal of the American Statistical Association*, 74, 1979, str. 427–431.

[5] GRANGER, C.W.J.: „Some Properties of Time Series Data and Their Use in Econometric Model Specification“, *Journal of Econometrics* 16, 1981, str. 121-130.

[6] HYLLEBERG, S., ENGLE, R.F., GRANGER, C.W.J., YOO, B.S.: „Seasonal Integration and Cointegration“, *Journal of Econometrics* 44, 1990, str. 215-238.

[7] HARVEY, D.I., VAN DIJK, D.: „Sample size, lag order and critical values of seasonal unit root tests“, *Loughborough University – Institutional Repository*, 2003.
Zdroj: <https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/352/3/erp03-09.pdf>
Publikováno: září 2003

Adresa autorů

Ondřej Šimpach, Ing.
VŠE v Praze, katedra demografie
Nám. W. Churchilla 4, 130 67 Praha 3
ondrej.simpach@vse.cz

Petra Dotlačilová, Ing.
VŠE v Praze, katedra demografie
Nám. W. Churchilla 4, 130 67 Praha 3
xdotp00@vse.cz

Jitka Langhamrová, doc., Ing., CSc.
VŠE v Praze, katedra demografie
Nám. W. Churchilla 4, 130 67 Praha 3
langhamj@vse.cz

Metódy zhlukovej analýzy založené na Bayesovej vete
Cluster analysis method based on Bayesian theorem

Lukáš Sobíšek, Mária Stachová

Abstrakt: Hlavným cieľom nášho príspevku je predstaviť a porovnať viaceré metódy zhlukovania, ktoré sú založené na Bayesovej vete. Zameriavame sa na modely, ktoré sú implementované v softvéroch AutoClass, LatentGold a v štatistickom systéme R.

Abstract: The aim of our contribution is to present and compare different clustering models, which are based on Bayesian theorem. We focus on models that are implemented in AutoClass, Latent GOLD and R software.

Kľúčové slová: AutoClass, Latent Gold, R softvér, zhluková analýza, Bayesova veta

Key words: AutoClass, Latent Gold, R software, cluster analysis, Bayes theorem

JEL classification: C11

1. Úvod

Úlohou zhlukovej analýzy dát, alebo skrátene zhlukovania, je zatriedenie jednotlivých objektov do navzájom disjunktných skupín (zhlukov) tak, aby objekt bol viac podobný ostatným objektom v rámci jedného zhluku, ako objektom vo zvyšných zhlukoch.
Metódy zhlukovania môžeme rozdeliť do dvoch kategórií: modelový prístup *model-based* a prístup založený na matici vzdialenosti/podobnosti *distance-based* (napr. hierarchické zhlukovanie, alebo metóda *k*-means). Metódy založené na matici vzdialenosti/podobnosti popisuje napr. Řezanková [13]. Bayesovské zhlukovanie modeluje dáta ako zmes latentných tried (ďalej iba tried).

Oproti metódam, ktoré priraďujú jednotlivé objekty jednoznačne do tried (ako sú metódy zhlukovej analýzy pre pevné zhlukovanie), bayesovský modelový prístup hľadá v dátach latentné triedy, ktoré čo „najlepšie“ charakterizujú príslušné objekty. Namiesto jednoznačného priradenia je pre objekty odhadnutá pravdepodobnosť príslušnosti k jednotlivým triedam. Pod pojmem trieda sa ukrýva parametrizované pravdepodobnostné rozdelenie.

V príspevku popisujeme všeobecný model latentných tried. Parametre modelu je možné odhadnúť dvomi prístupmi, pomocou naivného bayesovského klasifikátora, alebo pomocou bayesovských sietí [3]. Z dôvodu výrazne nižšej výpočtovej náročnosti je v štatistickom softvéri AutoClass, LatentGold a vo vybraných balíčkoch systému R aplikovaný prvý prístup. Tento prístup v článku popisujeme a ďalej sa zaoberáme špecifikáciami modelov implementovaných vo vymenovaných programov a ich využiteľnosť na rôzne typy dátových množín.

2. Model Latentných tried (LC – Latent Class)

Pravdepodobnostný model Gelman et al. [9] popisujú ako pravdepodobnostnú funkciu alebo hustotu pravdepodobnosti, z ktorej mohli empirické dáta vzniknúť. Pri odhade rozdelenia pravdepodobnosti parametrov modelu podmienenej empirickými dátami X[1] sa vychádza z *Bayesovej vety*

$$P(q|X) = \frac{P(q) \cdot P(X|q)}{P(X)}. \quad (1)$$

[1] Teória je vysvetlovaná pre dáta s jednou premennou.

Aposteriórna znalosť o parametroch modelu, obsiahnutá v aposteriórnom pravdepodobnostnom rozdelení $P(q|X)$, je úmerná súčinu apriórnej znalosti o parametroch a vierohodnosti. Apriórna znalosť o parametroch je vyjadrená pravdepodobnostným rozdelením parametrov $P(q)$. Vierohodnosť $P(X|q)$ obsahuje informácie o parametroch, ktoré sú obsiahnuté v premennej *X*. Ide teda o pravdepodobnosť, že dané dáta vzniknú, podmienená hodnotami parametrov modelu. Hodnotu parametra je možné interpretovať ako oblasť hodnôt premennej, v ktorej dominuje pravdepodobnosť príslušnosti do danej triedy. Marginálna vierohodnosť *P*(*X*) sa nachádza v menovateľovi.

Pre dané *X* existuje *T* potenciálnych modelov *t* = 1, 2, …, *T* s parametrami q_t. Pri odhade najpravdepodobnejšieho modelu, teda modelu s najväčšou aposteriórnou pravdepodobnosťou (MAP – maximum posteriory) θ_{MAP} sa často zo vzorca (1) vynecháva menovateľ. V prípade, že nie je dôvod favorizovať žiadny z modelov, zo vzorce sa vyradí $P(q)$ a $q_t = q_{MAP}$ je model s maximálnou vierohodnosťou

$$P(X|q) = \max_t P(X|q_t). \tag{2}$$

Aposteriórnu pravdepodobnosť pre viac premenných $P(q|X_1, X_2, ..., X_K)$ je možné vypočítať pomocou *naivného bayesovského klasifikátora*, ktorý vychádza z klasického modelu zmesí rozdelení (FMM – *Finite Mixture Model*) a predpokladá podmienenú nezávislosť jednotlivých premenných.

FMM algoritmus hľadá pre každú triedu C_j, j = 1, 2, …, J najpravdepodobnejší vektor hodnôt parametrov $\overset{\blacktriangledown}{q}_j$ a jeho model, teda pravdepodobnostné rozdelenie alebo hustotu pravdepodobnosti $P(X|q_j)$.

Axiómom FMM je *podmienená nezávislosť*[2] jednotlivých premenných X_1, X_2, …, X_K. Podmienenú nezávislosť premenných X_k, k = 1, 2, …, K, je možné vyjadriť zťahom

$$P(\overset{\blacktriangledown}{\mathbf{X}}_i | X_i \in C_j) = \prod_{k=1}^{K} P(X_{ik} | X_i \in C_j), \tag{3}$$

kde $\overset{\blacktriangledown}{\mathbf{X}}_i = \{X_{i1}, ..., X_{iK}\}$ je vektor hodnôt premenných pre *i*-ty objekt (*i* = 1, 2, …, *n*). Kombináciou vzorca podmienenej nezávislosti (3) a vzorca (1) získame FMM, viz Cheeseman [4].

$$P(q|\mathbf{X}) = \frac{P(q)}{P(\mathbf{X})} \cdot \prod_{k=1}^{K} P(X_k|q), \tag{4}$$

kde **X** je dátová matica obsahujúca všetkých *n* objektov. MAP odhad q a $P(X|q)$, potrebné pre špecifikáciu modelu, je oveľa jednoduchšie získať v prípade učenia s učiteľom. Menovateľ je v prípade hľadania MAP odhadov vynechaný. Algoritmus spočíta $P(X|q)$ a $P(q)$ zvlášť v každé triede na základe početnosti výskytov hodnôt metódou maximalizácie vierohodnosti. V prípade učenia bez učiteľa sa pre hľadanie $P(X|q)$ vo viacrozmernom priestore hodnôt $\mathbf{X}=\{x_{11}, …, x_{ik}, …, x_{IK}\}$ používa tzv. *Expectation Maximization (EM) algoritmus*, viz Dempster [5]

EM algoritmus prebieha v dvoch krokoch.

[2] Ak je známa náležitosť všetkých objektov do jednotlivých tried C_j modelu q_t, tieto premenné sú navzájom nezávislé. Premenné môžu byť selektivne korelované v rámci jednotlivých tried.

1. fáza: "E krok" (Expectation - očakávanie): na základe náhodne nastavených parametrov q_j vypočítame očakávané funkcie hustoty pravdepodobnosti $P(X|q)$ pre každý objekt a každý zhluk.

2. fáza: "M krok" (Maximization - maximalizácia): maximalizujeme funkciu vierohodnosti vzhľadom na náhodne zvolené parametre a určí sa θ_{MAP}. Ak došlo pri optimalizácii k zmene parametrov, algoritmus sa vráti do kroku 1.

Výstupom je nájdená množina tried a vektor príslušnosti do jednotlivých tried.

Iným prístupom ako je použitie metód z triedy Markov Chain Monte Carlo (MCMC) algoritmov, ako napríklad Metropolis Hastingsov (MH) algoritmus, alebo Gibbsovo vzorkovanie (GS – Gibbs sampling). Tieto metódy vychádzajú z teórie Markových reťazcov. Markove reťazce sú postupnosti náhodných veličín, ktorých rozdelenie každého elementu závisí na hodnote predchádzajúceho. Ak simuláciou získame výber vektora parametrov, tak potom môžeme ľahko spočítať jeho priemer, rozptyl a podobne. Ich hodnoty budú tým bližšie k správnym hodnotám, čím bude nasimulovaných hodnôt viac. Základ MCMC metód spočíva v tom, že ohraničujúcou vlastnosťou rozloženia hodnôt je rovnaká ako požadované mnohorozmerné rozdelenie. Takže výberový priemer simulovaných dát aproximuje správnu hodnotu aposteriórneho rozdelenia. Takýto spôsob maximalizácie vierohodnosti je efektívnejší ako pri bežných numerických metódach

3. Vybrané modely latentných tried

Modely LC v programe AutoClass

AutoClass [2] využíva normálne rozdelenie pre spojité premenné X a multinomické, Bernoulliho, Poissonovo rozdelenie pre diskrétne premenné X. Softvér odhaduje na základe vierohodnosti, či sa oplatí pridať parameter, teda či má parameter signifikantný prínos pre vysvetlenie dát na úkor zvýšenia počtu parametrov. Z dôvodu výpočtovej náročnosti (súčty a integrály v vzorci 4) program AutoClassu ponúka štyri relatívne jednoduché robustné štatistické modely.

Jednoduchý multinomický model modeluje jednu vierohodnostnú funkciu, ktorá sa riadi multinomickým rozdelením. Parametre sú vektor pravdepodobnosti každej hodnoty X_{ik}. Model je určený pre diskrétne podmienene nezávislé premenné, ktoré môžu obsahovať chýbajúce hodnoty.

Jednoduchý normálny CN model modeluje funkciu vierohodnosti s normálnym rozdelením s parametrami stredná hodnota μ a rozptyl σ^2, kde $P(\mu)$ ~ normálne rozdelenie a $P(\sigma)$ ~ logaritmické rovnomerné rozdelenie. Pripúšťa chybu výpočtu, ale predpokladá, že je konštantná a relatívne malá v porovnaní s variabilitou. Model je aplikovateľný pre spojité premenné bez chýbajúcich hodnôt. Nepriamo ho je možné použiť pre zdola ohraničené spojité premenné po predchádzajúcej logaritmickej transformácii a pre ohraničené spojité premenné po log-odds transformácii.

Viacrozmerný normálny CN model sa riadi viacrozmerným normálnym rozdelením a používame ho pre spojité premenné bez chýbajúcich hodnôt. Je vhodnejší ako jednoduchý normálny CN model v prípade, ak premenné sú korelované. Ak sa predpokladá korelácia premenných, tak sa odporúča vytvoriť viac modelov s vhodnou kombináciou premenných a vybrať model s najväčšou pravdepodobnosťou.

Jednoduchý normálny CM model – Model je možné použiť pre rovnaký typ premenných ako CN model. Oproti predchádzajúcemu modelu sa líši v tom, že umožňuje modelovať chýbajúce pozorovania. Parametrami modelu sú binárna pravdepodobnosť, μ a σ.

Modely bližšie definujú[3] Hanson et al.[11].

Modely LC v programe LatentGold (LG)

Pre kategorické premenné (nominálne a ordinálne) sa modelujú triedy s multinomickým rozdelením, v prípade spojitých premenných sa triedy riadia normálnym rozdelením a pre premenné (*counts*), vyjadrujúce počet výskytu daného javu je predpokladané Poissonove alebo binomické rozdelenie.

Do LG model je možné zahrnúť exogénnu premennú (*covariates*) z_i, ktorá delí objekty do viacerých skupín. FMM pre objekt *i* vyzerá nasledovne:

$$P(\mathbf{x}_i|\mathbf{z}_i) = \sum_{j=1}^{J} P(q_j|z_i) \cdot \prod_{k=1}^{K} P(\mathbf{x}_{ik}|q_j, \mathbf{z}_i). \tag{5}$$

$P(\mathbf{x}_i|\mathbf{z}_i)$ špecifikuje pravdepodobnosť vektora pozorovaní $\boldsymbol{x}_i$ pre daný objekt *i* podmienenou hodnotami $\mathbf{z}_i$. Všeobecne povedané, exogénne premenné môžu mať priamy vplyv na *q* a nepriamy vplyv na $\boldsymbol{x}_i$.

V LG je možné zmierniť požiadavku podmienenej nezávislosti. Nie je nutní modelovať rozdelenie zvlášť pre každé x_k vo vzťahu 3. Pre vzájomne závislé premenné je možné modelovať spojité (*joint*) multinomické alebo viacrozmerné normálne rozdelenie. V prípade, že neuvažujeme vplyv z_i, K=3 a x_1 a x_2 sú navzájom závislé, potom je možné vzorec zapísať nasledovne:

$$\prod_{k=1}^{3} P(x_{ik}|x_i \in C_j) = P(X_{i1,}X_{i2}|X_i \in C_j) \cdot P(X_{i3}|X_i \in C_j), \tag{6}$$

viac Hagenaars [12]. Parametre modelu sa odhadujú metódou maximálnej vierohodnosti podľa vzorca 2. V prípade, že chceme penalizovať niektoré modely (hodnoty parametrov), LG zahrnie do funkcie vierohodnosti apriórne pravdepodobnosti a odhaduje pomocou *posteriorne-modálnej* metódy (Posterior Mode). Na výpočet maximálnej vierohodnosti používa kombináciu algoritmov EM a Newton-Raphson (NR). Z dôvodu stability nájdenia optima LG najskôr spustí niekoľko EM iterácií a vo finálnej fáze, keď je blízko optima, prepne z dôvodu zrýchlenia výpočtu na NR.

Okrem modulu *Cluster*, ktorý umožňuje zhlukovať objekty (učenie bez učiteľa), softvér LG umožňuje zhlukovať premenné pomocou modulu *Dfactor*, či klasifikovať (učenie s učiteľom) premenné použitím modulu *Regression*. Viac o modeloch je možné nájsť v technickom manuáli LG.

Modely LC v balíčkoch štatistického systému R

Balíček *"mclust"*

Medzi pravdepodobnostné zhlukovacie algoritmy patrí aj algoritmus obsiahnutý v balíčku "*mclust*" [6], [7] štatistického systému R. Tento balíček obsahuje okrem nástroja na hierarchické zhlukovanie pre normálne rozdelenia aj EM algoritmus pre viacrozmerné normálne zmesi, ktorý sa dokáže vysporiadať aj s modelom obsahujúcim odľahlé pozorovania. V tomto prípade nezačína EM algoritmus náhodným výberom parametrov na optimalizáciu (priemer, kovariančná matica), ale tieto sú výsledkom hierarchického zhlukovania, ktoré prebehne na dátach ako prvé.

Funkcia *Mclust*(), ktorá je súčasťou tohto balíčka hľadá optimálny zmiešaný model vychádzajúc z rôznych kovariančných štruktúr a rôznych počtov zhlukov. Najlepší model je po-

[3] Parametre T, V, normovanú vierohodnosť a apriórnu znalosť o parametroch.

tom vybraný na základe Bayesovho informačného kritéria (BIC - Bayesian Information Criterion).
Kovariančné matice jednotlivých modelov sa môžu vyjadriť v tvare normalizovanej parametrizácie nasledovne:

$$\Sigma_i = \lambda_i D_i A_i D^T_i \quad (7)$$

$$|D_i| = 1, \lambda_i = |\Sigma_i|^{1/p}, A_i = \mathrm{diag}(\lambda_{i1}/\lambda_i , \ldots , \lambda_{i1}/\lambda_i) \quad (8)$$

Číslo λ_i riadi objem i-teho zhluku, matica A_i jeho tvar a matica D_i jeho orientáciu. Tieto charakteristiky sú väčšinou odhadované z dát, môžu sa meniť cez meniace sa zhluky, alebo môžu byť rovnaké pre všetky zhluky. Všetky možnosti, ktoré zahŕňa balíček *„mclust“* sú uvedené v Tabuľke1. V jednorozmernom prípade rozlišujeme len prípade E – komponenty s rovnakým rozptylom a V – komponenty s rôznym rozptylom. Pre viacrozmerný priestor identifikátory popisujú geometrickú charakteristiku modelu. Napríklad, EVI označuje model, v ktorom objem všetkých zhlukov je rovnaký (E), tvar zhluku sa môže meniť (V) a orientácia je identitou (I), t.j. sú stredovo súmerné podľa stredu súradnicovej sústavy.

Tabuľka 1: Sférické parametrizácie kovariančnej matice Σ_i, ktoré sú momentálne dostupné v balíčku mclust pre HC (hierarchické zhlukovanie) a EM algoritmus.

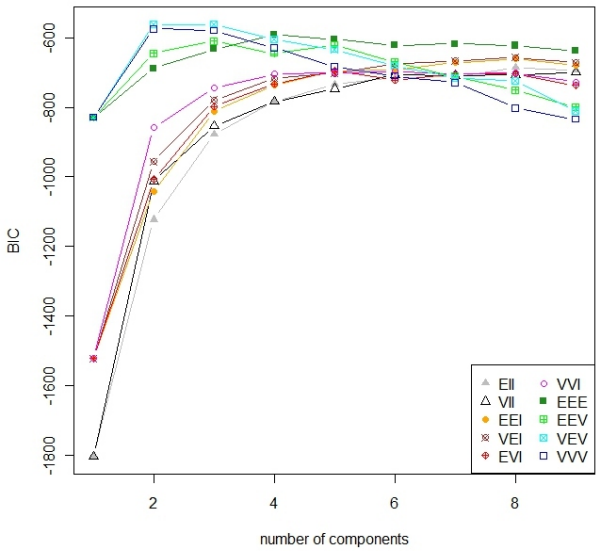
Názov	Model	HC	EM	Rozdelenie	Veľkosť	Tvar	Orientácia
E		+	+	(jednorozmerné)	Rovnaká		
V		+	+	(jednorozmerné)	Rôzna		
EII	λI	+	+	Sférické	Rovnaká	Rovnaký	Nie je
VII	$\lambda_k I$	+	+	Sférické	Rôzna	Rovnaký	Nie je
EEI	λA		+	Diagonálne	Rovnaká	Rovnaký	Osi SS
VEI	$\lambda_k A$		+	Diagonálne	Rôzna	Rovnaký	Osi SS
EVI	λA_k		+	Diagonálne	Rovnaká	Rôzny	Osi SS
VVI	$\lambda_k A_k$		+	Diagonálne	Rôzna	Rôzny	Osi SS
EEE	λDAD^T	+	+	Elipsoidné	Rovnaká	Rovnaký	Rovnaká
EEV	$\lambda D_k A D_k^T$		+	Elipsoidné	Rovnaká	Rovnaký	Rôzna
VEV	$\lambda_k D_k A D_k^T$		+	Elipsoidné	Rôzna	Rovnaký	Rôzna
VVV	$\lambda k D k A_k D k T$	+	+	Elipsoidné	Rôzna	Rôzny	Rôzna

Vstupom do funkcie *Mclust()*, ktorý si môžeme navoliť, je počet zhlukov a výber kovariančnej štruktúry. Primárne nastavených je 1-9 zhlukov a všetky dostupné kovariančné štruktúry. Výstup v sebe zahŕňa parametre modelu, ktorý dosiahol maximálne BIC (vzhľadom na všetky modely a počet zhlukov) a prislúchajúce zaradenie objektov do tried spolu s ich neistotami zaradenia.

Príklad:

```
> library(mclust)
> data = iris[,1:4]
> model1 = Mclust(data)
> BIC = mclustBIC(data,  G=NULL) #výpočet BIC
> BIC
BIC:
         EII        VII        EEI        VEI        EVI        VVI       EEE       EEV       VEV       VVV
1 -1804.0854 -1804.0854 -1522.1202 -1522.1202 -1522.1202 -1522.1202 -829.9782 -829.9782 -829.9782 -829.9782
2 -1123.4115 -1012.2352 -1042.9680  -956.2823 -1007.3082  -857.5515 -688.0972 -644.5997 -561.7285 -574.0178
3  -878.7651  -853.8145  -813.0506  -779.1565  -797.8356  -744.6356 -632.9658 -610.0853 -562.5514 -580.8399
4  -784.3102  -783.8267  -735.4820  -716.5253  -732.4576  -705.0688 -591.4097 -646.0011 -603.9266 -628.9650
5  -734.3865  -746.9931  -694.3922  -703.0523  -695.6736  -700.9100 -604.9299 -621.6906 -635.2087 -683.8206
6  -715.7148  -705.7813  -693.8005  -675.5832  -722.1517  -696.9024 -621.8177 -669.7188 -681.3062 -711.5726
7  -712.1014  -708.7210  -671.6757  -666.8672  -704.1649  -703.9925 -617.6212 -711.3150 -715.2100 -728.5508
8  -686.0967  -707.2610  -661.0846  -657.2447  -703.6602  -702.1138 -622.4221 -750.1897 -724.1750 -801.7295
9  -694.5242  -700.0220  -678.5986  -671.8247  -737.3109  -727.6346 -638.2076 -799.6408 -810.1318 -835.9095
```

```
> plot(BIC, what=BIC)
```



Obrázok 1: grafické znázornenie BIC jednotlivých modelov pre počet zhlukov 1-9 .

```
> print(model1) # zobrazenie najlepšieho modelu
best model: ellipsoidal, equal shape with 2 components
```

Balíček *"bayesclust"*

Funkcia *bayesclust()* je súčasťou rovnomenného balíčka systému R [10], [14]. Tento algoritmus na začiatku testuje existenciu viacerých zhlukov v dátach [8] a následne hľadá ich optimálny počet.

Začíname testovaním hypotéz:

H0: počet zhlukov v dátach je 1
H1: počet zhlukov v dátach je k

a snažíme sa nájsť model s najvyššou pravdepodobnosťou. Používame pritom Bayesovský prístup, pričom na výpočet Bayesovského faktora používame vzťah:

$$BF_{10} = \frac{P(X|H_1)}{P(X|H_0)} \tag{9}$$

kde P(X| H1) označuje pravdepodobnosť rozdelenie dát X do k zhlukov.

Ak označíme θ ako jeden z možných spôsobov rozdelenia dát do zhlukov, tak môžeme tento vzťah prepísať ako:

$$BF_{10} = \sum_{\theta \in S_{n,k}} \frac{P(X|\theta)P(\theta)}{P(X|\theta_1)P(\theta_1)} \tag{10}$$

kde θ_1 označuje prípad, keď počet zhlukov je 1 a $P(\theta)$, $P(\theta_1)$ sú označením pre apriórne pravdepodobnosti rozdelenia θ a θ_1 v tomto poradí. $S_{n,k}$ je množina všetkých spôsobov rozdelení n objektov do k zhlukov.

V prípade, že predpokladáme $P(H_1) = P(H_0) = ½$, tak aposteriórnu pravdepodobnoť hypotézy H_0 vypočítame ako:

$$P(H_0|X) = \frac{1}{1+BF_{10}} \tag{11}$$

Hypotézu H_0 zamietame ak je tento výraz malý.

Bayesov faktor je vzhľadom na jeho výpočtovú náročnosť hľadaý pomocou MCMC metódy.

V prípade, že sa hypotéza H_0 zamieta a teda predpokladáme existenciu viac ako jedného zhluku, pokračujeme hľadaním optimálneho počtu zhlukov. Účelovou funkcou, ktorú pre tento prípad optimalizujeme je marginálna vierohodnosť $P(X|\theta_k)$.

Optimálny počet zhlukov je taký, ktorý maximalizuje túto funkciu pre danú množinu objektov. Spolu s predpokladom, že $P(\theta_k) \propto 1$, je maximalizácia tejto funkcie ekvivalentná s maximalizáciu aposteriórnej pravdepodobnosti premennej θ_k. Znova je pri optimalizácii používaný MCM algoritmus, konkrétne Metropolis-Hastingsova metóda. Tento algoritmus poskytuje výber z pravdepodobnostného rozdelenia za účelom aproximácie rozdelenia odhadovaných parametrov. Takže narozdiel od EM algoritmu, neposkytuje iba bodový odhad neznámych parametrov, ale poskytuje priamo náhodný výber z ich spoločného posteriórneho rozdelenia.

Balíček *"BayesLCA"*

Balíček *"BayesLCA"* [15] v sebe zahŕňa funkciu, pomocou ktorej je možné bayesovsky triediť binárne dáta do latentných tried. V závislosti od nastavení užívateľa je možné použiť viacero algoritmov na maximalizovanie funkcie vierohodnosti, na odhad štandardnej odchýlky, alebo na odhad hustoty hľadaných parametrov. Medzi tieto algoritmy patria: EM algoritmus, rôzne bayesovké algoritmy, Gibbsovo vzorkovanie, alebo boot-straping techniky [1].

4. Záver

AutoClass a LatentGOLD umožňujú kombinovať vstupné premenné spojité a nespojité. Poradia si s chýbajúcimi hodnotami. AutoClass automaticky vyberá počet tried, zatiaľ čo v LG je nutné tento počet stanoviť. Algoritmy v oboch softvéroch vychádzajú z FMM, teda z predpokladu podmienenej nezávislosti. Pri odhade parametrov modelu postupujú iteratívne. AutoClass je zdarma stiahnuteľný z internetu a poradí si aj s objemnými dátami. LG je licencovaný softvér, ktorý má priateľskejšie užívateľské prostredie a oproti AutoClassu naviac aj grafické výstupy, avšak je použiteľný len na menej objemné dátové množiny.

Funkcia *mclust()*, ktorá je súčasťou rovnomenného balíčka systému R si dokáže poradiť aj s objemnými dátovými súbormi, je obmedzená iba vnútornou pamäťou počítača, na ktorom je výpočet vykonávaný (za predpokladu, že používame 64-bit verziu systému R) avšak nedokáže pracovať s kategorickými premennými.

Balíček *"bayesclust"* je alternatívou k balíčku *"mclust"*, dokáže pracovať s dátami, s viacerými premennými a narozdiel od iných algoritmov testuje aj hypotézu, či sa vôbec v daných dátach viac ako jeden zhluk vyskytuje.

Posledným balíčkom je *"BayesLCA"*, tento balíček dokáže pracovať s binárnymi dátami.

Výhodou všetkých balíčkov systému R je, že sú voľne dostupné rovnako, ako celý štatistický systém R.

Literatúra

[1] ASPAROUHOV, T., MUTH´EN, B., 2011. *Using Bayesian priors for more flexible latent class analysis*. Proceedings of the 2011 Joint Statistical Meetings.

[2] AutoClass homepage [online]: http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/. [cit. 14.11.2012]

[3] BERKA, P., 2003. *Dobývání znalostí z databází*. Praha: Academia, ISBN 80-200-1062-9.

[4] CHEESEMAN, P., STUTZ, J., 1996. *Bayesian Classification (AutoClass): Theory and Results*. Advances in Knowledge Discovery and Data Mining, Usama M. Fayyad, Gregory

Piatetsky-Shapiro, Padhraic Smyth, & Ramasamy Uthurusamy, Eds. AAAI Press/MIT Press.

[5] DEMPSTER, A.P., LAIRD, N.V., RUBIN, D.B., 1997. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society. Series B, 39(1), s. 1-38.

[6] FRALEY, CH., RAFTERY, A.E., 2006. MCLUST Version 3 for R*: Normal Mixture Modeling and Model-based Clustering*. Technical Report No. 504, Department of Statistics, University of Washington(revised 2009).

[7] FRALEY, CH., RAFTERY, A.E., 2007. *Model-based Methods of Classification: Using the mclust Software in Chemometrics*. Journal of Statistical Software, 18, 6.

[8] FUENTES, C., CASELLA, G., July 2009. *Testing for the existence of clusters*. Sort (Barc)., 33(2), s. 115–157.

[9] GELMAN, A. ET AL., 2004. *Bayesian Data Analysis*. Boca Raton: Chapman & Hall / CRC, ISBN: 1-58488-388-X.

[10] GOPAL, V., FUENTES, C., CASELLA, G., 2010. *bayesclust: Tests/Searches for significant clusters in genetic data* [online]. Dostupné na: R package, version 3.0, <http://CRAN.R-project.org/package=bayesclust>.

[11] HANSON, R., STUTZ, J., CHEESEMAN, P., 1991. *Bayesian Classification Theory*, Technical Report FIA-90-12-7-01, NASA Ames Research Center, Artificial Intelligence Branch.

[12] HAGENAARS, J.A., 1988. Latent structure models with direct effects between indicators: local dependence models. Sociological Methods and Research, 16, 379-405.

[13] ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V., 2009. *Shluková analýza dat.* 2. vyd. Praha: Professional Publishing, ISBN 978-80-86946-26-9.

[14] VIKNESWARAN, G., FUENTES, C., CASELLA, G., 2012. *bayesclust: An R Package for Testing and Searching for Significant Clusters* [online]. Journal of Statistical Software, 47(14), s. 1-21. Dostupné na internete: <http://www.jstatsoft.org/v47/i14/>.

[15] WHITE, A., MURPHY, B., 2012. *BayesLCA: Bayesian Latent Class Analysis* [online]. Dostupné na: R package version 1.0., <http://CRAN.R-project.org/package=BayesLCA>.

Adresa autorov:

Lukáš Sobíšek, Ing.
VŠE Praha
nám. W. Churchilla 4
130 67 Praha 3, ČR
lukas.sobisek@vse.cz

Mária Stachová, Mgr., PhD.
EF UMB
Tajovského 10
974 00 Banská Bystrica
maria.stachova@umb.sk

Poďakovanie: Tento príspevok bol vypracovaný za podpory projektu GAČR P202/10/0262 a s podporou Univerzitnej grantovej agentúry Univerzity Mateja Bela v Banskej Bystrici v rámci riešenia vedecko-výskumného projektu UGA I-10-005-07.

Medzinárodné porovnávania na základe mikroúdajov EU SILC
International comparisons on base of microdata EU SILC

Iveta Stankovičová, Tomáš Želinský

Abstract: The EU SILC[1] is an important source of comparative statistical data and provides data on income, living conditions, poverty, social exclusion, material deprivation. The surveys are fielded in 29 European countries and coordinated by Eurostat. Although the survey is harmonised, the individual microdata consists of many dissimilarities across participating countries because of different national conditions, methods of data collection and data processing. The aim of this article is to discuss the opportunities and limitations of EU SILC datasets for international comparison of income and poverty.

Abstrakt: Výberové zisťovanie EU SILC patrí medzi európske štatistické zisťovania a zozbierané údaje sú dôležitým zdrojom pre medzinárodné porovnávania. Poskytuje údaje o príjmoch, životných podmienkach, chudobe a sociálnom vylúčení a tiež materiálnej deprivácii. Zisťovanie sa uskutočňuje v 29 európskych krajinách a je koordinované Eurostatom. Hoci ide o harmonizované zisťovanie, individuálne údaje obsahujú veľa odlišností v dátach z jednotlivých krajín spôsobených rozdielnymi národnými podmienkami, metódami zberu a spracovania. Cieľom článku je poukázať na možnosti a obmedzenia údajov EU SILC pri medzinárodných porovnávaniach príjmov a chudoby.

Key words: EU SILC 2010, equivalised disposable income, poverty indicator, median.

Kľúčové slová: EU SILC 2010, ekvivalentný disponibilný príjem, indikátor chudoby, medián.

JEL classification: D63, I30, R11

1. Úvod

Na zasadnutí Európskej rady v Laekene v decembri 2001 predstavitelia štátov a vlád schválili prvý súbor spoločných štatistických ukazovateľov sociálneho vylúčenia a chudoby, ktoré sa stále zdokonaľujú a vyvíjajú. Výberové štatistické zisťovanie EU SILC sa zaviedlo na zabezpečenie podkladových údajov pre tieto ukazovatele. Je organizované v rámci nariadenia č. 1177/2003 Európskeho parlamentu a je povinné pre všetky členské štáty EÚ. Má za cieľ získať aktuálne prierezové i panelové mikroúdaje o príjmoch a sociálnych podmienkach osôb a domácností v krajine.

Údaje EU SILC sú kľúčovým zdrojom informácií nielen pre sociálny výskum, ale aj pre politické rozhodovanie, lebo prináša zlepšenie podmienok pre monitorovanie a reportovanie v sociálnej politike (pravidelnosť, zjednotená metodológia, veľký rozsah premenných, prídavné moduly). Sú zdrojom pre medzinárodné porovnávania a dôležitým rámcom pre hodnotenie efektívnosti sociálnej politiky. Ide o otvorený zdroj údajov: 1. otvorený pre odbornú verejnosť a 2. otvorený kritike a zmenám. V súčasnosti Eurostat opäť pripravuje významné zmeny pre toto zisťovanie (pravdepodobne sa zavedú v roku 2014) a diskutuje o nich s národnými Štatistickými úradmi.

Cieľom tohto článku je poukázať na niektoré zdroje problémov pri medzinárodných porovnávaniach a analýzach na základe údajov EU SILC. Pretože máme k dispozícii celoeurópsku databázu údajov EU SILC 2010 (tj. referenčné obdobie 2009) poskytneme aj výsledky exploratívnej analýzy odvodených premenných HX090 - ekvivalentný disponibilný príjem a HX080 – indikátor chudoby pre domácnosti, ktoré sa v tomto súbore nachádzajú pre 27 európskych krajín.

[1] EU SILC - The European Union Statistics on Income and Living Conditions

2. Metodika zberu údajov EU SILC

Jedným zo základných princípov zisťovania EU SILC je harmonizovaný prístup k zisťovaným údajom v rámci krajín Európskej únie, čo následne umožňuje medzinárodné porovnania a analýzy. V súčasnosti sa EU SILC realizuje vo všetkých 27 štátoch EÚ a okrem toho aj v Nórsku a na Islande a očakáva sa zapojenie Turecka a Švajčiarska. Na európskej úrovni, keďže ide o harmonizované zisťovanie, sa porovnanie údajov uskutočňuje na základe jednotného zoznamu povinných ukazovateľov, ich definícií, jednotných pravidiel, usmernení a postupov pri aplikovaní štatistických metód (váženie a imputácie) a pri výpočte základných indikátorov chudoby.

Zisťovanie EU SILC je síce harmonizované Eurostatom, ale národné štatistické úrady majú určitú voľnosť, napríklad pri organizácii tohto zisťovania, pri spôsobe výberu spravodajských jednotiek do vzorky, ale aj pri spôsobe naplnenia niektorých premenných zisťovania. Týka sa to hlavne výpočtu príjmových premenných. Medzinárodný súbor údajov EU SILC tak obsahuje veľa rozdielov, ktoré majú vplyv na porovnateľnosť údajov.

Tabuľka 1: Spôsoby získania dát EU SILC 2010 (osoby v %)

Krajina	interview	register	obidva	imputácie	nekompletné	nezistené	Spolu
AT	81.16%			0.43%		18.40%	100%
BE	78.55%			1.54%		19.91%	100%
BG	88.41%			0.02%	0.22%	11.35%	100%
CZ	85.17%					14.83%	100%
DE	84.11%				0.56%	15.34%	100%
DK			79.58%			20.42%	100%
EE	82.48%			0.78%		16.74%	100%
ES	82.33%			1.27%		16.40%	100%
FI			80.33%			19.67%	100%
FR	78.79%			0.61%		20.60%	100%
GR	83.97%				0.61%	15.42%	100%
HU	83.40%			0.04%		16.56%	100%
IS		0.31%	76.50%			23.19%	100%
IT	84.88%					15.12%	100%
LT	87.64%			0.05%		12.31%	100%
LU	76.27%					23.73%	100%
LV			84.19%	0.70%		15.11%	100%
MT	83.95%					16.05%	100%
NL	0.18%		77.47%			22.34%	100%
NO	0.00%	1.19%	75.84%			22.98%	100%
PL	76.39%			6.03%		17.59%	100%
PT	85.13%				0.61%	14.27%	100%
RO	88.10%				0.26%	11.64%	100%
SE			80.09%			19.91%	100%
SI		53.78%	31.72%			14.50%	100%
SK	86.52%					13.48%	100%
UK	80.80%					19.20%	100%
Spolu	60.20%	2.90%	19.12%	0.61%	0.08%	17.08%	100%

Zdroj: Vlastný výpočet z EU SILC 2010 podľa premennej RB250 – Status údajov (Data status) zo súboru R.
Poznámka: Stĺpec „obidva“ znamená kombináciu interview a register, v stĺpci „nezistené“ sú zahrnuté neodpovede respondentov v roku 2010.

Vo väčšine krajín sa zisťujú všetky údaje priamo pomocou rozhovorov s respondentmi. Údaje o zložení domácnosti a ďalšie informácie na úrovni domácnosti poskytuje jeden jej

člen. Na príjmové, základné a špecifické premenné o osobách sa opytovatelia pýtajú všetkých členov domácnosti jednotlivo (resp. je prípustné aj „proxy“ vyplnenie, teda vyplnenie dotazníka inou osobou za neprítomného člena domácnosti).

V štátoch, v ktorých majú dobre vybudované registre o obyvateľoch, je možné získať určité informácie z nich alebo aj z rôznych administratívnych zdrojov. Ide hlavne o údaje o príjmoch. Tieto krajiny potom používajú iný spôsob výberu do vzorky. Náhodne vyberajú spravidla jednotlivca a predmetom zisťovania je jeho domácnosť. Údaje o zložení domácnosti a iné premenné na úrovni domácnosti sa dajú získať aj kombináciou rozhovoru a údajov z registrov. Príjmové a základné premenné o jednotlivých členoch domácnosti sú z registrov a na základe nich sa konštruujú agregované premenné za domácnosť. Špecifické údaje o osobách (napr. detailné pracovné charakteristiky, história pracovnej aktivity, informácie o zdraví a bývaní) sa však z registrov nedajú zistiť a musia sa získať priamo od respondentov. *Tabuľka 1* poskytuje prehľad o krajinách, kde ako zdroj údajov používajú registre a kde nie (premenná RB250 – Status údajov). Využívanie informácií z registrov má podstatný vplyv na kvalitu údajov a to hlavne na príjmové premenné, ktoré sú potom presnejšie. Registre využíva 9 európskych krajín. V najvyššej miere je to Slovinsko (SI: 53,78% len z registrov), ale aj Dánsko (DK), Fínsko (FI), Litva (LV), Holandsko (NL), Nórsko (NO), Švédsko (SE) a Island (IS).

Tabuľka 2: Spôsoby zberu dát EU SILC 2010 (osoby v %)

Krajina	PAPI	CAPI	CATI	samovyplnenie	proxy	nezistené	Spolu
AT		43.18%	26.89%		11.09%	18.84%	100%
BE		71.77%			6.78%	21.45%	100%
BG	70.46%				17.96%	11.59%	100%
CZ	46.36%	22.27%		0.06%	16.48%	14.83%	100%
DE				68.30%	15.81%	15.89%	100%
DK			38.02%	2.86%	38.71%	20.42%	100%
EE	1.05%	61.07%	0.29%	0.01%	20.07%	17.52%	100%
ES		50.46%	13.87%		18.00%	17.67%	100%
FI		1.48%	44.54%		34.30%	19.67%	100%
FR		57.00%			21.78%	21.21%	100%
GR	65.49%	6.48%	5.26%	0.03%	6.71%	16.03%	100%
HU	66.86%				16.54%	16.60%	100%
IS			76.50%			23.50%	100%
IT	68.77%				16.11%	15.12%	100%
LT	45.87%		27.68%	0.41%	13.68%	12.36%	100%
LU	60.96%				15.31%	23.73%	100%
LV	2.37%	46.10%	15.70%	0.01%	19.99%	15.83%	100%
MT		59.67%			24.28%	16.05%	100%
NL			76.70%		0.95%	22.34%	100%
NO		0.44%	57.31%		18.07%	24.18%	100%
PL	61.73%				14.66%	23.61%	100%
PT	3.39%	64.32%			17.42%	14.87%	100%
RO	74.60%				13.51%	11.90%	100%
SE	0.15%		78.30%		1.64%	19.91%	100%
SI		10.42%	13.49%		7.82%	68.28%	100%
SK	82.53%			0.29%	3.69%	13.48%	100%
UK		70.60%			8.48%	20.93%	100%
Spolu	26.66%	18.63%	15.32%	3.55%	15.10%	20.73%	100%

Zdroj: Vlastný výpočet z EU SIC 2010 podľa premennej RB260 – Typ opytovania (Type of interview).
Poznámka: V stĺpci „nezistené“ sú zahrnuté neodpovede respondentov v súbore R v roku 2010.

Krajiny si môžu voliť aj spôsob zberu údajov (premenná RB260 – Typ opytovania). V štatistickej praxi sa používa až 5 rôznych spôsobov: 1. osobný rozhovor pomocou papierového dotazníka (PAPI – paper and pen interviewing), 2. osobný rozhovor pomocou elektronického dotazníka (CAPI – computer assisted personal interviewing), 3. osobný telefonický rozhovor (CATI – computer assisted telephone interviewing), 4. ponechanie dotazníka k samostatnému vyplneniu respondentom, alebo 5. tzv. „proxy" vyplnenie, čiže opytovanie so zástupcom[2]. *Tabuľka 2* uvádza prehľad spôsobov zberu údajov v jednotlivých krajinách. Telefonický rozhovor uplatňujú hlavne v krajinách, kde získavajú príjmové údaje z registrov obyvateľstva. Každý spôsob zberu má svoje výhody i nevýhody a môže mať veľký vplyv na kvalitu údajov a ich porovnávanie.

Významnou výhodou CAPI oproti PAPI je, že naprogramované logické kontroly upozornia na nekonzistentnú odpoveď a je možná okamžitá oprava za prítomnosti respondenta. U PAPI sa logické kontroly robia až v neprítomnosti respondenta, a tak nie vždy sú odstránené, lebo sa už nepodarí kontaktovať znovu respondenta. CATI sa využíva hlavne v krajinách, kde sa získavajú základné údaje z registrov a rozhovorom sa zisťuje len časť údajov. Nevýhody samovyplnenia a proxy vyplnenia sú zrejmé a v EU SILC sú požívané ako posledná možnosť (okrem Nemecka).

Zisťovanie EU SILC sa na Slovensku (SK) realizuje formou terénneho zberu dát v náhodne vybraných domácnostiach, kde opytovatelia prevažne prostredníctvom papierových dotazníkov (PAPI) priamo zisťujú požadované údaje od jej členov. Podobne je to aj v ostatných bývalých socialistických krajinách, kde registre nie sú väčšinou dobre alebo vôbec vybudované.

3. Metodika zisťovania príjmových premenných

Príjmové premenné zisťované v EU SILC znamenajú pre užívateľov niekoľko problémov. Ide o ročné príjmy a tie majú vo väčšine krajín časové meškanie v porovnaní s ukazovateľmi práce a aj referenčné obdobia môžu byť iné. Čiže údaje EU SILC sú ideálne na ročné analýzy príjmov, ale nie sú vhodné na analýzy mesačných alebo dokonca hodinových príjmov, resp. miezd.

Príjmové premenné sú zisťované v národných menách a v dátach EU SILC za všetky krajiny sú uvádzané v eurách. Výmenný kurz je určený Eurostatom (premenná HX010, resp. PX010). V roku 2010 Eurostat doplnil v dokumentácii pre užívateľa aj tabuľku s prepočtovými koeficientmi na paritu kúpnej sily (PPP) a tak je potrebné premennú HX010 prepočítať na PPS.

V minulosti mohli byť príjmové premenné uvádzané ako hrubé alebo čisté a tak dochádzalo k problémom pri medzinárodných porovnaniach. Od roku 2007 majú krajiny povinnosť uvádzať hrubé hodnoty pre všetky príjmové premenné, a tak zase chýbajú údaje pre čisté hodnoty v niektorých krajinách. Napríklad v roku 2007 až 10 krajín neuviedlo čisté hodnoty, čo robí problémy pri časových analýzach príjmov. Navyše 2 krajiny (DE a UK) od roku 2007 prestali úplne uvádzať čisté hodnoty pre príjmové premenné.

Problémom je aj to, že niektoré príjmové premenné obsahujú aj záporné hodnoty. Kým u príjmov zo závislej činnosti (PY010) sa záporné hodnoty vyskytujú len zriedka (Holandsko), stratu z podnikania (PY050) povoľujú takmer v dvoch tretinách štátov. Záporné hodnoty pri príjmových premenných sa objavujú hlavne v krajinách, ktoré získavajú údaje z registrov, a to robí problém pri porovnávaní týchto príjmov s krajinami, ktoré získavajú údaje rozhovormi s respondentmi. Záporné a tiež nulové hodnoty sa často vyskytujú v premennej HY010 - celkové hrubé príjmy domácnosti a aj v premennej HY020 - celkový

[2] Proxy vyplnenie nie je vhodné na subjektívne otázky, napr. o zdraví.

disponibilný príjem domácnosti. Podľa Eurostatu je síce výskyt nekladných hodnôt vo vzorkách malý (spolu je to menej ako 0,5% a v žiadnej krajine neprekročil 2%, *Tabuľka 3*), ale aj tak to robí problémy. Spomínané príjmové premenné slúžia ako základ pre výpočet mnohých ukazovateľov chudoby a príjmovej nerovnosti a pri ich konštrukcii delenie nulou je neprípustné z matematického hľadiska. Aj práca s negatívnymi hodnotami spôsobuje rôzne ťažkosti. Výpočet ekvivalentného disponibilného príjmu domácnosti (HX090) v intervale záporných hodnôt stráca význam. Z nižšie uvádzanej tabuľky (*Tabuľka 3*) je zrejmé, že nezáporné hodnoty premennej HX090 mali v EU SILC 2010 len dve krajiny: Rakúsko (AT) a Portugalsko (PT). U ostatných krajín sa záporné a nulové hodnoty vyskytovali, a to najviac v Španielsku (ES: 1,66% zo všetkých domácností vo vzorke údajov), ale aj v Dánsku (DK: 1,07%), v pobaltských krajinách (LT: 1,07%, LV: 0,85%) a Taliansku (IT: 0,74%).

Tabuľka 3: Výskyt nekladných hodnôt v odvodenej premennej HX090 podľa krajín (N - počet domácnosti vo výberovom súbore)

štát	HX090≤0 počet	HX090>0 počet	N počet	HX090≤0 % z N
AT	0	6188	6188	0.00%
PT	0	5182	5182	0.00%
SI	1	9363	9364	0.01%
CZ	2	9096	9098	0.02%
SK	2	5374	5376	0.04%
HU	4	9809	9813	0.04%
PL	9	12921	12930	0.07%
FR	10	11033	11043	0.09%
NO	5	5222	5227	0.10%
FI	11	10978	10989	0.10%
BG	10	6161	6171	0.16%
BE	10	6122	6132	0.16%
DE	23	13056	13079	0.18%
MT	7	3774	3781	0.19%
IS	6	3015	3021	0.20%
NL	24	10110	10134	0.24%
LU	12	4864	4876	0.25%
EE	13	4959	4972	0.26%
SE	22	7151	7173	0.31%
RO	38	7680	7718	0.49%
UK	48	8061	8109	0.59%
GR	51	6954	7005	0.73%
IT	142	19005	19147	0.74%
LV	53	6202	6255	0.85%
LT	57	5257	5314	1.07%
DK	63	5804	5867	1.07%
ES	226	13371	13597	1.66%
Spolu	849	216712	217561	0.39%

Zdroj: Vlastné výpočty z dát EU SILC 2010

Tabuľka 4: Domácnosti ohrozené rizikom chudoby podľa krajín (v %) – usporiadané podľa stĺpca D (premenná HX080=1)

A	B	C	D	E	F	G
štát	1	Spolu	1	Spolu	1	Spolu
CZ	0.20%	1.99%	9.93%	100%	1.13%	1.99%
HU	0.20%	1.82%	11.00%	100%	1.15%	1.82%
NL	0.40%	3.54%	11.32%	100%	2.30%	3.54%
IS	0.01%	0.06%	11.82%	100%	0.04%	0.06%
SK	0.11%	0.92%	11.89%	100%	0.62%	0.92%
FR	1.74%	13.09%	13.28%	100%	9.95%	13.09%
LU	0.01%	0.09%	13.90%	100%	0.07%	0.09%
AT	0.25%	1.74%	14.44%	100%	1.44%	1.74%
NO	0.17%	1.13%	15.33%	100%	1.00%	1.13%
BE	0.35%	2.24%	15.37%	100%	1.98%	2.24%
SE	0.36%	2.16%	16.67%	100%	2.06%	2.16%
MT	0.01%	0.07%	16.84%	100%	0.07%	0.07%
DK	0.23%	1.34%	17.41%	100%	1.33%	1.34%
FI	0.21%	1.21%	17.48%	100%	1.22%	1.21%
PL	1.12%	6.33%	17.66%	100%	6.40%	6.33%
SI	0.07%	0.37%	17.69%	100%	0.38%	0.37%
IT	2.24%	12.08%	18.53%	100%	12.81%	12.08%
EE	0.05%	0.28%	18.61%	100%	0.30%	0.28%
UK	2.34%	12.55%	18.63%	100%	13.39%	12.55%
PT	0.36%	1.89%	19.12%	100%	2.06%	1.89%
DE	3.68%	19.05%	19.32%	100%	21.07%	19.05%
RO	0.71%	3.55%	19.88%	100%	4.04%	3.55%
ES	1.70%	8.21%	20.75%	100%	9.75%	8.21%
GR	0.42%	1.98%	21.06%	100%	2.38%	1.98%
LT	0.14%	0.65%	21.88%	100%	0.81%	0.65%
LV	0.10%	0.41%	23.18%	100%	0.55%	0.41%
BG	0.30%	1.25%	23.89%	100%	1.71%	1.25%
Spolu	17.47%	100.00%	17.47%	100%	100.00%	100.00%

Zdroj: Vlastné výpočty z dát EU SILC 2010

Výpočet odvodených príjmových premenných, ktoré sa už nachádzajú v poskytnutom súbore údajov EU SILC 2010 je nasledovný [6]:

Celkový disponibilný príjem domácnosti (HY020 – Total disposable household income) sa vypočíta ako suma zložiek hrubého osobného príjmu všetkých členov domácnosti, plus zložky hrubého príjmu na úrovni domácnosti (napr. príjem z prenájmu majetku, prijaté transfery od iných domácností), mínus pravidelné dane z majetku, pravidelné platené transfery medzi domácnosťami (napr. výživné, pravidelná peňažná pomoc od iných domácností), daň z príjmu a príspevky na sociálne poistenie. V národných metodikách výpočtu tejto príjmovej premennej sú rozdiely, hlavne v odpočítateľných položkách.

Ekvivalentná škála sa používa na výpočet *ekvivalentnej veľkosti domácnosti (HX050).* V súlade s metodikou Eurostatu sa používa tzv. modifikovaná OECD škála [5], v ktorej sa používa koeficient 1 pre prvého dospelého člena domácnosti, 0,5 pre druhého a každého dospelého člena domácnosti a pre 14 ročných a starších a 0,3 pre každé dieťa mladšie ako 14 rokov.

Ekvivalentný disponibilný príjem (HX090 - Equivalised disposable income) sa vypočíta ako podiel celkového disponibilného príjmu domácnosti (HY020) a ekvivalentnej veľkosti domácnosti (HX050) podľa vzorca: HX090=(HY020*HY025)/HX050. Celkový disponibilný príjem domácnosti (HY020) je však upravený o neodpovede na danú otázku vo vnútri domácnosti (HY025 – *Within household non response inflation factor*). Tento príjem je potom priradený každému členovi domácnosti. Ide o ročnú hodnotu v EUR a pôvodne zistené v národných menách jednotlivých krajín.

Medián ekvivalentného disponibilného príjmu je hodnota ekvivalentného disponibilného príjmu, ktorá rozdeľuje súbor podľa výšky príjmu na dve rovnako početné časti podľa počtu osôb. *Miera rizika chudoby (Risk of poverty rate)* vyjadruje podiel osôb s ekvivalentným disponibilným príjmom pod hranicou 60% národného mediánu ekvivalentného príjmu. *Hranica rizika chudoby* je definovaná ako hodnota 60% mediánu ekvivalentného disponibilného príjmu. Binárna premená HX080 (*Poverty indicator*) v dátach EU SILC nadobúda hodnotu 1 v prípade, že domácnosť je pod takto definovanou hranicou chudoby.

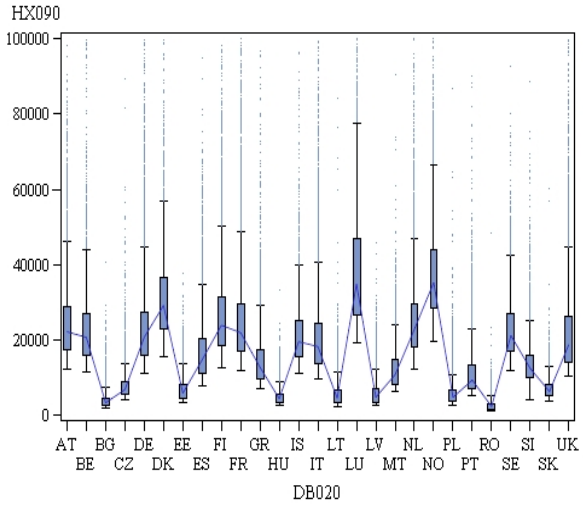
4. Výsledky

Na základe binárnej premennej HX080 – indikátor chudoby je zostavená tabuľka s percentami domácností ohrozených rizikom chudoby v jednotlivých krajinách (*Tabuľka 4*, *Obrázok 3*). Najväčšou krajinou v Európe je Nemecko, kde sa nachádza vyše 19% domácností zo všetkých domácností analyzovaných 27 krajín (*Tabuľka 4*, stĺpec B), o ktorých nám poskytuje údaje súbor EU SILC 2010. Je tam pochopiteľne aj najvyšší počet domácností ohrozených rizikom chudoby (3,68%, *Tabuľka 4*, stĺpec A). Najvyššia miera rizika chudoby v roku 2009 bola v Bulharsku (BG: 23,89%), Litve (LV: 23,18%), Lotyšsku (LT: 21,88%) a Grécku (GR: 21,06%). Najnižšiu hodnotu miery rizika chudoby dosiahla opäť Česká republika (CZ: 9,93%). Slovensko obsadilo piatu priečku rebríčka (SK: 11,89%), čo znamená pre nás zhoršenie pozície, lebo na základe dát EU SILC 2009 sme boli na 3. mieste. Pred Slovensko sa dostali 2 krajiny: Maďarsko (HU: 11%) a Holandsko (NL: 11,32%).

Na základe premenných HX090 a HX080 sme zostrojili box ploty pre 27 krajín Európy (*Obrázok 1*, *Obrázok 2*). Z grafov je zrejmé, že najvyššiu úroveň ekvivalentných disponibilných príjmov v nominálnej hodnote (EUR) dosahujú domácnosti v Luxembursku (LU) a v Nórsku (NO). Naopak najnižšie hodnoty sú v Bulharsku (BG) a Rumunsku (RO) a potom nasledujú všetky krajiny z bývalého socialistického bloku. Na box-plotoch je jasne viditeľný aj výskyt záporných hodnôt v domácnostiach ohrozených rizikom chudoby (*Obrázok 2*: HX080=1). Pre medzinárodné analýzy je potrebné prepočítať nominálne hodnoty HX090 v EUR na PPS.

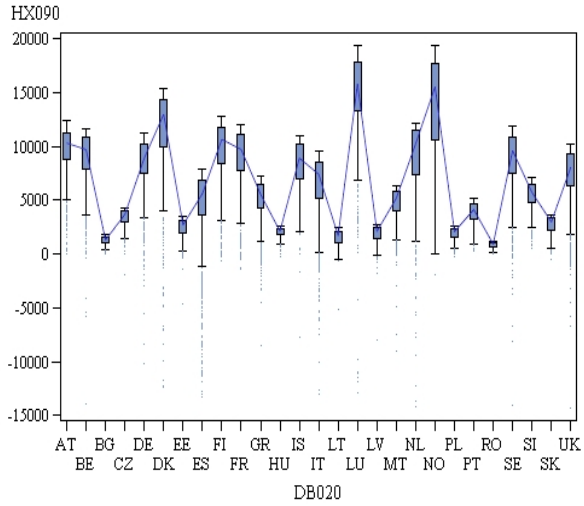
Záverečné tabuľky (*Tabuľka 5*, *Tabuľka 6*) poskytujú prehľad o populáciách počtu domácností v jednotlivých krajinách a základných popisných štatistikách premennej HX090. Zoznam 27 krajín je zoradený podľa nominálnej hodnoty hranice rizika chudoby, čiže podľa mediánu. Najnižšie hodnoty mediánu tejto príjmovej premennej boli v Rumunsku a Bulharsku a naopak najvyššie boli v Luxembursku a Nórsku. Slovensko sa nachádza v našom rebríčku na 8. pozícii.

HX080=0



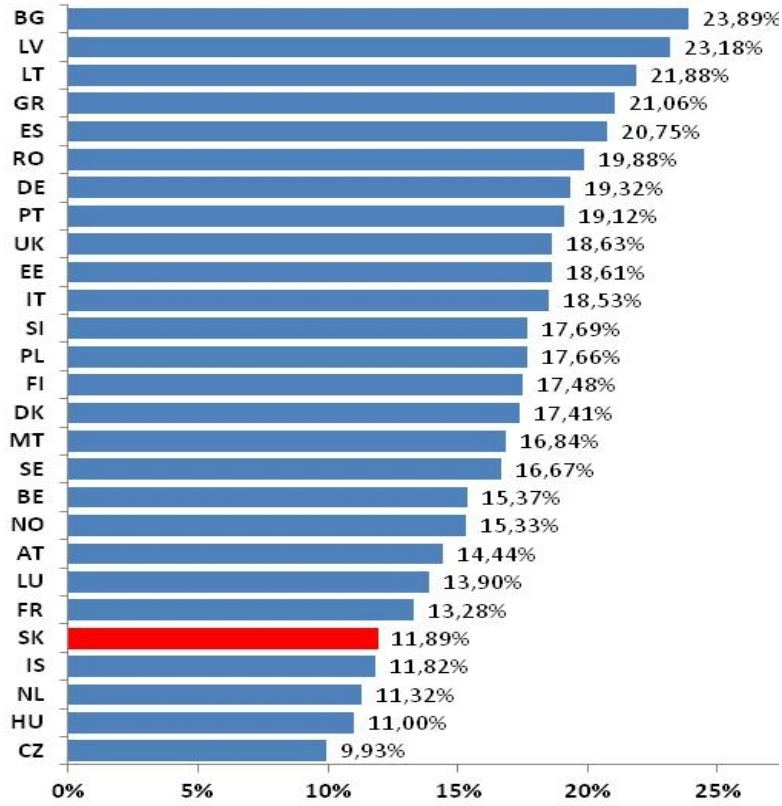
Obrázok 1: Box ploty pre HX090 podľa krajín
Zdroj: Vlastné zobrazenie v SAS EG

HX080=1



Obrázok 2: Box ploty pre HX090 podľa krajín
Zdroj: Vlastné zobrazenie v SAS EG

Poznámka: Pre lepšie grafické rozlíšenie sme nastavili interval hodnôt premennej HX090 pre zobrazenie na vertikálnej osi takto: (-20 000, 100 000). Z maximálnych hodnôt HX090 je zrejmé, že v niektorých krajinách sa vyskytujú aj hodnoty vyššie ako 500 000 Eur (za rok).



Obrázok 3: Miera rizika chudoby v krajinách EÚ (27 krajín)
Zdroj: Vlastné zobrazenie z mikrodát EU SILC 2010

P.č.	Skratka	Krajina
1	BE	Belgicko
2	BG	Bulharsko
3	CZ	Česko
4	DK	Dánsko
5	EE	Estónsko
6	FI	Fínsko
7	FR	Francúzsko
8	GR	Grécko
9	NL	Holandsko
10	IS	Island
11	LV	Litva
12	LT	Lotyšsko
13	LU	Luxembursko
14	HU	Maďarsko
15	MT	Malta
16	DE	Nemecko
17	NO	Nórsko
18	PL	Poľsko
19	PT	Portugalsko
20	AT	Rakúsko
21	RO	Rumunsko
22	SK	Slovensko
23	SI	Slovinsko
24	ES	Španielsko
25	SE	Švédsko
26	IT	Taliansko
27	UK	V. Británia

Poznámka: V mikrodátach EU SILC 2010 poskytnutých Eurostatom chýbajú údaje za 2 krajiny EÚ: Írsko (IR) a Cyprus (CY), ale sú tam údaje za 2 nečlenské krajiny: Nórsko (NO) a Island (IS).

5. Záver

V údajoch EU SILC sa nachádzajú rôzne odlišnosti, ktoré spôsobujú problémy pri medzinárodných porovnávaniach. Sú spôsobné rozdielmi v metodike, spôsobe zberu, ale aj v postupe spracovania údajov v jednotlivých krajinách. Porovnateľnosť sa môže zvýšiť len ďalšou štandardizáciou postupov v jednotlivých krajinách. Eurostat sa stále intenzívne zaoberá zjednocovaním postupov pre toto zisťovanie, organizuje rôzne semináre a konferencie, kde si členské krajiny vymieňajú informácie a skúsenosti z danej oblasti.

Predložený článok predstavuje východiskovú komparatívnu analýzu základných premenných súboru EU SILC 2010 a treba ho považovať za metodický odrazový mostík pre následné zložitejšie medzinárodné analýzy.

Literatúra

[1] BARTOŠOVÁ, J. 2009. *Analysis and Modelling of Financial Power of Czech Households,* APLIMAT – Journal of Applied Mathematics, Vol. 2 (2009), Nr. 3, Slovak University of Technology, Bratislava, 2009, s. 31-36, ISSN 1337-6365.

[2] EUROSTAT. 2010. *Algorithms to compute Social Inclusion Indicators based on EU SILC and adopted under the Open Method of Coordination (OMC).* Working Group meeting "Statistics on Living Conditions", 10-12 May 2010. Luxembourg: Eurostat.

[3] LABUDOVÁ, V. 2012. Miery príjmovej nerovnosti. In: Pauhofová, I., Želinský, T. (eds.): *Nerovnosť a chudoba v Európskej únii a na Slovensku*. Košice: Ekonomická fakulta TUKE. s. 107-112.

[4] MYSÍKOVÁ, M. 2011. *EU SILC a jeho metodologická úskalí: mezinárodní srovnatelnost a příjmové proměnné.* Data a výzkum SDA Info, Vol. 5, No. 2, pp. 147-170.

[5] SIPKOVÁ, Ľ. 2009. Ekvivalentná škála v EU SILC analýzach príjmovej nerovnosti a chudoby. In: Pacáková, V. (ed.): *Štatistické metódy v ekonómii so zameraním na sociálne analýzy*. Bratislava: EKONÓM. ISBN 978-80-225-2704-0. s. 81-126.

[6] ŠÚ SR. 2011. Informatívne správy Štatistického úradu SR. *Zisťovanie o príjmoch a životných podmienkach EU SILC 2010.* Bratislava 7. októbra 2011. Dostupné online (12. 11. 2012): http://portal.statistics.sk/files/Sekcie/sek_600/eu-silc-2010.pdf

Adresy autorov:

Iveta Stankovičová, doc. Ing. PhD.
Katedra informačných systémov FM UK
Odbojárov 10, 820 05 Bratislava
iveta.stankovicova@fm.uniba.sk

Tomáš Želinský, Ing. PhD.
Ekonomická fakulta, TU Košice
Němcovej 32, 040 01 Košice
tomas.zelinsky@tuke.sk

Príspevok bol napísaný s podporou Vedeckej grantovej agentúry MŠ SR a SAV v rámci riešenia vedecko výskumného projektu VEGA 1/0127/11: *Priestorová distribúcia chudoby v EÚ. Mikroúdaje EU-SILC boli poskytnuté na výskumné účely na základe kontraktu no. EU-SILC/2011/33, podpísaného medzi Európskou komisiou, Eurostatom a Technickou univerzitou v Košiciach. Eurostat nenesie žiadnu zodpovednosť za výsledky a závery, ku ktorým autor dospel*

Tabuľka 5: Popisné štatistiky premennej HX090 podľa krajín v súbore dát EU SILC 2010 – usporiadané podľa výšky mediánu (N - počet domácností v krajine, použitá váha DB090)

	Analysis Variable : HX090							
	DB020	N	Mean	Std Dev	Minimum	Maximum	Median	Median*0.6
1	RO	7396731	2416	1639	-59	48527	2063	1238
2	BG	2610459	3403	2640	0	104921	2860	1716
3	LT	1351757	4793	4016	-5213	84331	3822	2293
4	LV	861528	5287	3980	-7976	45740	4199	2520
5	HU	3792559	4676	2438	-214	104280	4249	2549
6	PL	13193890	5220	3745	-259	86583	4383	2630
7	EE	581393	6459	4192	-4641	37823	5267	3160
8	SK	1908933	6655	4228	-515	173046	5918	3551
9	CZ	4145058	7788	4695	-1879	120169	6799	4079
10	PT	3929457	10513	7882	207	90117	8488	5093
11	MT	141687	11697	7193	-9094	150150	10147	6088
12	SI	773124	12051	6048	-29794	88537	11046	6628
13	GR	4121444	13693	9892	-8467	152836	11560	6936
14	ES	17107627	14788	9969	-42600	149385	12955	7773
15	IT	25166288	18254	13145	-13000	571587	15903	9542
16	UK	26157762	20151	15750	-52018	527772	16721	10033
17	IS	123267	20255	12856	-7746	185778	17565	10539
18	DE	39712127	21060	17420	-89078	838842	18223	10934
19	BE	4677103	20946	13968	-33450	504383	18703	11222
20	SE	4489624	20097	11858	-73347	498088	18812	11287
21	NL	7381077	22285	12692	-135455	549589	19831	11899
22	FR	27280696	23528	17369	-87960	422930	19980	11988
23	AT	3621179	22873	12991	67	204627	20267	12160
24	FI	2526062	22645	14654	-703	544886	20307	12184
25	DK	2784164	25534	18642	-397926	518911	23670	14202
26	NO	2362772	33362	19623	-104493	705185	31327	18796
27	LU	193295	37384	24357	-39742	727485	33035	19821

Zdroj: Vlastné výpočty z mikrodát EU SILC 2010.

Tabuľka 6: Priemerné a mediánové ekvivalentné disponibilné príjmy (HX090) podľa krajín pre domácnosti ohrozené rizikom chudoby (HX080=1) a mimo tohto rizika (HX080=0) ***Usporiadané podľa posledného stĺpca – Rozdiel mediánov***

Krajina DB020	N 0	N 1	Mean 0	Mean 1	Rozdiel Mean	Median 0	Median 1	Rozdiel Median
RO	5906543	1470691	2810.6	830.7	1979.9	2407.6	878.4	1529.2
BG	1981634	623602	4078.0	1259.3	2818.7	3416.8	1295.4	2121.3
HU	3375328	417231	5006.0	2006.2	2999.7	4473.4	2110.2	2363.1
LT	1055844	295913	5719.5	1488.0	4231.5	4489.6	1658.6	2831.0
PL	10864450	2329440	5927.4	1919.7	4007.7	4943.5	2057.7	2885.7
LV	661867	199661	6343.4	1784.6	4558.8	5134.0	2018.5	3115.5
EE	473181	108212	7377.0	2446.7	4930.3	6134.6	2748.4	3386.2
SK	1681947	226986	7198.0	2630.5	4567.6	6302.5	2855.9	3446.5
CZ	3733583	411475	8284.0	3288.8	4995.2	7119.4	3524.5	3594.9
PT	3178001	751456	12096.0	3820.3	8275.7	9796.0	4084.3	5711.7
MT	117830	23857	13132.3	4607.0	8525.3	11273.6	5226.9	6046.7
SI	636158	136966	13490.2	5364.8	8125.4	12188.1	5600.0	6588.1
GR	3253665	867779	15988.9	5085.9	10903.0	13584.8	5600.0	7984.8
ES	13558139	3549488	17489.7	4469.4	13020.3	14952.0	5570.0	9382.0
IS	108377	14620	21912.2	7970.6	13941.6	18823.3	8787.5	10035.8
IT	20502010	4664278	20953.9	6387.3	14566.6	18040.7	7350.0	10690.7
BE	3958098	719005	23141.0	8862.7	14278.2	20636.0	9758.4	10877.6
NL	6545244	835833	24037.4	8559.0	15478.4	21105.0	10072.0	11033.0
UK	21283403	4874359	23108.3	7240.5	15867.8	19086.7	8001.5	11085.2
SE	3741206	748418	22437.4	8398.1	14039.3	20637.1	9521.2	11115.8
AT	3098402	522777	25114.2	9586.6	15527.6	21914.5	10293.2	11621.3
FI	2084206	441856	25302.6	10111.2	15191.4	22567.0	10888.0	11679.0
DE	32039134	7672993	24126.6	8255.4	15871.2	20622.0	8736.0	11886.0
FR	23657481	3623215	25796.3	8719.0	17077.3	21510.0	9540.0	11970.0
DK	2299199	484965	29135.2	8460.5	20674.7	26302.0	11932.9	14369.1
NO	2000457	362315	36978.3	13394.7	23583.6	33729.9	15639.6	18090.3
LU	166446	26849	41059.2	14598.8	26460.3	35967.6	15794.0	20173.6
Spolu	171961833	36404240						
Spolu (%)	82.53%	17.47%						

Zdroj: Vlastné výpočty z mikrodát EU SILC 2010.

Vysvetlivky:

- N = počet domácností v populácii danej krajiny
- Mean = priemerná hodnota premennej HX090 v krajine (vážený aritmetický priemer, váha DB090: Household cross sectional weight)
- Rozdiel Mean = Mean 0 – Mean 1
- Median = mediánová hodnota premennej HX090 v krajine (vážený medián, váha DB090)
- Rozdiel Median = Median 0 – Median 1
- 0 = domácnosti mimo ohrozenia rizikom chudoby (HX080=0)
- 1 = domácnosti ohrozené rizikom chudoby (HX080=1)

Využitie kopula funkcií v štatistickom programe R
Copula functions in statistical software R

Gábor Szűcs

Abstract: This article deals with multivariate modeling using copulas in statistical software R. A copula is a function that links univariate marginal distributions to their joint multivariate distribution. Copulas provide a suitable tool to describe the dependence structure among random variables. The article contains definition of most frequently used Archimedean, elliptical and extreme-value copulas as well as features and some implementation particularities of the R package copula.

Abstrakt: Príspevok sa zaoberá multivariačnými kopula-modelmi v rámci štatistického programu R. Kopule sú také funkcie, ktoré spájajú marginálne distribúcie so združeným rozdelením nejakého náhodného vektora. Predstavujú veľmi účinný štatistický nástroj, ktorý sa využíva pri modelovaní štruktúry závislosti medzi náhodnými premennými. V článku sú uvedené najdôležitejšie triedy kopula funkcií a funkčné možnosti balíka s názvom copula, ktorý je súčasťou štatistického softvéru R.

Key words: copula, multivariate distributions, statistical software R, R package

Kľúčové slová: kopula, multivariačné rozdelenia, štatistický program R, balík programu R

JEL classification: C46, C88

Úvod

Kopula funkcie sa stali v poslednom čase veľmi populárnym nástrojom aplikovanej štatistiky, pretože majú široké využitie v rôznych vedeckých odboroch. Začiatky výskumu týchto funkcií siahajú do polovice 20-teho storočia, keď svoje práce publikovali takí významní autori ako M. Fréchet, A. Sklar alebo E. J. Gumbel. Nová vlna výskumu kopúl nastala až v 90-tych rokoch, po rozšírení osobných počítačov. Odvtedy sa aplikujú predovšetkým v ekonómii, bioštatistike, finančnej matematike, poisťovníctve a mnohých ďalších oblastiach.

Výskumné metódy založené na kopulách sú vo väčšine prípadov výpočtovo náročné, pri vytváraní modelov a vyhodnocovaní dátových súborov sa používajú rôzne počítačové (matematické alebo štatistické) programy. Cieľom tohto článku je zhrnúť vlastnosti a aplikácie kopula funkcií, vytvoriť prehľad o najznámejších triedach kopúl a ukázať ich implementáciu v štatistickom programe R.

1. Idea kopula funkcií

Uvažujme základný pravdepodobnostný priestor (Ω, S, P) a jednorozmerné náhodné premenné $X_1, X_2, \mathbf{K}\, X_d$ definované na tomto pravdepodobnostnom priestore. Ak označíme symbolom F združenú distribučnú funkciu náhodných premenných $X_1, X_2, \mathbf{K}\, X_d$, tak platí

$$F(x_1, x_2, \mathbf{K}\, x_d) = P(X_1 \le x_1, X_2 \le x_2, \mathbf{K}, X_d \le x_d). \qquad (1)$$

Z teórie pravdepodobnosti je známe, že združená distribučná funkcia úplne popisuje marginálne charakteristiky a silu závislosti medzi náhodnými premennými $X_1, X_2, \mathbf{K}, X_d$. Ukázalo sa, že v niektorých situáciách by bolo vhodné, keby sme funkciu F rozdelili na dve časti: prvá časť by popísala len štruktúru závislosti medzi $X_1, X_2, \mathbf{K}, X_d$, kým druhá časť by vyjadrila marginálne vlastnosti uvedených náhodných premenných. Práve takáto úvaha vedie

k tzv. **kopula** funkciám. Kopule sú také funkcie, ktoré prepájajú marginálne rozdelenia náhodných premenných s ich združenou distribúciou, pričom popisujú výlučne ich vzájomný vzťah (závislosť). Formálne teda môžeme písať

$$F(x_1, x_2, \mathbf{K}\, x_d) = C(F_1(x_1), F_2(x_2), \mathbf{K}, F_d(x_d)), \tag{2}$$

kde F_i je marginálna distribučná funkcia náhodnej premennej X_i pre $i = 1,2,\mathbf{K},d$ a C je kopula závislá od združenej distribučnej funkcie F. Existencia takejto funkcie C vyplýva zo Sklarovej vety (1959) [9]. Ak navyše F_i pre $i = 1,2,\mathbf{K},d$ sú spojité funkcie, tak kopula C je jediná. Rovnica (2) sa dá prepísať aj do tvaru:

$$C(u_1, u_2, \mathbf{K}\, u_d) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2), \mathbf{K}, F_d^{-1}(u_d)\right), \tag{3}$$

kde $u_i \in (0;1)$ pre $i = 1,2,\mathbf{K},d$. Z rovnice (3) vyplýva, že kopula C je vlastne distribučná funkcia multivariačného rozdelenia pravdepodobnosti na $\langle 0;1\rangle^d$ s rovnomernými marginálnymi rozdeleniami na intervale $(0;1)$.

2. Najdôležitejšie triedy kopúl

Kopula funkcie sa dajú zaradiť do rôznych skupín na základe ich vlastností. Najznámejšie a najpoužívanejšie sú archimedovské kopule, eliptické kopule, archimax kopule a kopule s extrémnymi hodnotami (*Extreme-Value copulas*).

Archimedovské kopule sa dajú vyjadriť v explicitnej forme, ich tvar je preto pomerne jednoduchý. Majú bohaté možnosti na vyjadrenie štruktúry závislosti, a preto sa využívajú v mnohých vedeckých odboroch. Podľa rozmeru ich delíme na bivariačné ($d = 2$) a multivariačné ($d > 2$) archimedovské kopule. Ďalšou dôležitou vlastnosťou týchto kopúl je, že nie sú odvodené z multivariačnej distribučnej funkcie; pri ich konštrukcii sa používa tzv. generátor, ktorý sa označuje symbolom j. Všeobecné vyjadrenie bivariačnej archimedovskej kopule je

$$C(u,v) = j^{-1}(j(u) + j(v)), \tag{4}$$

pre $u, v \in (0;1\rangle$, pričom $j : (0;1\rangle \to \langle 0;\infty)$ je konvexná klesajúca funkcia s vlastnosťou $j(1) = 0$. Najznámejšie triedy archimedovských kopúl sú uvedené v Tabuľke 1.

Tab. 1: Archimedovské kopule

Trieda, parameter	*Generátor*	*Tvar kopula funkcie*
Gumbelova trieda	$j(t) = (-\ln(t))^q$ $q \geq 1$	$C_q^{Gu}(u,v) = \exp\left\{-\left[(-\ln(u))^q + (-\ln(v))^q\right]^{1/q}\right\}$
Gumbelova-Hougardova trieda	$j(t) = (-\ln(t))^q$ $q \geq 1$	$C_q^{GH}(u_1, \mathbf{K}, u_d) = \exp\left\{-\left[\sum_{i=1}^{d}(-\ln(u_i))^q\right]^{1/q}\right\}$
Claytonova trieda	$j(t) = t^{-q} - 1$ $q \geq 0$	$C_q^{Cl}(u,v) = \left[u^{-q} + v^{-q} - 1\right]^{-1/q}$
Cookova-Johnsonova trieda	$j(t) = t^{-q} - 1$ $q \geq 0$	$C_q^{CJ}(u_1, \mathbf{K}, u_d) = \left[\sum_{i=1}^{d} u_i^{-d} - d + 1\right]^{-1/q}$

Frankova trieda	$j(t) = -\ln\left(\frac{\exp(-qt)-1}{\exp(-q)-1}\right)$ $q > 0$	$C_q^{Fr}(u,v) = -\frac{1}{q}\ln\left[1+\frac{(\exp(-qu)-1)(\exp(-qv)-1)}{\exp(-q)-1}\right]$

Poznámka: Gumbelova, Claytonova a Frankova trieda sú bivariačné triedy kopula funkcií. Gumbelova-Hougardova trieda je zovšeobecnením Gumbelovej triedy, kým Cookova-Johnsonova trieda je zovšeobecnením Claytonovej triedy pre rozmer $d > 2$. Ďalšie významné archimedovské triedy sú napríklad Aliova-Mikhailova-Haqova trieda, Gumbelova-Barnettova trieda alebo Joe-ova trieda.

Na rozdiel od archimedovských kopúl **eliptické kopule** sú vyjadrené v implicitnom tvare. Sú odvodené od tzv. eliptických rozdelení pravdepodobnosti, akými sú napríklad Gaussovo normálne rozdelenie a Studentovo t-rozdelenie. Eliptické kopule majú široké využitie predovšetkým v oblasti ekonomickej a finančnej matematiky.

Gaussova normálna kopula (v bivariačnom prípade) má tvar

$$C_r^{Ga}(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2p\sqrt{1-r^2}} \exp\left\{-\frac{s^2-2rst+t^2}{2(1-r^2)}\right\} \mathrm{d}s\,\mathrm{d}t\,, \qquad (5)$$

kde Φ je distribučná funkcia štandardného normálneho rozdelenia a r je Pearsonov korelačný koeficient.

Bivariačná Studentova t-kopula sa dá písať v tvare

$$C_{r,n}^{t}(u,w) = \int_{-\infty}^{t_n^{-1}(u)} \int_{-\infty}^{t_n^{-1}(w)} \frac{1}{2p\sqrt{1-r^2}} \left(1+\frac{s^2-2rst+t^2}{n(1-r^2)}\right)^{-(n+2)/2} \mathrm{d}s\,\mathrm{d}t\,, \qquad (6)$$

kde t_n je distribučná funkcia t-rozdelenia, n je počet stupňov voľnosti a r je Pearsonov koeficient lineárnej korelácie.

Ďalšiu dôležitú skupinu tvoria **kopula funkcie s extrémnymi hodnotami** (*EV copulas*). Tieto kopule vzniknú ako potenciálne limity kopúl, ktoré sú spojené rozdelením maxima nezávislých, rovnako rozdelených náhodných premenných (podrobnejšie viď v [8]). Patrí sem napríklad Tawnova trieda, Galambosova trieda, trieda kopúl s ťažkým pravým chvostom (*Heavy Right Tail copulas*), Hüslerova-Reissova trieda a Studentova t-kopula s extrémnymi hodnotami. EV-kopule sa využívajú hlavne v poistnej a finančnej matematike, napríklad v oblasti riadenia rizík (*risk management*) resp. pri modelovaní extrémnych poistných udalostí.

V neposlednom rade spomenieme aj niektoré ďalšie triedy, ktoré nepatria do predchádzajúcich skupín. Plackettova trieda kopula funkcií bola odvodená od Plackettovej bivariačnej distribúcie (1965). Podobne aj Farlieho-Gumbelova-Morgensternova trieda (*FGM family*) je úzko prepojená s príslušným FGM-rozdelením, ktorá sa používa pri modelovaní a testovaní zoskupení (*tests of association*) a pri štúdiu efektivity neparametrických testov [4].

3. Kopula funkcie v štatistickom programe R

Ako to už bolo spomenuté v úvodnej časti, výkonné osobné počítače značne zjednodušujú komplikované výpočty spojené s kopula-modelmi. Pri kvalitnom vyhodnocovaní údajov, ale aj pri vytváraní a interpretácii modelov, špeciálne matematické resp. štatistické programy hrajú veľmi dôležitú rolu. Jedným z týchto programov je voľne dostupný softvér R [5]. Významnú časť tohto programu tvoria tzv. balíky (*packages*, *libraries*), pomocou ktorých sa dá rozšíriť základná databáza matematických, štatistických a iných funkcií.

Základné triedy kopula funkcií a súvisiace testy sú implementované vo viacerých balíkoch štatistického softvéru R. Pred niekoľkými rokmi sa odštartoval nový projekt s cieľom

zjednotiť tieto knižnice a vytvoriť nový, vylepšený balík kopula-programov v prostredí R. Prvá publikovaná verzia balíka s názvom copula vyšla v roku 2007 a spolupracovali na nej autori I. Kojandovic a J. Yan [10]. Odvtedy sa táto knižnica stále aktualizovala, v každom roku boli pridané rôzne vylepšenia a ďalšie funkčné možnosti. Doposiaľ posledná verzia balíka bola zverejnená 13. augusta 2012 a okrem spomínaných dvoch autorov sa na jeho vypracovaní podieľali aj M. Hofert a M. Mächler. Počas celého procesu sa postupne zlúčili knižnice copulab a nacopula do nového balíka copula.

Používanie balíka vyžaduje aktuálnu verziu softvéru R (>= 2.14.2). Po spustení balíka sa automaticky aktivujú aj ďalšie knižnice ako napríklad stats4, graphics, gsl, ADGofTest, mvtnorm, pspline a iné. Balík copula obsahuje nasledovné funkcie, metódy a testy:[1]

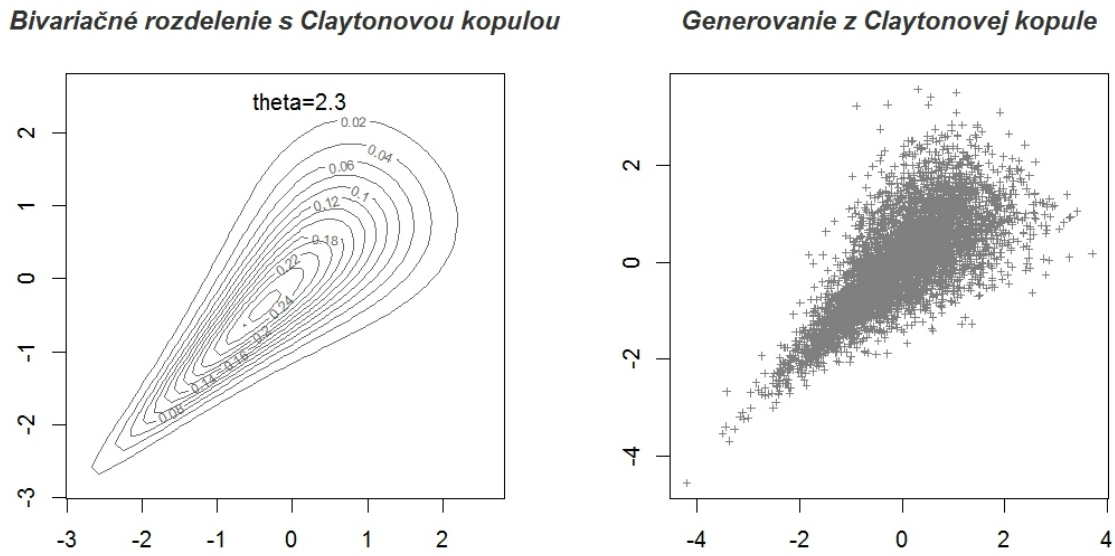
- archimedovské kopule: claytonCopula, frankCopula, gumbelCopula, amhCopula, joeCopula,
- eliptické kopule: normalCopula, tCopula,
- kopule s extrémnymi hodnotami: galambosCopula, huslerReissCopula, tawnCopula, tevCopula,
- ďalšie triedy: plackettCopula, fgmCopula,
- funkcie pre hustotu (dCopula), distribučnú funkciu (pCopula) a generátor náhodných čísel (rCopula) pre všetky spomínané triedy,
- metódu pre fitovanie kopula-modelov: fitCopula,
- testy dobrej zhody, testy zameniteľnosti, testy nezávislosti, test sériovej nezávislosti,
- bivariačný a multivariačný test závislosti na extrémnych hodnotách,
- a mnohé ďalšie.

V nasledujúcej časti príspevku ukážeme konkrétne aplikácie niektorých funkcií z balíka copula. Vysvetlenie a komentáre k jednotlivým príkazom uvedieme v takom tvare, aby celý skript bol vykonateľný v štatistickom softvéri.

```
# načítanie knižnice copula
library(copula)
# definovanie (bivariačnej) Claytonovej kopule
kop.clayton <- claytonCopula(dim=2, param=2.3)
# zadanie združeného rozdelenia: Claytonova kopula a normálne marginálne rozdelenia
mvr.clayton <- mvdc(copula=kop.clayton, margins=c("norm", "norm"),
paramMargins=list(list(mean=0,sd=1), list(mean=0,sd=1)))
# generovanie náhodných čísel z Claytonovej kopule
gen.clayton <- rMvdc(mvr.clayton, n=5000)
# grafické znázornenie kontúr bivariačného rozdelenia a zobrazenie generovaných čísel
par(mfrow=c(1,2))
contour(mvr.clayton, dMvdc, xlim=c(-2.8, 2.6), ylim=c(-2.8, 2.6),
col=rgb(0.4,0.4,0.4), lwd=1.5, xlab="")
title("Bivariačné rozdelenie s Claytonovou kopulou", font.main= 4,
col.main=rgb(0.2,0.2,0.2), cex.sub=1, font.sub=2, col.sub="black")
legend("top",legend="theta=2.3", bty="n")
plot(gen.clayton, type="p", pch=3, cex=0.5, col=rgb(0.5,0.5,0.5), main="",
```

[1] Vypracované na základe oficiálnej príručky balíka copula, ktorá je dostupná na adrese http://cran.open-source-solution.org/web/packages/copula/copula.pdf

```
xlab="", ylab="")
title("Generovanie z Claytonovej kopule", font.main=4,
col.main=rgb(0.2,0.2,0.2))
```



Obr. 1: Claytonova kopula s parametrom $q = 2{,}3$

V poslednej časti príspevku ukážeme ako sa hľadá vhodná kopula pre konkrétne dáta pomocou testu dobrej zhody `gofCopula`. Dáta budeme umelo generovať, maticu „pozorovaní“ vytvoríme ako zmes náhodne generovaných čísel z Gaussovej a Gumbelovej-Hougardovej kopule v pomere 80%:20%. Náš postup je uvedený v nasledovnom skripte:[2]

Zadefinujeme dve pomocné multivariačné kopule (d=3) a príslušné multivariačné združené rozdelenia, pričom marginálne distribúcie budú v každom prípade normálne rozdelenia.

```
kop.normal <- normalCopula(dim=3, dispstr="ex", param=0.4)
kop.gumbel <- gumbelCopula(dim=3, param = 1.9)
mvr.normal <- mvdc(copula=kop.normal, margins = c("norm", "norm", "norm"),
paramMargins = list(list(mean=0,sd=4), list(mean=0,sd=3),
list(mean=0,sd=2)))
mvr.gumbel <- mvdc(copula=kop.gumbel, margins = c("norm", "norm", "norm"),
paramMargins = list(list(mean=0,sd=4), list(mean=0,sd=3),
list(mean=0,sd=2)))
```

umelo vytvoríme dáta z Gaussovej a Gumbelovej; dáta z normálnej kopule s parametrom $\rho = 0{,}4$ kontaminujeme dátami z Gumbelovej kopule s parametrom $q = 1{,}9$

```
x<-matrix(c(rMvdc(mvr.normal, n=400), rMvdc(mvr.gumbel, n=100)), ncol=3)
```

zadáme hypotetickú Gaussovu kopulu s parametrom $\rho = 0{,}3$

```
kop.hypo <- normalCopula(dim=3, dispstr="ex", param=0.3)
```

spustíme test dobrej zhody a overíme, či naše dáta pochádzajú z Gaussovej kopule s parametrom $\rho = 0{,}3$

```
gofCopula(copula=kop.hypo, x=x, N=100, method="Sn", simulation="pb")
```

výstup funkcie, výsledok: `statistic=0.0201, parameter=0.045, p-value=0.8465`

[2] Podrobnejšie viď v oficiálnej príručke balíka `copula`, ktorá je dostupná na adrese http://cran.open-source-solution.org/web/packages/copula/copula.pdf

Výsledok testu naznačuje, že dáta by mohli pochádzať z normálnej kopule s parametrom $\rho = 0{,}3$, ale zároveň funkcia vypočítala aj vlastný odhad parametra ρ, ktorý je $\hat{\rho} = 0{,}045$.

4. Záver

Balík `copula` v štatistickom programe `R` obsahuje významný počet implementovaných kopula funkcií, najdôležitejšie štatistické testy súvisiace s kopulami a ďalšie užitočné funkcie, ktoré nám môžu pomôcť napríklad pri fitovaní viacrozmerných dátových súborov. V rámci tohto príspevku sme aspoň čiastočne ukázali, ako sa pracuje s kopula-modelmi v prostredí softvéru `R`.

Literatúra

[1] BERG, D. 2008. *Using and selecting among copulae*. Oslo: University of Oslo & Norwegian Computing Center. 2008.

[2] HOFERT, M. - MÄCHLER, M. 2011. Nested Archimedean Copulas Meet R: The nacopula Package. *Journal of Statistical Software* 39(9), s. 1–20. http://www.jstatsoft.org/v39/i09/.

[3] KOJADINOVIC, I. - YAN, J. 2010. Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. *Journal of Statistical Software* 34(9), s. 1–20. http://www.jstatsoft.org/v34/i09/.

[4] NELSEN, R. B. 2006. *An Introduction to Copulas*. Springer Science+Business Media, Inc. Second Edition. 2006. ISBN-10: 0-387-28659-4.

[5] R CORE TEAM. 2012. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[6] RONCALLI, T. 2000. *Financial Applications of Copulas*. CREREG, Rennes, France. 2000.

[7] SANDSTRÖM, A. 2011. *Handbook of solvency for actuaries and risk managers: Theory and practice.* Chapman & Hall/CRC. Taylor & Francis Group. 2011. s. 167-194. ISBN 978-1-4398-2130-5.

[8] SEGERS, J. 2004. *Non-Parametric Inference for Bivariate Extreme-Value Copulas*. CentER Discussion Paper 91, 2004. SSRN: http://ssrn.com/abstract=61661.

[9] SKLAR, A. W. 1959. *Fonctions de repartition en dimensions et leurs marges*. Publications de l'Institut de Statistique de l'Universit ´ede Paris, 8, s. 229–231.

[10] YAN, J. 2007. Enjoy the Joy of Copulas: With a Package copula. *Journal of Statistical Software*, 21(4), s. 1–21. http://www.jstatsoft.org/v21/i04/.

Adresa autora:

Gábor Szűcs, Mgr.
Katedra aplikovanej matematiky a štatistiky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave
Mlynská dolina, 842 48 Bratislava 4
szucs@fmph.uniba.sk

Práca bola podporená grantom VEGA č. 2/0038/12.

Dagumovo a Singh-Maddalovo rozdelenie pre modelovanie príjmov
Dagum and Singh-Maddala distribution in income distribution modelling

Alena Tartaľová

Abstract: In this paper the income distribution in Slovakia was studied. Lognormal, Dagum and Singh-Maddala distribution are fitted into data of equalised household income from the survey Statistics of Income and Living Conditions (EU SILC). The obtained estimates are compared with the use of Akaike information criterion (AIC). According to AIC the model of Dagum distribution a Singh-Maddala distribution provides better fit than classical lognormal distribution. The estimated and sample quantiles are calculated in order to compare three different models.

Abstrakt: V príspevku sa venujeme modelovaniu príjmov na Slovensku. Model lognormálneho, Dagumovho a Singh-Maddalovho rozdelenia je použitý na odhad hustoty rozdelenia ekvivalentného ročného príjmu domácnosti z databázy EU SILC. Získané odhady sú porovnané s využitím Akaikeho informačného kritéria(AIC), podľa ktoré je Dagumovo alebo Singh-Maddalovo rozdelenie vhodnejšie než lognormálne. Tiež sú vypočítané a porovnané empirické a odhadnuté kvantily na základe troch modelov.

Key words: density estimation, dagum distribution, singh-maddala distribution, lognormal distribution, the income of households, EU SILC

Kľúčové slová: odhad rozdelenia, dagumovo rozdelenie, singh-maddalovo rozdelenie, lognormálne rozdelenie, príjem domácnosti, EU SILC

JEL classification: C13**,** C16 , O15

1. Úvod

Analýza a modelovanie rozdelenia príjmov domácností je dlhodobo aktuálnou a populárnou témou. Na modelovanie príjmov sa používajú rôzne metódy a analyzuje sa príjem z rôznych uhlov pohľadu. O aktuálnosti tejto témy svedčí aj množstvo zahraničných a domácich publikácií [1],[2],[7],[11],[13],[14]. Znalosť pravdepodobnostného modelu nám poskytne všetky informácie o rozdelení príjmov. Niekoľko užitočných informácií je možno získať aj priamo z odhadnutých parametrov rozdelení, ktoré reprezentujú základné charakteristiky ako napríklad priemer, či smerodajnú odchýlka. Pomocou vhodne odhadnutého rozdelenia pravdepodobnosti potom môžeme odpovedať aj na zložitejšie otázky týkajúce sa rizika chudoby, či merať nerovnosť rozdelenia príjmov využitím indexov nerovnosti napr. Giniho index. Znalosť rozdelenia nám umožní výpočet všetkých dôležitých charakteristík základného súboru, kvantilov, pravdepodobností ľubovoľných intervalov hodnôt apod. Z teórie pravdepodobnosti poznáme dva základné prístupy k odhadu hustoty výberových údajov: Parametrický prístup, ktorý je v súčasnosti najviac používaný a menej známy neparametrický prístup, ktorý sme podrobnejšie popísali v [11].

Ako model pravdepodobnostného rozdelenia príjmov sa v literatúre používajú rôzne typy rozdelení. Medzi tie najčastejšie patria Paretovo a lognormálne rozdelenie s dvomi resp. tromi parametrami. Cieľom tohto príspevku, je predstaviť použitie Dagumovho a Singh-Maddalovho rozdelenia na modelovanie príjmov a poukázať na výhody resp. nevýhody použitia týchto rozdelení oproti tým, ktoré sa bežne používajú.

2. Pravdepodobnostný model rozdelenia

Pravdepodobnostný model rozdelenia sledovanej veličiny umožňuje aproximáciu a zjednodušenie často komplikovaného výberového rozdelenia. Keďže chýbajú logické kritériá, ktoré by viedli k voľbe určitého typu rozdelenia, tak ako najvhodnejší model sa volí ten, ktorý maximalizuje zhodu empirického a teoretického rozdelenia. Modelovanie distribučnej funkcie príjmu má dlhú históriu. Ako prvý začal modelovať rozdelenie príjmov VilfredoPareto ešte v roku 1985. Vtedy vynašiel Paretovo rozdelenie, ktoré môže byť prvého, druhého alebo tretieho resp. štvrtého typu (podľa [3]), no najčastejšie (aj vo väčšine štatistických programov) je v tvare s dvoma parametrami:

$$F(x,s,a)=\left(\frac{x}{s}\right)^{=a},\ x>s\ . \tag{1}$$

Ďalšími najčastejším modelmi sú lognormálne rozdelenie, exponenciálne, gama alebo tiež Weibullovo rozdelenie. Menej známymi rozdeleniami sú Dagumovo alebo Singh-Maddalovo rozdelenie, ktoré sú známe najmä z aplikácií v aktuárstve, ale ich použitím pre modelovanie príjmov možno dostať taktiež veľmi dobré odhady neznámych charakteristík (podľa [3]).

Dagumovo rozdelenie

Dagum vytvoril model pravdepodobnostnej funkcie na základe pozorovania, že elasticita príjmu , založená na distribučnej funkcii F, je klesajúcou a ohraničenou funkciou F. Analytický tvar distribučnej funkcie je možné získať riešením diferenciálnej rovnice:

$$h(F,x)=\frac{d\log F(x)}{d\log x}=ap\left\{1-\left[F(x)^{1/p}\right]\right\},\ x\geq 0\,, \tag{2}$$

kde $p>0$ a $ap>0$, z toho dostávame

$$F(x)=\left[1+\left(\frac{x}{b}\right)^{-a}\right]^{-p} \tag{3}$$

Parametre a, p sú parametre tvaru rozdelenia parameter b je tzv. parameter škály.

Dagumovo rozdelenie, má viacero tvarov, ten ktorý sme uviedli my, je typu I, ale je možné nájsť aj iný tvar s viacerými parametrami, typu II alebo typu III, pozri [6]. Dagum odvodil rozdelenie na základe experimentu s loglogistickým rozdelením, preto pre parameter $p=1$ dostávame práve loglogistické rozdelenie. Okrem toho, Dagumovo rozdelenie dostaneme aj pridaním ďalšieho parametra k Burrovmu rozdelenie. Všetky spomínané rozdelenia patria do systéme zovšeobecneného rozdelenia druhého typu, tzv. GB2 systém, pozri [3].

Singh-Maddalovo rozdelenie

Aj Singh-Maddalo(SM) rozdelenie patrí do systému rozdelení GB2, preto je možné odvodiť vzťah medzi Dagumovým a SM rozdelením

$$X\sim D(a,b,p)\Leftrightarrow\frac{1}{X}\sim SM(a,1/b,p) \tag{4}$$

Na základe tohto vzťahu je zrejmý názov Inverzné Burrovo rozdelenie, pod ktorým je SM rozdelenie známe v aktuárskej literatúre.

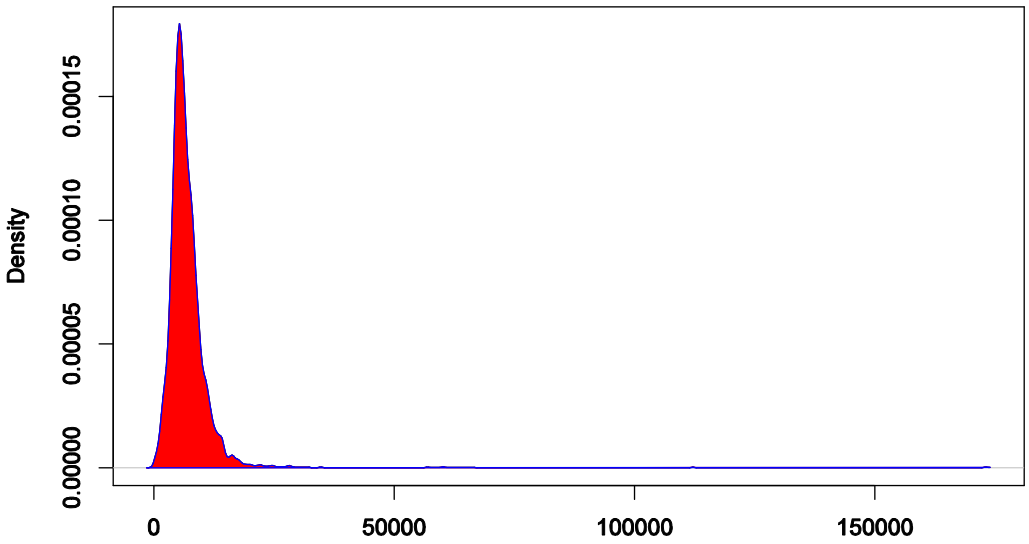
Základné vlastnosti oboch rozdelení, odvodenie vzťahov pre Giniho koeficient a Lorenzovu krivku je možné nájsť v [5].

3. Modelovanie príjmov domácností

Viaceré publikácie poukazujú na to, že použitie týchto rozdelení miesto klasických modelov, ako je napr. lognormálne rozdelenie vedie k lepším odhadom neznámeho tvaru rozdelenia príjmov (pozri [3]).

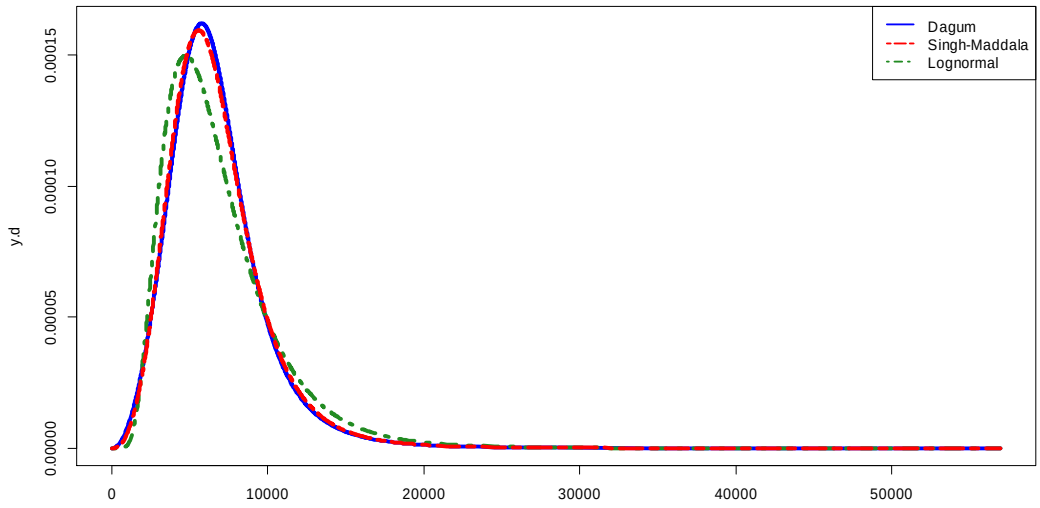
Ako vstupné dáta použijeme údaje výberového zisťovanie o príjmov a životných podmienkach domácnosti EU SILC 2010 (ŠÚ SR, 2011). EU SILC je ročné výberové zisťovanie, ktorého cieľom je získať informácie o príjmoch, o úrovni chudoby a ďalších premenných. Obdobie, za ktoré sa sledujú príjmy v zisťovaní EU SILC (príjmové referenčné obdobie), je kalendárny rok predchádzajúci roku zisťovania, t.j. pre zisťovanie EU SILC 2010 predstavovalo príjmové referenčné obdobie kalendárny rok 2009. Jednotkami výberu v EU SILC sú hospodáriace domácnosti. Hospodáriace domácnosti sú podľa metodiky Eurostatu definované ako súkromné domácnosti tvorené osobami v byte, ktoré spoločne žijú a spoločne hospodária, vrátane spoločného zabezpečovania životných potrieb. Za znak spoločného hospodárenia sa považuje spoločná úhrada základných výdavkov domácnosti (strava, úhrada nákladov na bývanie, elektrina, plyn a pod.).

Prvý odhad o tvare rozdelenia príjmov poskytuje jadrový odhad hustoty na Obr.1, z neho je možné vidieť, rozdelenie je unimodálne, výrazne špicaté a vpravo zošikmené. Takéto vlastnosti majú aj tri rozdelenia, ktoré sme vybrali na analýzu.

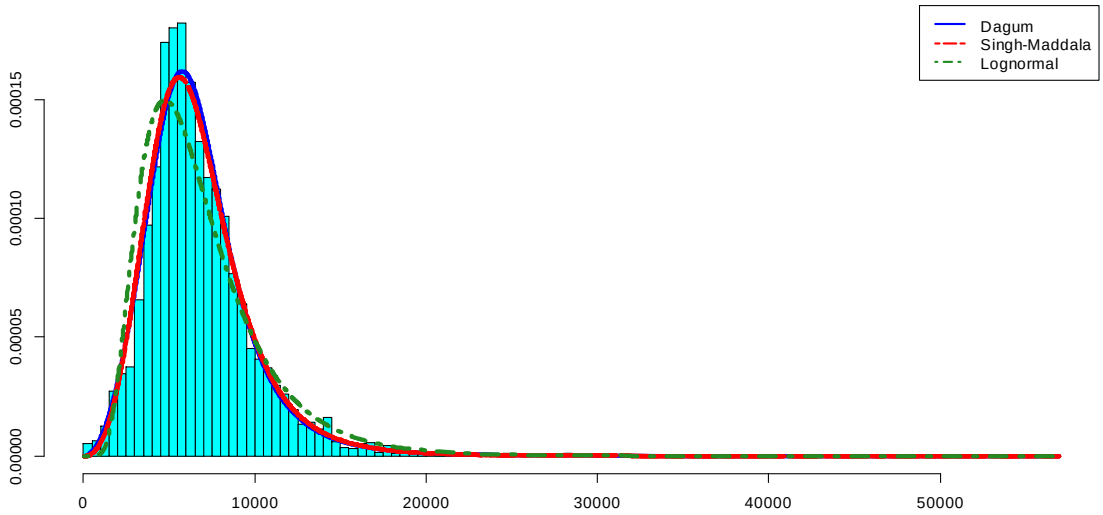


Obr. 1:Odhad jadrovej hustoty empirického rozdelenia príjmov domácností

Výsledky odhadnutých parametrov pre lognormálne, Dagumovo a SM rozdelenie sú v Tab.1. Na obrázku nižšie (Obr. 2) je porovnanie hustôt rozdelení. Pre výber vhodného rozdelenia, nám poslúži Obr. 3, kde sme zvolené teoretické rozdelenia porovnali s empirickým, s histogramom. Už z tohto grafického porovnania je vidieť, že najlepším modelom je Dagumovo rozdelenie, ktoré najlepšie modeluje stredovú časť rozdelenia a má pomalšie klesanie k osi x ako lognormálne rozdelenie. Potvrdzuje to aj hodnota tzv. Akaikeho informačného kritéria (AIC), ktorá je práve pre Dagumovo rozdelenie najnižšia (pozri Tab.1).



Obr. 2: Porovnanie hustoty rozdelenia pre Dagumovo, SM a lognormálne rozdelenie



Obr. 3: Porovnanie hustoty rozdelenia pre Dagumovo, SM a lognormálne rozdelenie s histogramom

Tab. 1: Hodnoty odhadnutých parametrov pre lognormálne, Dagumovo a Singh-Maddalovo rozdelenie a AIC

Model rozdelenia	**Odhadnuté parametre**			**AIC**
Lognormálne	8,712	0,495		307393,6
Dagum	4,653	7206,452	0,648	305050
Singh-Maddala	3,432	7366,246	1,538	305176,1

Tab. 2: Porovnanie kvantilov pre lognormálne, Dagumovo a Singh-Maddalovo rozdelenie s kvantilmi výberového súboru

Model rozdelenia	**Odhadnuté kvantily rozdelenia**				
	10%	***25%***	***Medián***	***75%***	***90%***
Empirické	3586,943	4748,593	6188,319	8165,105	10650,39
Lognormálne	3219,75	4349,53	6078,422	8490,278	11469,5
Dagum	3379,662	4675,18	6288,248	8166,823	10462,39
Singh-Maddala	3407,059	4646,71	6252,033	8231,051	10586,01

4. Záver

Najvhodnejším postupom pre analýzu diferenciácie príjmov domácností je nájdenie funkčného vyjadrenia rozdelenia príjmov. V príspevku sme sa venovali metódam hľadania najvhodnejšieho tvaru rozdelenia. Paretovo, lognormálne a gamma rozdelenie sú modely, ktoré sa najčastejšie používajú na modelovanie príjmov. Ich výhodou je jednoduchosť analytického tvaru funkcie hustoty a tiež ekonomická interpretácia parametrov. Viacero publikácií (pozri napr. [3],[6]) ukazuje, že Paretovo rozdelenie je vhodné ako model vysokých príjmov a jeho tvar je pre celé rozdelenie nevhodný, teda dobre modeluje iba pravý koniec rozdelenia príjmov. Naopak lognormálne a gamma rozdelenie dobre modelujú stredovú časť rozdelenia, ale kvôli rýchlej konvergencii k osi x nie sú dostatočným modelom na pravom konci rozdelenia príjmov. Vhodnú alternatívu k tým rozdelenia tak poskytuje Dagumovo a Singh-Maddalovo (SM) rozdelenie.Dagumovo a SM rozdelenie sú známe z aktuárskych aplikácií, alebo sú známe pod inými názvami, ako Burrovo, či inverzné Burrovo rozdelenie. Práce [3],[4] a [5] však ukazujú na výhodné použitie týchto rozdelení na modelovanie príjmov. Použitie neparametrických metód je výhodné vtedy, ak dopredu nepoznáme tvar a vlastnosti rozdelenia, pretože nevyžaduje žiaden predpoklad. Rozdelenie ročných príjmov je asymetrické rozdelenie a navyše môže byť aj viac modálne (viac vrcholové). V takom prípade, všetky bežne používane distribučné funkcie nie sú vhodné a ako vhodný model sa ponúka zmes pravdepodobnostných rozdelení.

Príspevok bol vytvorený s podporou vedeckovýskumného projektu VEGA 1/0127/11
„*Priestorová distribúcia chudoby v Európskej únii*“

Literatúra

[1] BARTOŠOVÁ, J.2009. Výberové šetření příjmu domácností v České republice, In: *Forum Statisticum Slovacum*, č.7, 2009, s. 4-9

[2] BÍLKOVÁ, D., MALÁ, I. 2012. Modelling the Income Distributions in the Czech Republic since 1992. *Österreichische Zeitschrift für Statistik [elektronickýzdroj] : Organ der Österreichischen Statistischen Gesellschaft.*. sv. 41, č. 2, s. 133--152. ISSN 1026-597X. URL: http://www.stat.tugraz.at/AJS/ausg122/122Bilkova2.pdf

[3] CHOTIKAPANICH, D. 2008. *Modelling Income Distributions and Lorenz Curves*. Springer Science+Business Media, LLC, 2008, ISBN: 978-0-387-72756-1

[4] KLEIBER, C. 1996. Dagum vs. Singh-Maddala income distributions, *Economic Letters* 53, pp. 265-268

[5] KLEIBER, C. 2007. *A Guide to Dagum Distribution*, WWZ Working Paper, URL: http://wwz.unibas.ch/uploads/tx_x4epublication/23_07.pdf

[6] KLEIBER,C.; KOTZ, S. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences*. New York :Wiley-Interscience

[7] PACÁKOVÁ, V., SIPKOVÁ, Ľ., SODOMOVÁ, E. 2005. Štatistické modelovanie príjmov domácností v Slovenskej republike, In: *Ekonomický časopis*. 2005, č. 4, d. 427- 439.

[8] R DEVELOPMENT CORE TEAM. (2012). *R: A language and environment forstatistical computing*. Viedeň: R Foundation for Statistical Computing. ISBN 3-900051-07-0. URL http://www.R-project.org/.

[9] SILVERMAN, B. W. 1996: *Density estimation for Statistics and Data Analysis*. NewYork: CHAPMAN AND HALL, 1996, 76 S. ISBN 978-0412246203

[10] STANKOVIČOVÁ, I. 2009. Analýza monetárnej chudoby v domácnostiach Českej republiky, In: *Forum Statisticum Slovacum*, č.7, 2009, s. 151-156

[11] ŠÚ SR. (2011). Zisťovanie o príjmoch a životných podmienkach EU SILC 2010 (UDB_31/08/11). [databáza s mikroúdajmi]. Bratislava: Štatistický úrad SR.

[12] TARTAĽOVÁ, A. 2010:Neparametrické metódy odhadu hustoty rozdelenia. In: *Forum Statisticum Slovacum*. č.5, 2010, s. 250-255. , ISSN 1336-7420

[13] TARTAĽOVÁ, A. 2012.Modelling income distribution in Slovakia In: *Mezinárodní statisticko-ekonomické dny v Praze* : sborník příspěvků 6. ročníku mezinárodní konference : Praha, 13-15 září 2012. - Praha : Melandrium, 2012 S. 1-10. - ISBN 978-80-86175-79-9

[14] ŽELINSKÝ, T. (2010). Pohľad na regióny Slovenska cez prizmu chudoby. In: Pauhofová, I., Hudec, O., Želinský, T. (eds.):*Sociálny kapitál, ľudský kapitál a chudoba v regiónoch Slovenska*. Košice: TU Košice. s. 37-50. ISBN 978-80-553-0573-8.

Adresa autorky:

Alena Tartaľová, Mgr., PhD.
Technická univerzita v Košiciach, Ekonomická fakulta
Katedra aplikovanej matematiky a hospodárskej informatiky
Nemcovej 32
040 01 Košice
alena.tartalova@tuke.sk

Citlivosť vybraných mier príjmovej nerovnosti na voľbu ekvivalentnej stupnice

Sensitivity of Selected Income Inequality Measures to the Choice of Equivalence Scale

Tomáš Želinský

Abstract: The aim of this paper is to analyse the impact of equivalence scales on selected income inequality measures. In the study we simulate application of various combinations of adult/child household members' weights to a linear type equivalence scale at national and regional level of Slovakia.

Abstrakt: Cieľom príspevku je analyzovať vplyv ekvivalentných stupníc na vybrané miery príjmovej nerovnosti. V štúdii je simulovaná aplikácia rôznych kombinácií váh dospelých a detských členov domácností na lineárny typ ekvivalentnej stupnice na národnej a regionálnej úrovni Slovenska.

Key words: Equivalence scales, Gini coefficient, S80/S20, EU SILC, Slovakia.

Kľúčové slová: Ekvivalentné škály, Giniho koeficient, S80/S20, EU SILC, Slovensko.

JEL classification: D63, I30, R11.

Úvod

Príjem možno považovať za jeden z najpoužívanejších ukazovateľov individuálneho blahobytu domácností/osôb. Použitie tohto ukazovateľa je však spojené s viacerými problémami – koncepčnými, ako aj metodologickými.

Každá osoba je spravidla súčasťou určitej domácnosti, za problematické možno považovať porovnávanie príjmu medzi domácnosťami rôznej veľkosti a štruktúry. Na účely porovnávania je preto potrebné každú domácnosť charakterizovať akýmsi „priemerným" príjmom [2], [10]. Na zohľadnenie vekovej štruktúry členov domácnosti sa v praxi používajú rôzne formy ekvivalentných stupníc, na základe ktorých možno určiť ekvivalentnú veľkosť domácnosti a následne ekvivalentný príjem.

Cieľom príspevku je ilustrovať, ako voľba ekvivalentnej stupnice vplýva na výšku ekvivalentného príjmu a následne na odhady vybraných mier príjmovej nerovnosti (Giniho koeficient [5] a pomer príjmov horného a dolného kvintilu).

1. Ekvivalentná stupnica a ekvivalentný príjem

Logika ekvivalentných škál je založená na myšlienke úspor z rozsahu v domácnosti. Spotrebu domácnosti totiž možno rozdeliť na *kolektívnu*, na ktorej sa podieľajú všetci členovia domácnosti (napr. náklady na bývanie) a *individuálnu*, ktorá zodpovedá jednotlivým členom domácnosti. Podľa [3] možno elasticitu takmer všetkých ekvivalentných stupníc aproximovať výrazom h^e, kde h je veľkosť domácnosti a $e \in [0; 1]$ je parameter elasticity ekvivalentnej stupnice. Pri nízkych hodnotách e je vyšší dôraz kladený na kolektívnu spotrebu a pri nízkych hodnotách na individuálnu.

Za najpoužívanejší typ stupnice možno považovať lineárny typ ekvivalentnej stupnice [8]:

$$S_i = 1 + a(A_i - 1) + bK_i, \qquad (1)$$

kde

A_i je počet dospelých v i-tej domácnosti;

K_i je počet detí v *i*-tej domácnosti;

a je parameter reprezentujúci proporciu nákladov pri ďalších dospelých členoch v *i*-tej domácnosti (resp. ich váha), $a \in (0, 1)$;

b je parameter reprezentujúci proporciu nákladov u detí v *i*-tej domácnosti (resp. ich váha), $b \in (0, 1)$.

Ekvivalentná veľkosť domácnosti je nevyhnutným vstupom na odhad ekvivalentného disponibilného príjmu domácnosti, ktorý je definovaný ako podiel celkového disponibilného príjmu domácnosti a ekvivalentnej veľkosti domácnosti, teda pre ekvivalentný disponibilný príjem *i*-tej domácnosti platí:

$$I_{E,i} = \frac{I_{H,i}}{S_i}, \text{ kde} \tag{2}$$

kde $I_{H,i}$ je celkový disponibilný príjem *i*-tej domácnosti; S_i je ekvivalentná veľkosť *i*-tej domácnosti.

2. Metodika

V štúdii sú použité mikroúdaje zisťovania EU SILC 2010 za SR [9]. Referenčným obdobím použitých údajov je rok 2009.

V príspevku je uskutočnená jednoduchá simulácia, v ktorej pomocou vzťahu (1) sú pre každú domácnosť vypočítané všetky možné kombinácie ekvivalentnej veľkosti, ktoré môžu nastať (pre zjednodušenie: a_i, b_i = {0,00; 0,05; 0,10; 0,15; ...; 1,00}). Následne je pre každú domácnosť v každom kroku odhadnutá nová úroveň ekvivalentného disponibilného príjmu, ktorá je priradená každej osobe v domácnosti.

Odhadované sú dve miery nerovnosti rozdeľovania príjmov:

(1) Giniho koeficient:

$$G_w = 100\left(1 - \frac{\sum_{i=1}^{n}\left[\left(2\sum_{j=1}^{i}(y_i w_i) - y_i w_i\right)w_i\right]}{\sum_{i=1}^{n} y_i w_i \cdot \sum_{i=1}^{n} w_i}\right), \tag{3}$$

kde *n* je veľkosť súboru, y_i je príjem *i*-tej osoby, w_i je osobná prierezová váha *i*-tej osoby.

(2) S80/S20 pomer príjmov horného a dolného kvintilu:

$$S80/S20 = \frac{\sum_{i=x_{80}}^{n} y_i w_i}{\sum_{i=1}^{x_{20}} y_i w_i}, \tag{4}$$

kde y_i je príjem *i*-tej osoby, w_i je osobná prierezová váha *i*-tej osoby, *n* je veľkosť súboru, x_{20} je 20-percentný kvantil rozdelenia domácností podľa príjmov, x_{80} je 80-percentný kvantil rozdelenia domácností podľa príjmov.

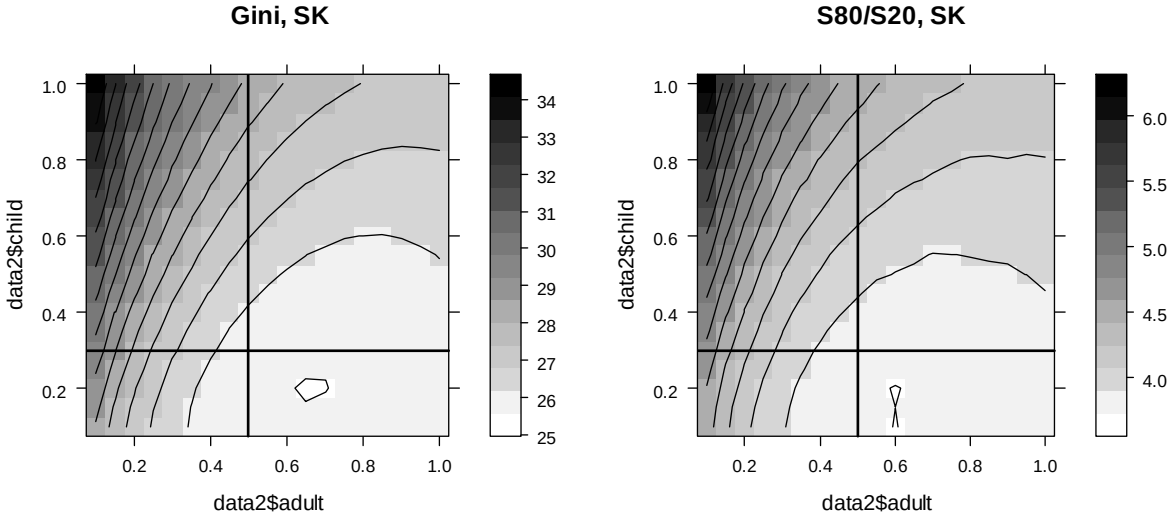
Kvôli lepšej názornosti sú výsledky prezentované graficky vo forme tzv. úrovňových grafov, kde na osi *x* je parameter (váha) pre člena domácnosti vo veku do 14 rokov vrátane (v texte ďalej uvádzané aj ako „deti“) a na osi *y* je parameter (váha) pre člena domácnosti staršieho ako 14 rokov (v texte ďalej uvádzané aj ako „dospelí“). Prvému dospelému členovi

domácnosti je vždy priradená váha 1. Vo všetkých výstupoch je plnou čiarou znázornený bod zodpovedajúci oficiálnej ekvivalentnej škále (t. j. $a = 0{,}5$ a $b = 0{,}3$).

Odhady indexov sú uskutočnené v súlade s metodikou Eurostatu [4]. Všetky odhady a výpočty v štúdii sú uskutočnené v prostredí softvéru **R** [9] s použitím knižníc „*laeken*" [1] a „*lattice*" [7].

3. Výsledky a diskusia

Použitím metodiky opísanej v predchádzajúcej kapitole dostávame obraz, aké rôzne hodnoty Giniho koeficientu a pomeru príjmov horného a dolného kvintilu (S80/S20) by sme dostali, ak by sme použili rôzne váhy pre členov domácností mladších/starších ako 14 rokov (Obr. 1).



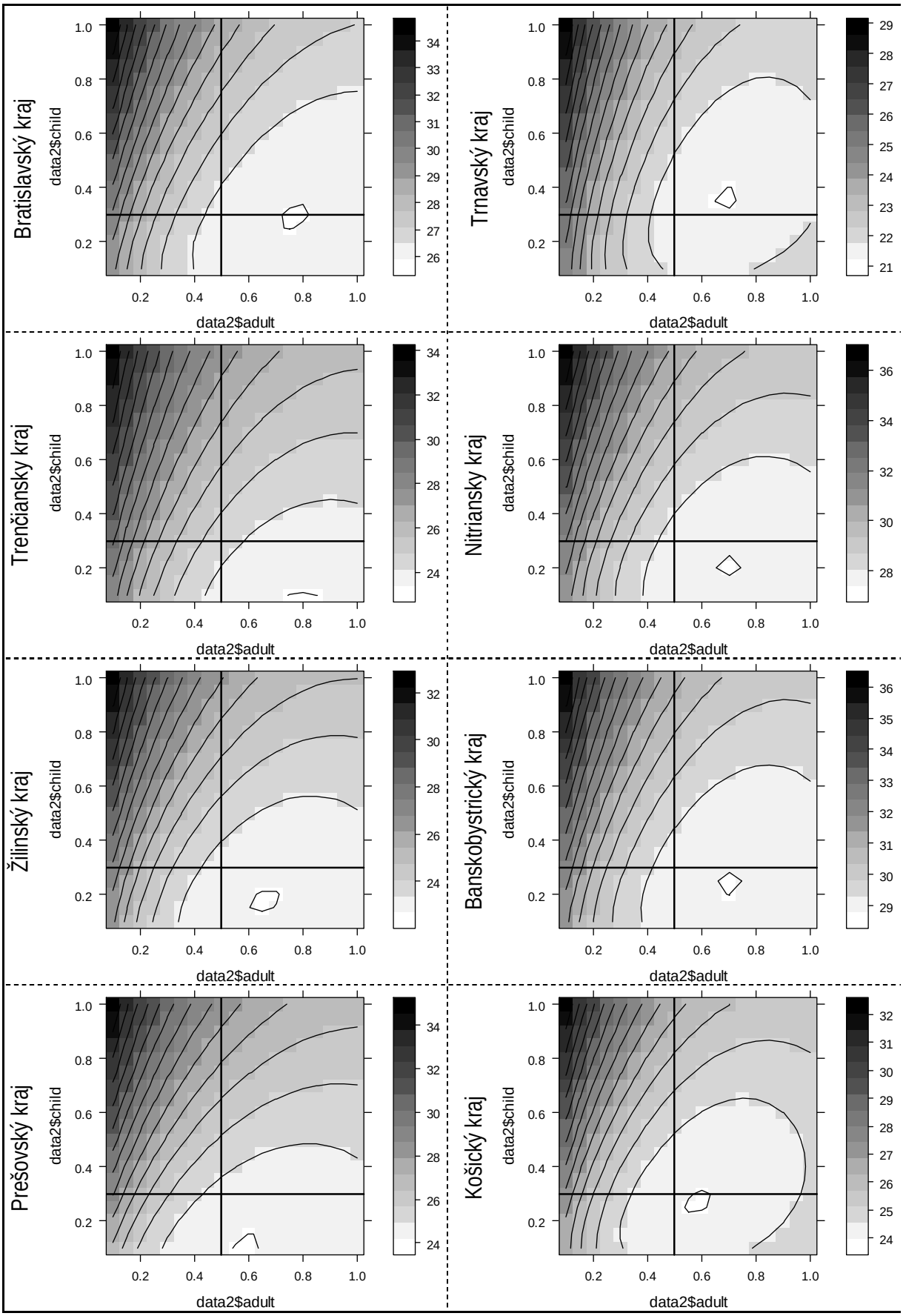
Obr. 1: Giniho koeficient a S80/S20 (Slovensko, r. 2010)

Zdroj: vlastné spracovanie podľa údajov EU SILC

Podľa údajov EU SILC 2010 bol odhadnutý Giniho koeficient v SR na úrovni približne 26 a odhadnutý pomer príjmov horného a dolného kvintilu približne 3,8. Je ale zrejmé, že použitím iných ekvivalentných škál by sa odhadnuté hodnoty indexov líšili. Najvyššie hodnoty sú zaznamenané pri maximalizovaní váhy detí a minimalizovaní váhy dospelých, kedy Giniho koeficient prevyšuje hodnotu 34 a S80/S20 hodnotu 6.

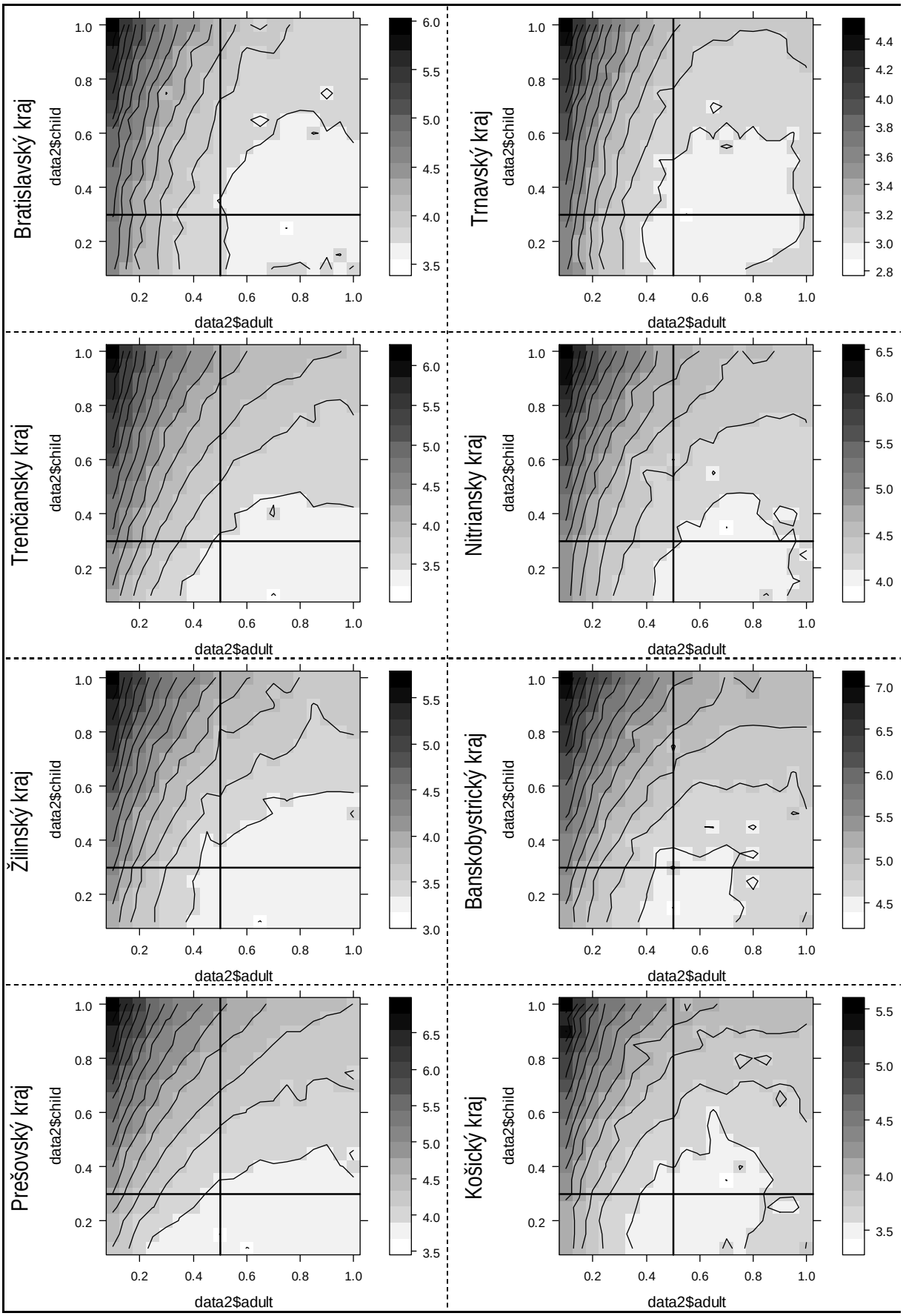
Podobne, ako v prípade prezentácie výsledkov odhadnutých mier rizika chudoby pre SR s použitím kombinácií váh pre dospelých a detských členov domácností, aj v prípade krajov sú výsledky prezentované vo forme úrovňových grafov (Obr. 2 a 3). Z uvedených úrovňových grafov je zrejmé, že takmer vo všetkých prípadoch je zachovaná schéma rozdelenia analyzovaných indexov odhadnutých pre celú SR. Tento predpoklad je potvrdený aj hodnotami Spearmanovho koeficientu poradovej korelácie, ktoré sa pohybujú na úrovni nad 0,9[1]. Znamená to teda, že aj po prechode na nižšiu teritoriálnu úroveň je zachovaná konzistentnosť s národnou úrovňou. Takéto zistenie súvisí so skutočnosťou, že medzi krajmi SR neexistujú výrazné rozdiely v základnej štruktúre spotrebných výdavkov domácností, a teda uplatnenie rôznych ekvivalentných škál pre rôzne kraje nie je opodstatnené.

[1] *Gini:* BA: 0.980, TT: 0.935, TN: 0.984, NR: 0.997, ZA: 0.999, BB: 0.995, PO: 0.987, KE: 0.965
S80/S20: BA: 0.907, TT: 0.943, TN: 0.988, NR: 0.988, ZA: 0.987, BB: 0.986, PO: 0.980, KE: 0.971



Obr. 2: Gini koeficient, kraje SR, 2010

Zdroj: vlastné spracovanie podľa údajov EU SILC



Obr. 3: Pomer príjmov horného a dolného kvintilu, kraje SR, 2010
Zdroj: vlastné spracovanie podľa údajov EU SILC

4. Záver

Odhad ukazovateľov príjmovej nerovnosti (príp. chudoby) je závislý od použitej ekvivalentnej veľkosti domácnosti. Je totiž potrebné uvedomiť si, že aplikácia „správnej" ekvivalentnej stupnice je nevyhnutná na získanie neskreslenej informácie o príjmovej situácii v krajine. V príspevku je uskutočnená analýza, ako ovplyvní použitá ekvivalentná stupnica (lineárny typ s obmieňaním váh dospelých/detských členov domácností) výslednú hodnotu vybraných mier príjmovej nerovnosti.

Poďakovanie

Príspevok bol napísaný s podporou Vedeckej grantovej agentúry MŠ SR a SAV v rámci riešenia vedecko-výskumného projektu *VEGA 1/0127/11 Priestorová distribúcia chudoby v EÚ* a s podporou *Stipendia Husovy nadace a Nadácie UPJŠ*.

Literatúra

[1] ALFONS, A., HOLZER, J., TEMPL, M. 2012. laeken: Estimation of indicators on social exclusion and poverty. R package version 0.3.3.

[2] BARTOŠOVÁ, J., STANKOVIČOVÁ, I. 2009. Deferenciace příjmú a chudoba v českých a slovenských domácnostech. In: *MSED 2009*. Sborník přispěvků : VŠE Praha. s. 1-6. ISBN 978-80-86175-66-9.

[3] BUHMANN, B. ET AL. 1988. Equivalence Scales, Well-being, Inequality and Poverty: Sensitivity Estimates Across Ten Countries Using the Luxembourg Income Study Database. In: *The Review of Income and Wealth*. Vol.34, No.2, s. 115–142.

[4] EUROSTAT. 2010. Algorithms to compute Social Inclusion Indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC). Working Group meeting "Statistics on Living Conditions", 10-12 May 2010. Luxembourg: Eurostat.

[5] LABUDOVÁ,V. 2012. Miery príjmovej nerovnosti. In: Pauhofová, I., Želinský, T. (eds.): *Nerovnosť a chudoba v Európskej únii a na Slovensku*. Košice: Ekonomická fakulta TUKE. s. 107-112.

[6] R DEVELOPMENT CORE TEAM. 2012. R: A language and environment for statistical computing. Viedeň: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

[7] SARKAR, D. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer. ISBN 978-0-387-75968-5.

[8] SIPKOVÁ, Ľ. 2009. Ekvivalentná škála v EU-SILC analýzach príjmovej nerovnosti a chudoby. In: Pacáková, V. (ed.): *Štatistické metódy v ekonómii so zameraním na sociálne analýzy*. Bratislava: EKONÓM. ISBN 978-80-225-2704-0. s. 81-126.

[9] ŠÚ SR. 2011. Zisťovanie o príjmoch a životných podmienkach EU SILC 2010 (UDB_31/08/11). [databáza s mikroúdajmi]. Bratislava: Štatistický úrad SR.

[10] TARTAĽOVÁ, A. 2011. Odhad hustoty rozdelenia zmesou exponenciálnych funkcií. In: *Forum Statisticum Slovacum*. Vol. 7, No. 7, s. 251-256.

Adresa autora:

Tomáš Želinský, Ing. PhD.
Ekonomická fakulta, TU Košice
Němcovej 32, 040 01 Košice
tomas.zelinsky@tuke.sk

Analýza příjmů domácností v závislosti na věku a vzdělání v mezinárodním srovnání

The analysis of the household income depending on age and education in international comparison

Jitka Bartošová, Klára Siegelová

Abstract: This article deals with the linear regression of household income, depending on several factors like age structure of the population and education. The article is based on the results of the survey EU-SILC for households from Czech Republic, Slovakia and Germany. General linear model were used for linear regression, because of the data file properties. The model parameters were estimated in statistical program called SAS.

Abstrakt: Tento článek se zabývá lineární regresí příjmů domácností v závislosti na několika faktorech, zejména z pohledu věkové struktury obyvatelstva a jeho dosaženého vzdělání. Článek vychází z výsledků statistického šetření EU-SILC v roce 2009 a věnuje se domácnostem České Republiky, Slovenska a Německa. Pro lineární regresi je použit zobecněný lineární model, který lépe vyhovuje datovým vlastnostem souboru EU-SILC a parametry modelu byly odhadnuty v prostředí statistického programu SAS.

Key words: Household income, education, age structure, general linear model

Klíčová slova: Příjem domácnosti, vzdělání, věková struktura, zobecněný lineární model

JEL classification: J31

Úvod

Modelování příjmů domácností z hlediska věkové struktury, vzdělání a dalších faktorů umožňuje sledování ekonomické a sociální situace obyvatelstva a dále slouží k mezinárodnímu srovnání. Datová základna pro model je výběrové šetření příjmů a životních podmínek domácností s názvem EU-SILC z roku 2009. Poprvé bylo toto šetření provedeno Českým statistickým úřadem v roce 2005 pod názvem Životní podmínky 2005.

Základní jednotkou členění je hospodařící domácnost a její osoby, které mají ve vybraném bytě jediné či hlavní bydliště. Konstrukce hospodařící domácnosti je v duchu § 115 občanského zákoníku založena na prohlášení osob bydlících ve vybraném bytě, že spolu trvale žijí a společně uhrazují náklady na své potřeby. Mezi 16leté osoby byly zahrnuty osoby, které tento věk dovršily ke konci roku 2008.[2] Výběrový plán je založen na náhodném dvoustupňovém výběru pro každý kraj nezávisle tak, aby celkový počet vybraných domácností byl úměrný velikosti jednotlivých krajů. Pro šetření se používá, v souladu s metodikou doporučenou Eurostatem, systém tzv. integrovaných vah, tj. jediná sada přepočítacích koeficientů, vhodný pro souběžné zpracování výstupů jak za hospodařící domácnosti, tak za jednotlivé osoby. Protože šetření podléhaly pouze osoby žijící v bytech, byly na úrovni ČR od všech údajů z demografie odečteny odhady počtu osob (mladistvých, důchodců, atd.) žijících v tzv. ústavních domácnostech (ústavy sociální péče, nápravná zařízení, domovy důchodců). Odhad počtu byl stanoven podle statistik sociálního zabezpečení za rok 2008. Ohledně příjmů dochází v šetření ke zkreslení informací, vzhledem k tomu, že domácnosti záměrně i nezáměrně podhodnocují peněžní příjmy nebo údaje o příjmech chybí úplně. V takových situacích jsou hodnoty statisticky dopočítány nebo se chybějící příjmy

[2] Zdroj: http://www.czso.cz/csu/2010edicniplan.nsf/publ/3012-10-

doplní od jiné náhodně vybrané osoby z podobné domácnosti. Proto je nutné brát v úvahu chybu zjišťování.

1. Zobecněný lineární model

Klasický lineární model se pro modelování datové základny nehodí, pro účely byl zvolen zobecněný lineární regresní model. Některá zobecnění dovolující předpoklady klasického modelu zeslabit jsou následující:

- Heteroskedasticita
- Korelované náhodné složky v modelu
- Regresory mohou být stochastické
- Náhodná složka může mít jiné než normální rozdělení

Zobecněný lineární model (dále GLM) poskytuje rámec vytváření jednotné třídy modelů, které pracují se spojitými i kategorizovanými nezávislými proměnnými. GLM zahrnuje lineární regresi, ANOVA modely, logistickou regresi, loglineární modely, probitové modely.
Zobecněným lineárním modelem se rozumí klasický lineární model se změněnou podmínkou týkající se kovarianční matice nepozorovatelné náhodné složky ε a při nenáhodnosti X i y. Rozdíl mezi KLM a ZLM je v tom, že místo kovarianční matice

$$C(\varepsilon) = C(y) = \sigma^2 I_n \qquad (1)$$

se zavádí obecnější matice

$$C(\varepsilon) = C(y) = \Omega = \sigma^2 W. \qquad (2)$$

Kde Ω a W jsou symetrické pozitivně definitivní matice.
Náhodná veličina Y, tedy závislá proměnná je sloupcový vektor náhodných veličin a je typu $n \times 1$, tedy $y = [y_1, y_2, \dots y_n]^T$. Matice X nezávislých proměnných je typu $n \times p$. Její j-tý sloupec označujeme x_j. Vektor parametrů je následovný $\beta = [\beta_1, \dots \beta_p]$. Náhodná složka modelu má vektor středních hodnot $E(Y) = \mu$ typu $n \times 1$ a kovarianční matici cov(Y). Lineární predátor η je systematická složka v lineárním modelu, tedy

$$\eta = \sum_{j=1}^{p} x_j \beta_j, \qquad (3)$$

kde x_j je j-tý sloupec matice X, tj vektor $n \times 1$

2. Příjmy domácností

Do modelu vstupuje jako závisle proměnná logaritmované celkové příjmy domácností. Jedná se o celkové příjmy domácnosti skládající se z příjmů z hlavního či vedlejšího pracovního poměru, z příjmů z podnikání a jiné samostatně výdělečné činnosti, příjmů státní sociální podpory nebo jiných dávek sociální podpory, příjmy z pronájmu, příjmy z kapitálového majetku atd. Mezi kvantitativní nezávislé proměnné byly zařazeny celkové výdaje domácnosti a do kvalitativních byly zvoleny proměnné Pohlaví, Věk, Vzdělání, Ekonomický status. Celkové výdaje domácnosti jsou míněny jako výdaje na bydlení, zahrnující veškerou režii, nájemné a zároveň výdaje spojené s půjčkami vázajícími se k bydlení. Pohlaví je u hodnoty 1 muž a u hodnoty 2 žena.
Věkové kategorie byly rozděleny do čtyř skupin:

1-24	1
25-39	2
40-65	3
66+	5

Vzdělání je nejvyšší dosažená úroveň podle ISCED a je rozdělena do pěti skupin:

Preprimární vzdělání (bez vzdělání) a Primární vzdělání	1
Nižší sekundární vzdělání	2
Vyšší sekundární vzdělání a Postsekundární vzdělání	3

První stupeň terciárního vzdělání	5
Druhý Stupeň terciárního vzdělání	6

Ekonomický status definovaný jako subjektivní názor osoby vyplňují dotazník a je rozdělen do 9 skupin:

Plný úvazek	1
Částečný úvazek	2
OSVČ plný i částečný úvazek	3
Nezaměstnaný	5
Student, žák	6
Důchodce	7
Invalidní osoba nebo nezpůsobilá pracovat	8
Základní vojenská nebo civilní služba	9
Osoba v domácnosti nebo jiná neaktivní osoba	10

Výsledky modelu pro jednotlivé země

Česká Republika

Tabulka 1 Významnost modelu

Koeficient determinace	Variační koeficient	lnPříjem průměr
0.293460	5.610325	9.353662

Tabulka 2 Významnosti parametrů

Parametr	Odhady	Standardní odchylka	Testová statistika	Pr > |t|
Konstanta	9.597246228	0.00642670	1493.34	<.0001
Výdaje	-0.000016025	0.00000199	-8.07	<.0001
EkStatus 10	-0.039699506	0.01290870	-3.08	0.0021
EkStatus 2	-0.119498535	0.02151709	-5.55	<.0001
EkStatus 3	0.088844924	0.01026946	8.65	<.0001
EkStatus 5	-0.535867008	0.01251619	-42.81	<.0001
EkStatus 6	-0.257588088	0.03779258	-6.82	<.0001
EkStatus 7	-0.423741320	0.00985696	-42.99	<.0001
EkStatus 8	-0.403462546	0.01366387	-29.53	<.0001
EkStatus 1	0.000000000	.	.	.
Pohlaví žena	0.038311031	0.00596422	6.42	<.0001
Pohlaví muž	0.000000000	.	.	.
VěkKategorie 1	-0.452482188	0.02695896	-16.78	<.0001
VěkKategorie 2	-0.095435450	0.00736296	-12.96	<.0001
VěkKategorie 5	-0.263574152	0.00988801	-26.66	<.0001
VěkKategorie 3	0.000000000	.	.	.
Vzdělání 1	-0.354973025	0.06126417	-5.79	<.0001
Vzdělání 2	-0.182872865	0.00844932	-21.64	<.0001
Vzdělání 5	0.270100885	0.00822461	32.84	<.0001
Vzdělání 3	0.000000000	.	.	.

Referenční skupina pro všechny modely je muž s vyšším sekundárním vzděláním, pracující na plný úvazek a ve věkové kategorii 40-65 let, byla zvolena na základě nejvyšších četností. Příjem této domácnosti je 14724 Euro ročně.

Za Českou republiku nejvyšší příjem ve věkové kategorii 1-24 let u mužů mají vysokoškolsky vzdělaní muži pracující na plný úvazek ve výši 18352 Euro ročně. Jedná se ale pouze o pět takových domácností. V nejmladší věkové kategorii je nejvíce mužů, studentů s vyšším sekundárním vzděláním s celkovým příjmem 7925 Euro a dále mužů se sekundárním vzděláním pracujících na plný úvazek s průměrným platem 12838 Euro. U žen je největší podíl sekundárně vzdělaných s ekonomickou aktivitou na plný pracovní úvazek. V kategorii věku 25-39 mají největší celkové příjmy terciárně vzdělané ženy samostatně výdělečně činné a jejich podíl je na celkovém počtu domácností s osobou v čele žena pouze 2%. Jejich celkový příjem je 28290 Euro. Nejpočetnější skupinou, jak u žen, tak u mužů jsou zase sekundárně vzdělané domácnosti s plným pracovním úvazkem. Výrazný počet žen v této věkové kategorii je osobou v domácnosti, kvůli dětem. U další věkové kategorie 40-65 let výrazně stoupl počet nezaměstnaných osob a nejvíce u sekundárně vzdělaných lidí. Celkově počet nezaměstnaných k ekonomicky aktivním (i neaktivním) je 8%. V poslední věkové kategorii nad 66 let je jasné, že vzrostl počet důchodců, zajímavé je, že není výrazný rozdíl v celkových příjmech důchodců se sekundárním nebo terciárním vzděláním.

Slovensko

Tabulka 3 Významnost modelu

Koeficient determinace	Variační koeficient	lnPříjem Průměr
0.261539	6.305411	9.146892

Tabulka 4 Významnosti parametrů

Parametr	Odhady	Standardní odchylka	t hodnota	Pr > |t|
Konstanta	9.429466587	0.01067724	883.14	<.0001
Výdaje	-0.000016384	0.00000237	-6.92	<.0001
Pohlaví 2	-0.011896783	0.01000531	-1.19	0.2344
Pohlaví 1	0.000000000	.	.	.
VěkKategorie 1	-0.430522437	0.05973492	-7.21	<.0001
VěkKategorie 2	-0.155716924	0.01233266	-12.63	<.0001
VěkKategorie 5	-0.378907788	0.01629419	-23.25	<.0001
VěkKategorie 3	0.000000000	.	.	.
EkStatus 10	-0.178529132	0.02575505	-6.93	<.0001
EkStatus 2	-0.134968611	0.03196680	-4.22	<.0001
EkStatus 3	-0.150353412	0.02110994	-7.12	<.0001
EkStatus 5	-0.565052504	0.02314426	-24.41	<.0001
EkStatus 6	0.329530733	0.11366828	2.90	0.0037
EkStatus 7	-0.347971813	0.01515643	-22.96	<.0001
EkStatus 8	-0.391416548	0.02725793	-14.36	<.0001
EkStatus 1	0.000000000	.	.	.
Vzdělání 1	-0.262767610	0.04315535	-6.09	<.0001
Vzdělání 2	-0.165753377	0.01595994	-10.39	<.0001
Vzdělání 5	0.249374194	0.01221665	20.41	<.0001
Vzdělání 3	0.000000000	.	.	.

Za stejnou referenční skupinu jsou roční příjmy v hodnotě 12449 Euro.

U věkové skupiny do 24 let se na Slovensku u mužů neobjevují vůbec nezaměstnaní junioři a oproti České Republice je větší počet žen než mužů v této věkové kategorii. V datovém souboru se objevují i osoby, které nemají žádné vzdělání, což u dat z ČR není.

Osoby bez vzdělání se především nacházejí v nejstarší věkové skupině, ale jejich počet vzhledem k celku je zanedbatelný. Ve věkové kategorii 25-39 let je menší poměr terciárně k sekundárně vzdělaným osobám než v ČR. Výrazně vyšší celkové příjmy mají vysokoškoláci OSVČ a dokonce u žen je hodnota vyšší než u mužů 24493 Euro. Stejně tak jako v předchozím souboru je největší podíl osob se sekundárním vzděláním a pracujících na plný úvazek. Poměr nezaměstnaných k ekonomicky aktivním (i neaktivním) je pouze 5,5%, což je výrazně lepší výsledek než u ČR u věkové skupiny 40-65 let. Průměrný celkový příjem u vysokoškoláků pracujících na plný úvazek je 19342 Euro u mužů, což je o cca 4000 nižší než v Čechách. U osob nad 66 let je podobné rozložení jako v ČR a není významný rozdíl v příjmech podle vzdělání.

Německo

Tabulka 5 Významnost modelu

Koeficient determinace	Variační koeficient	lnPříjem průměr
0.342132	5.845752	10.09506

Tabulka 6 Významnosti parametrů

Parametr	Odhady	Standardní odchylka	Testová statistika	Pr > |t|
Konstanta	9.899984155	0.00277464	3568.02	<.0001
Výdaje	0.000524815	0.00000201	261.18	<.0001
Pohlaví 2	-0.051403742	0.00203179	-25.30	<.0001
Pohlaví 1	0.000000000	.	.	.
VěkKategorie 1	-0.398346280	0.00733227	-54.33	<.0001
VěkKategorie 2	-0.029229310	0.00278455	-10.50	<.0001
VěkKategorie 5	-0.020048693	0.00477541	-4.20	<.0001
VěkKategorie 3	0.000000000	.	.	.
EkStatus 10	-0.050003881	0.00525433	-9.52	<.0001
EkStatus 2	-0.128345590	0.00272131	-47.16	<.0001
EkStatus 3	-0.019940480	0.00605470	-3.29	0.0010
EkStatus 5	-0.746669120	0.00438983	-170.09	<.0001
EkStatus 6	-0.772390448	0.00718061	-107.57	<.0001
EkStatus 7	-0.301025779	0.00480124	-62.70	<.0001
EkStatus 8	-0.687172794	0.00720319	-95.40	<.0001
EkStatus 9	0.126717757	0.06872069	1.84	0.0652
EkStatus 1	0.000000000	.	.	.
Vzdělání 1	-0.157944409	0.01088648	-14.51	<.0001
Vzdělání 2	-0.143027662	0.00398534	-35.89	<.0001
Vzdělání 5	0.240112779	0.00224201	107.10	<.0001
Vzdělání 3	0.000000000	.	.	.

Ve vybrané referenční skupině jsou roční celkové příjmy rovny 19930 Euro.

Koeficient determinace je ze všech tří modelů nejvýstižnější.

V Německu je výrazně vyšší celkový příjem než v obou předchozích zemích. Podíl vysokoškoláku oproti sekundárně vzdělaným je také vyšší než v ČR. V Německu můžeme pozorovat znatelnější rozdíl v celkových příjmech osob nad 66 let podle vzdělání. Osoby s vyšším vzděláním vykazují vyšší příjmy.

I podle modelu je zřejmé, že výsledky v Německém souboru dat za domácnosti jsou oproti ČR a SR lepší.

3. Závěr

Provedený zobecněný lineární model prokázal vyšší celkové příjmy v Německu za referenční skupinu, ale i za jiné skupiny domácností. Zároveň všechny parametry proměnných vstupují do modelu jako významné, podle vypočítaných p-hodnot. Tedy je určitá závislost celkových příjmů na ekonomickém statusu, vzdělání, pohlaví a věkové kategorii a výdajích na bydlení. Koeficienty determinace jsou dost nízké, nejedná se tedy o úplně kvalitní modely, důvodem může být narušení podmínek homoskedasticity, nezávislosti či normálnímu rozdělení odezvy.

4. Literatura

[1] ČSÚ: Příjmy a životní podmínky domácností 2009. [online]. [cit. 2012-11-20]. Dostupné z: http://www.czso.cz/csu/2010edicniplan.nsf/publ/3012-10-

[2] HEBÁK, Petr. Vícerozměrné statistické metody. Praha: Informatorium, 2010. ISBN 9788073330569

[3] STANKOVIČOVÁ, Iveta a VOJTKOVÁ, Mária. Viacrozmerné štatistické metódy s aplikáciami. Iura Edition, 2007. ISBN 978-80-8078-152-1.

Poděkování: Příspěvek byl vytvořen s podporou vědeckovýzkumného projektu Interní grantové agentury Vysoké školy ekonomické v Praze IG F6/3/2012 *"Kvantitativní studie sociální situace juniorů a seniorů"*.

Mikrodata EU-SILC byla poskytnuta na vědecké účely na základě kontraktu no. EU-SILC/2011/33, podepsaného mezi Evropskou komisií, Eurostatem a Technickou univerzitou v Košicích. Eurostat nenese žádnou odpovědnost za výsledky a závěry, ke kterým autorky dospěly.

Adresa autorů:

Klára Siegelová
Vysoká škola ekonomická v Praze
Fakulta managementu
Jarošovská 1117/2
377 01 Jindřichův Hradec
klara_siegelova@hotmail.com

RNDr. Jitka Bartošová, Ph.D
Vysoká škola ekonomická v Praze,
Fakulta managementu
Jarošovská 1117/2
377 01 Jindřichův Hradec
bartosov@fm.vse.cz

I N F O R M Á C I E

90. VÝROČIE MAĎARSKEJ ŠTATISTICKEJ SPOLOČNOSTI
90TH ANNIVERSARY OF HUNGARIAN STATISTICAL ASSOCIATION

Abstract: Hungarian Statistical Association organized in November 2012 the conference on the occasion of the 90th anniversary of its foundation. This article informs about the content of the conference.

Abstrakt: Maďarská štatistická spoločnosť organizovala v novembri 2012 konferenciu pri príležitosti 90. výročia svojho vzniku. Príspevok informuje o obsahu konferencie.

Vo vládnom rekreačnom zariadení v Balatonöszöd pri Balatóne oslavovala 15. a 16.novembra 2012 Maďarská štatistická spoločnosť (MŠS) 90. výročie svojho založenia. Oslava sa konala formou konferencie, na ktorej sa zúčastnilo vyše 200 domácich účastníkov a zahraniční hostia. Zo zahraničia sa na konferencii zúčastnila námestníčka generálneho riaditeľa Eurostatu Marie Bohatá, prezident Federácie národných štatistických spoločností Maurizio Vichi, delegácie českej a rakúskej štatistickej spoločnosti a delegácia Slovenskej štatistickej a demografickej spoločnosti.

Rokovanie konferencie bolo rozdelené do štyroch častí. V prvej - úvodnej časti hovoril bývalý predseda MŠS Sándor Herman o deväťdesiatročnej histórii spoločnosti a o úlohe politiky v jej činnosti. V histórii sa striedali obdobia, kedy bola spoločnosť nezávislou odbornou inštitúciou s obdobiami, keď bol vplyv politiky na činnosť spoločnosti významné. Predseda MŠS Lőrinc Soós hovoril o aktuálnom mieste spoločnosti v štatistickej komunite a o najbližších plánoch. Venoval pritom pozornosť najmä úsiliu, ktoré spoločnosť vyvíja v oblasti štatistickej etiky a podielu spoločnosti na implementácii Code of practice. Na záver prvej časti odzneli blahoželania k výročiu spoločnosti. Podpredseda SŠSD pre medzinárodné styky Peter Mach ocenil vzájomnú spoluprácu najmä v rámci regionálnej iniciatívy V6 a odovzdal pozdravný list predsedu SŠDS Jozefa Chajdiaka k výročiu.

Druhá časť konferencie bola venovaná pamiatke zakladateľa maďarskej štatistiky Károlyho Keleti. Jeho prácu najmä v oblasti poľnohospodárskych cenzov priblížila generálna tajomníčka MŠS a podpredsedníčka Maďarského štatistického úradu Éva Laczka. Predseda historickej sekcie MŠS Tamás Faragó hovoril o práci Keletiho ako demografa. Riaditeľka knižnice Maďarského štatistického úradu Erzsébet Nemes predstavila Keletiho vo svetle archívnych dokumentov. K spomienkam v tomto bloku sa pripojil aj zástupca Maďarskej reformnej cirkvi.

Tretia časť konferencie bola venovaná pohľadom na štatistiku zo strany partnerov MŠS. Bývalý predseda Maďarskej ekonomickej spoločnosti József Veress poukázal na dôležitosť dôveryhodnosti štatistiky. Člen Maďarskej akadémie vied Gábor Tusnády sa vo svojom vystúpení zamyslel nad tým, čo ponúkajú matematici štatistike. Lajos Besenyei, predseda výboru pre štatistiku a prognózy Maďarskej akadémie vied, hovoril o profesionálnych výzvach súvisiacich s kvantitatívnou explóziou informácií. Autor ju vidí v štatistickom spracovaní a prezentovaní informácii. Aj tento prístup má však svoje úskalia.

Posledná časť konferencie bola venovaná štatistike v Európe a v Maďarsku. Zástupkyňa generálneho riaditeľa Eurostatu Marie Bohatá hovorila o úlohách štatistiky v EÚ.

Predsedníčka Maďarského štatistického úradu Gabriella Vukovich hovorila o práci úradu z pohľadu tradícií a inovácií. Predsedníčka Rakúskej štatistickej spoločnosti Margit Epler hovorila o nezávislosti úradných štatistík. Gábor Rappai, predseda etickej komisie MŠS, sa zamyslel nad postavením štatistiky vo vzdelávacom systéme. Predseda európskeho regionálneho výboru Bernouliho spoločnosti László Márkus priblížil aktivity tejto spoločnosti, ktorá je jednou zo sekcií ISI. Šefredaktor László Hunyadi pripomenul, že 90 rokov oslavuje aj časopis Maďarského štatistického úradu - Magyar statisztikai szemle (Maďarská štatistická revue).

Záverom možno konštatovať, že konferencia bola dôstojnou oslavou významného výročia MŠS. Účasť delegácie našej spoločnosti bola potvrdením dobrých kontaktov našich spoločností a poskytla aj námety pre prípravu podobných podujatí u nás (aj keď na deväťdesiate výročie si ešte budeme musieť 45 rokov počkať).

Foto: Česká a slovenská delegácia na konferencii s predsedom Maďarskej štatistickej spoločnosti (zľava: Dohnal CZ, Soós HU, Žambochová CZ, Mach SK, Wimmer SK).

RNDr. Peter Mach
podpredseda SŠDS pre medzinárodné styky
petermach1951@yahoo.com

Z HISTÓRIE SEMINÁROV VÝPOČTOVÁ ŠTATISTIKA 2012
FROM THE HISTORY OF SEMINARS COMPUTATIONAL STATISTICS 2012

Pri príležitosti 21. ročníka seminára Výpočtová štatistika uvádzame stručnú chronológiu minulých ročníkov.

Prvý seminár sa uskutočnil 9. - 10. 12. 1986 z iniciatívy zamestnancov Katedry štatistiky VŠE v Bratislave a Katedry statistiky VŠE v Prahe zaoberajúcimi sa problematikou využitia výpočtovej techniky v riešení štatistických úloh. Príspevky účastníkov boli uverejnené v Informáciách SDŠS č. 3 a č. 4 v roku 1986.

Miestom konania Seminárov bola vždy budova Infostat-u. Väčšina seminárov sa organizovala v spolupráci so Štatistickým úradom SR (resp. SŠU v Bratislave) a Infostat-om Bratislava (resp. VUSEIaR Bratislava). **V aktuálnom 21. ročníku seminára je miesto konania Kongresová hala ŠÚ SR na Hanulovej 5/c v Bratislave a druhá časť akcie pre mladých štatistikov a demografov: Pohľady do analytiky** - Analytika očami profesionálov **- pásmo prednášok sa koná v spoločnosti SAS Slovakia, s. r. o., Lazaretská ul. 12, 811 08 Bratislava.**

Druhý seminár prebehol 8. 12. 1987, tretí seminár 11. - 12. 12.1990. Potom nastala prestávka v organizácii seminárov Výpočtovej štatistiky a 4. seminár sa uskutočnil 7. - 8. 12. 1994.

Od 5. seminára uskutočneného 5. - 6. 12. 1996 sa už realizuje každoročne ako medzinárodný seminár.

6. medzinárodný seminár Výpočtová štatistika sa uskutočnil 4.- 5. 12. 1997,

7. medzinárodný seminár Výpočtová štatistika sa uskutočnil 3. - 4. 12. 1998,

8. medzinárodný seminár Výpočtová štatistika sa uskutočnil 2. - 3. 12. 1999,

9. medzinárodný seminár Výpočtová štatistika sa uskutočnil 7. – 8. 12. 2000,

10. medzinárodný seminár Výpočtová štatistika sa uskutočnil 6. – 7. 12. 2001,

11. medzinárodný seminár Výpočtová štatistika sa uskutočnil 5. - 6. 12. 2002,

12. medzinárodný seminár Výpočtová štatistika sa uskutočnil 4. - 5. 12. 2003,

13. medzinárodný seminár Výpočtová štatistika sa uskutočnil 2. - 3. 12. 2004,

14. medzinárodný seminár Výpočtová štatistika sa uskutočnil 1. - 2. 12. 2005,

15. medzinárodný seminár Výpočtová štatistika sa uskutočnil 7. - 8. 12. 2006,

16. medzinárodný seminár Výpočtová štatistika sa uskutočnil 6. - 7. 12. 2007,

17. medzinárodný seminár Výpočtová štatistika sa uskutočnil 4. - 5. 12. 2008,

18. medzinárodný seminár Výpočtová štatistika sa uskutočnil 3. - 4. 12. 2009,

19. medzinárodný seminár Výpočtová štatistika sa uskutočnil 2. - 3. 12. 2010,

20. medzinárodný seminár Výpočtová štatistika sa uskutočnil 1. - 2. 12. 2011 a

21. medzinárodný seminár Výpočtová štatistika sa uskutočnil 6. - 7. 12. 2012.

Príspevky 2. seminára boli opublikované v Informáciách SDŠS č. 1/1989 a od 3. seminára sa publikujú v samostatnom Zborníku príspevkov príslušného seminára. Od 14.

seminára sú príspevky publikované vo vedeckom časopise SŠDS FORUM STATISTICUM SLOVACUM. Príspevky mladých v rámci prehliadky prác mladých štatistikov a demografov boli publikované spolu s príspevkami účastníkov semimára Výpočtová štatistika. Počnúc 19-tym ročníkom seminára Výpočtová štatistika je vydávaný, z príspevkov zaslaných do Prehliadky prác mladých štatistikov a demografov, samostatný zborník, v spolupráci s Klubom Dispersus.

Zameraním seminára je problematika na rozhraní počítačových vied a štatistiky. Tematické okruhy posledných seminárov sa nemenia:

- praktické využitie paketov štatistických programov,
- práca s rozsiahlymi súbormi údajov,
- vyučovanie výpočtovej štatistiky a príbuzných predmetov,
- praktické aplikácie výpočtovej štatistiky,
- iné.

V čase konania seminára Výpočtová štatistika sa uskutočňuje aj ***prehliadka prác mladých štatistikov a demografov***. Táto akcia prebieha od 7. seminára. Na 8. medzinárodnom seminári prezentovalo svoje práce 5 mladých štatistikov a demografov, na 9. medzinárodnom seminári už bolo 20 prác mladých štatistikov a demografov, na 10. bolo prihlásených 26 prác a na 11. bolo prihlásených 18 prác, ale vzhľadom na niekoľko prác vypracovaných skupinou autorov bol počet účastníkov vyšší než predošlý rok. Na 12. seminári bolo prihlásených 19 prác, pričom niektoré sú prácou viacerých autorov. Na ďalšom 13. seminári bolo prihlásených 9 prác od 12 autorov. V rámci 14. seminára bolo prihlásených 15 sólových prác mladých autorov. Na 15. seminári bolo prihlásených 20 prác mladých autorov. V rámci 16. seminára bolo prihlásených 17 sólových prác mladých autorov. V rámci 17. seminára bolo prihlásených 15 sólových prác mladých autorov. V 18. ročníku bolo prihlásených 12 sólových prác mladých autorov. V 19. ročníku bolo prihlásených 15 prác autorov. V 20. ročníku seminára bolo prihlásených 15 prác mladých autorov. V aktuálnom ročníku 21. seminára bolo prihlásených 19 prác mladých autorov.

Prípadní záujemcovia z radov mladých štatistikov a demografov (za mladých považujeme štatistikov a demografov pred ukončením 2. stupňa vysokoškolského štúdia) môžu získať informácie na www.ssds.sk , blok akcie a na e-mailových adresách:

chajdiak@statis.biz ; jan.luha@fmed.uniba.sk ; iveta.stankovicova@fm.uniba.sk.

Informácie o najbližšom seminári získate na webovskej stránke SŠDS, resp. na portáli ŠÚ SR http://portal.statistics.sk/showdoc.do?docid=1014 v bloku Slovenská štatistická a demografická spoločnosť.

doc. Ing. Jozef Chajdiak, CSc.
STU Bratislava
predseda SŠDS

RNDr. Ján Luha, CSc.
LFUK Bratislava
vedecký tajomník SŠDS

doc. Ing. Iveta Stankovičová, PhD.
FM UK Bratislava
predsedníčka Programového a organizačného výboru seminára Výpočtová štatistika

OBSAH / CONTENTS

SAS® Business Analytics Software

Data Management | Analytics | Reporting | Business and Industry Solutions

Môžu sa vaše štatistické odhady **zlepšiť o 96%,** ak využijete kvalitný analytický nástroj?

Áno, môžu. SAS vám dáva The Power to Know®.

SAS Business Analytics pomáha organizáciám zo všetkých odvetví objavovať inovatívne spôsoby ako zvyšovať ziskovosť, znižovať riziká, predikovať trendy, meniť dáta na informácie a získavať tým skutočnú konkurenčnú výhodu.

sas | THE POWER TO KNOW.

Čo ak existuje **softvér**, ktorý zabezpečí vašim študentom úspešnú **budúcnosť?**

Áno, existuje.

Pripojte sa k vyše 3000 univerzitám a začnite využívať SAS pre vašu výuku:

- poskytnete tým svojim študentom prístup k špičkovým technológiám pre riadenie rizík, marketingovú analytiku, finančné riadenie a pod.,
- pripravíte ich pre prácu v najväčších bankách, poisťovniach a telekomunikáciách, nielen na Slovensku, ale i na celom svete,
- získate prístup k medzinárodným konferenciám a početnej SAS komunite učiteľov a študentov.

▸▸ **www.sas.com/academic**

sas | THE POWER TO KNOW.

Pokyny pre autorov

Proces tvorby jednotlivých čísiel vedeckého recenzovaného časopisu FORUM STATISTICUM SLOVACUM (FSS) sa riadi dokumentom **Pu**blikačná **etika**. Jednotlivé čísla sú prevažne tematicky zamerané zhodne s tematickým zameraním akcií SŠDS. Príspevky v elektronickej podobe prijíma zástupca redakčnej rady na elektronickej adrese uvedenej v pozvánke na konkrétne odborné podujatie Slovenskej štatistickej a demografickej spoločnosti. Akceptujeme príspevky v slovenčine, češtine, angličtine, nemčine, ruštine a výnimočne po schválení redakčnou radou aj inom jazyku. Názov word-súboru uvádzajte a posielajte v tvare: ***priezvisko_nazovakcie.docx, resp. doc.***

Forma: Príspevky písané výlučne len v textovom editore MS WORD, verzia 6 a vyššia, písmo Times New Roman CE 12, riadkovanie jednoduché (1), formát strany A4, všetky okraje 2,5 cm, strany nečíslovať. Tabuľky a grafy v čierno-bielom prevedení zaradiť priamo do textu článku a označiť podľa šablóny. Bibliografické odkazy uvádzať v súlade s normou STN ISO 690 a v súlade s medzinárodnými štandardami. Citácie s poradovým číslom z bibliografického zoznamu uvádzať priamo v texte.

Rozsah: Maximálny rozsah príspevku je 6 strán.

Príspevky sú recenzované. Redakčná rada zabezpečí posúdenie príspevku oponentom.

Príspevky nie sú honorované, poplatok za uverejnenie akceptovaného príspevku je minimálne 30 €. Za každú stranu naviac je poplatok 5 €.

Štruktúra príspevku: ***(Pri písaní príspevku využite elektronickú šablónu: http://www.ssds.sk/ v časti Vedecký časopis, Pokyny pre autorov.). Časti v angličtine sú povinné!***

Názov príspevku v slovenskom jazyku **(štýl Názov: Time New Roman 14, Bold, centrovať)**

Názov príspevku v anglickom jazyku **(štýl Názov: Time New Roman 14, Bold, centrovať)**

Vynechať riadok

Meno1 Priezvisko1, Meno2 Priezvisko2 (štýl normálny: Time New Roman 12, bold, centrovať)

Vynechať riadok

Abstrakt: Text v slovenskom jazyku, max.10 riadkov (štýl normálny: Time New Roman 12).

Abstract: Text v anglickom jazyku, max. 10 riadkov (štýl normálny: Time New Roman 12).

Kľúčové slová: V slovenskom jazyku, max. 2 riadky (štýl normálny: Time New Roman 12).

Key words: V anglickom jazyku, max. 2 riadky (štýl normálny: Time New Roman 12).

JEL classification: Uviesť kódy klasifikácie podľa pokynov v:
<http://www.aeaweb.org/journal/jel_class_system.php>

Vynechať riadok a nastaviť si medzery odseku pre nadpisy takto: medzera pred 12 pt a po 3 pt. Poznámky pod čiarou: Time New Roman 10, zarovnať. Nasleduje vlastný text príspevku v členení:

1. **Úvod** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať,)*
2. **Názov časti 1** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*
3. **Názov časti . . .**
4. **Záver** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

Vlastný text jednotlivých častí je písaný štýlom Normal: písmo Time New Roman 12, prvý riadok odseku je odsadený vždy na 1 cm, odsek je zarovnaný s pevným okrajom. Riadky medzi časťami a odsekmi nevynechávajte. Nastavte si medzi odsekmi medzeru pred 0 pt a po 3 pt.

5. **Literatúra** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

[1] Písať podľa normy STN ISO 690

[2] GRANGER, C.W. – NEWBOLD, P. 1974. Spurious Regression in Econometrics. In: Journal of Econometrics, č. 2, 1974, s. 111 – 120.

Adresa autora (-ov): ***Uveďte svoju pracovnú adresu!!!*** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, adresy vpísať do tabuľky bez orámovania s potrebným počtom stĺpcov a s 1 riadkom):*

Meno1 Priezvisko1, tituly1 (študenti ročník)
Pracovisko1 (študenti škola1)
Ulica1, 970 00 Mesto1
meno1.priezvisko1@mail.sk

Meno2 Priezvisko2 , tituly2 (študenti ročník)
Pracovisko2 (študenti škola2)
Ulica2, 970 00 Mesto2
meno2.priezvisko2@mail.sk

FORUM STATISTICUM SLOVACUM

vedecký recenzovaný časopis Slovenskej štatistickej a demografickej spoločnosti

Vydavateľ:
Slovenská štatistická a demografická spoločnosť
Miletičova 3
824 67 Bratislava 24
Slovenská republika

Redakcia:
Miletičova 3
824 67 Bratislava 24
Slovenská republika

Fax:
02/39004009

e-mail:
chajdiak@statis.biz
jan.luha@fmed.uniba.sk

Registráciu vykonalo:
Ministerstvo kultúry Slovenskej republiky

Registračné číslo:
3416/2005

Evidenčné číslo:
EV 3287/09

Tematická skupina:
B1

Dátum registrácie:
22. 7. 2005

Objednávky:
Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava 24
Slovenská republika
IČO: 178764
DIČ: 2021504276
Číslo účtu: 0011469672/0900

ISSN 1336-7420

Ročník:
VIII.

Číslo:
7/2012

Cena výtlačku: 30 EUR
Ročné predplatné: 120 EUR